

# Multivariate Stochastic Volatility with Machine Learning-Augmented Regime Detection for Cross-Market Financial Contagion: A Bayesian Nonparametric and Ensemble-Learning Comparative Framework

**Joshua Christopher**

MBA in Artificial Intelligence and Data Science, CGPA 9.02/10  
SRM Institute of Science and Technology, Kattankulathur, Chennai, India  
Postgraduate Award in Finance & Investment Management, University of Liverpool, UK  
*Founder, Algol Ventures | Former Junior Trader, City Investment Solutions (UK, Remote)*

[algolventures@gmail.com](mailto:algolventures@gmail.com)

## Abstract

This paper extends the empirical framework of Christopher (2026), which documents a 49% contagion intensification between the S&P 500 and Nifty 50 during joint high-volatility regimes using GJR-GARCH and rolling-threshold regime classification, along two dimensions: replacing the univariate GARCH specification with a jointly-estimated multivariate stochastic volatility (MSV) model, and replacing the fixed three-regime classification with a data-driven layer combining an infinite Hidden Markov Model (iHMM) with an ensemble machine-learning classifier. A full Bayesian implementation — a Gibbs/FFBS-estimated MSV with ancillarity-sufficiency interweaving, and an iHMM estimated via truncated Dirichlet-process, hierarchical Dirichlet process, and beam-sampling approximations — finds a posterior mean correlation of 0.824 between the two markets' latent volatility innovations, and a contagion-intensification estimate ranging from -49.5% to +295.7% depending on estimator, regime-partition choice, sample period, and bias-correction. This sensitivity is itself a central finding: correlation-based estimates prove too methodology-dependent to report as a single magnitude, whereas the bias-immune Bayesian volatility-innovation correlation remains stable under subsample and prior perturbation. Freed from a fixed state-count ceiling, beam sampling recovers 11-12 active regimes versus 14-15 under truncated alternatives. Full derivations and convergence diagnostics are provided in the appendices.

**Keywords:** multivariate stochastic volatility; hierarchical Dirichlet process; infinite hidden Markov model; beam sampling; Markov chain Monte Carlo; ancillarity-sufficiency interweaving; financial contagion; regime detection; ensemble machine learning.

**JEL Classification:** C11 (Bayesian Analysis), C22 (Time-Series Models), C15 (Simulation Methods), C45 (Neural Networks and Related Topics), G01 (Financial Crises), G15 (International Financial Markets).

## 1. Introduction

Financial contagion research has evolved along two largely separate methodological tracks. The first, rooted in classical econometrics, uses GARCH-family models and researcher-defined regime thresholds to characterise volatility and classify market states (as in the author's own prior work). The second, emerging more recently from the machine learning literature, applies flexible, data-driven methods - hidden Markov models, ensemble classifiers, and increasingly graph neural networks - to detect regimes and model contagion without imposing a fixed structure in advance. These two tracks rarely intersect directly within a single study, and where they do, the comparison is rarely conducted on a controlled, identical dataset.

This paper is designed explicitly as that controlled comparison. By reusing the exact dataset and time period from the author's first paper (S&P 500 and Nifty 50, 2007-2024), any differences in regime classification or contagion estimates can be attributed to the methodological change itself, rather than to differences in data. As Section 2 documents, existing work tends to adopt either the classical, fixed-threshold track or the flexible, data-driven track, and rarely combines a genuinely nonparametric Bayesian volatility-and-regime specification -- multivariate

stochastic volatility jointly estimated with an infinite Hidden Markov Model -- with a supervised ensemble robustness check, on a controlled developed-emerging market pair.

This paper's contribution is fourfold. First, methodological: it specifies and, in Section 6, actually estimates a bivariate multivariate stochastic volatility (MSV) model jointly with a hierarchical-Dirichlet-process hidden Markov model (HDP-HMM), estimated via Gibbs sampling with ancillarity-sufficiency interweaving (ASIS) and, ultimately, literal beam sampling -- a combination not previously documented in the reviewed literature (Section 2). Second, empirical: it provides a direct, same-data comparison between a fixed-threshold GARCH approach (the author's prior work) and a jointly-estimated, ML-augmented Bayesian nonparametric approach, isolating the effect of methodology from the effect of data. Third, diagnostic: it uses supervised ensemble classifiers as an independent robustness check on the unsupervised Bayesian regime detection, a comparative approach not identified elsewhere in the reviewed literature. Fourth, methodological-cautionary: by reporting eight structurally different estimates of the same contagion-intensification effect side by side (Table 17) -- including a direct application of Forbes & Rigobon's (2002) standard bias correction, which reverses the effect's sign rather than resolving it (Section 8.1) -- it documents empirically both how sensitive this class of statistic is to estimator and regime-partition choice, and a concrete case in which the literature's own standard remedy for that sensitivity is itself unreliable, with implications for how this class of statistic should be reported beyond this paper's specific dataset. Fifth, reproducibility: Appendix B derives every conjugate-posterior and filtering step actually used from Bayes' rule, reports real convergence diagnostics computed on this paper's own saved posterior draws, and checks the sensitivity of the headline estimates to the samplers' own concentration-hyperparameter and prior choices -- an attempt at making the paper independently re-implementable rather than only citable.

## 2. Literature Review

### 2.1 GARCH-Deep Learning Hybrids

A substantial and fast-growing literature combines classical GARCH-family models with deep learning architectures to improve volatility forecasting. Araya, Aduda, Berhane & Wong (2024) integrate SARIMA, GARCH-family models, CNN, and Bidirectional LSTM, using GARCH residuals as network inputs; their hybrid designs report significant error reductions on volatile assets relative to standalone GARCH. Kakade, Jain & Mishra (2022) build a hybrid ensemble-learning GARCH-LSTM model specifically for Value-at-Risk forecasting, finding the hybrid materially outperforms either component alone during stress periods. Garcia-Medina & Aguayo-Moreno (2024) apply an LSTM-GARCH hybrid to cryptocurrency portfolio volatility, and Petrosino, Bacco, Salvati, Merone & Papi (2025) go a step further, replacing the LSTM stage with a temporal fusion transformer for ETF volatility prediction -- evidence that the GARCH-neural-hybrid literature is actively migrating toward attention-based architectures. Chung, Espinoza & Quispe (2025) take the complementary methodological position: using a direct GARCH-versus-deep-learning comparison under structural breaks, they show that neglecting regime change systematically overstates volatility persistence across every model class tested, including the deep-learning hybrids -- and explicitly call for future work integrating regime-change detection into hybrid architectures, which is the gap this paper addresses directly.

The hybridisation strategy itself varies considerably across this literature, which matters for interpreting what a 'GARCH-DL hybrid' result actually shows. Xu, Liechty, Benthall, Skar-Gislinge & McComb (2024) propose GARCH-Informed Neural Networks (GINN), which couple an LSTM with a GARCH-derived loss term in the style of physics-informed neural networks -- using GARCH as a soft constraint on training rather than as a feature-engineering preprocessing step, a structurally different fusion mechanism from the sequential architectures above. Ben Romdhane & Boubaker (2026) invert the usual ordering entirely, cascading HAR then LSTM then GARCH(1,1) on Brent Crude futures, using GARCH as a final stage that models the volatility of the LSTM's own forecast errors rather than as an input signal. Nsengiyumva, Mung'atu & Ruranga (2025) extend the standard GARCH-then-LSTM ordering to a frontier-economy foreign exchange setting (Rwandan Franc against USD/EUR/GBP), reporting up to a 10% Value-at-Risk accuracy gain during high-volatility periods, and

Amirshahi & Lahmiri (2023) test hybrid DL-GARCH-family combinations across 27 cryptocurrencies simultaneously, a considerably broader cross-sectional scope than single-asset crypto studies. Taneva-Angelova & Granchev (2025) offer a useful counterweight to the hybridisation trend: benchmarking GARCH, ARIMA, and HAR-RV directly against LSTM, CNN-LSTM, a vanilla Transformer, and PatchTST-lite on major U.S. equity indices -- rather than fusing GARCH with a neural network, they compare them head-to-head, finding the transformer models win on accuracy, especially in crisis periods, but at a genuine cost in interpretability that this literature does not always acknowledge.

## **2.2 Regime Detection: Hidden Markov Models and Ensemble Methods**

Hidden Markov Models remain the dominant tool for latent regime detection in financial time series (Hamilton, 1989, and a large subsequent literature). Ang & Bekaert (2002) is the foundational application to cross-market asset allocation, showing that a Markov-switching correlation structure materially changes optimal international portfolio weights relative to a single-regime model -- an early demonstration that regime-conditional dependence, not just regime-conditional volatility, matters for cross-market analysis. Nguyen & Nguyen (2015) apply an HMM directly to stock selection, and Boukardagha (2026) proposes an explainable regime-aware investing framework using a Wasserstein HMM with interpretable regime tracking. More recently, ensemble machine learning has been combined with HMMs to improve robustness: Zolotareva (2021) uses a two-stage XGBoost model to classify trending-versus-flat market regimes for long-term investment decisions, and Sanchez-Fernandez, Diez-Gonzalez, Huerga-Perez & Perez (2026) combine Gaussian mixture regime identification with gradient boosting for multivariate financial regime prediction. Gupta, Kapoor, Gupta & Natesan (2025) generalise this combination into an explicit multi-model ensemble-HMM voting framework -- bagging and boosting tree-based classifiers voting alongside HMM state output on Russell 3000 and S&P 500 ETF data -- directly motivating the ML-augmentation layer proposed in Section 3.4 of this paper.

A second, more critical strand of this literature questions the standard HMM's own specification rather than augmenting it. Nasri, Remillard & Thioub (2024) develop formal goodness-of-fit tests for HMMs and show numerically that classical information criteria (AIC, BIC, ICL, DIC, WAIC) are frequently unreliable for recovering the true number of regimes -- direct, methodology-level support for this paper's decision to treat regime count as a quantity to be estimated nonparametrically (Section 3.3) rather than selected by information criterion, and the same conclusion Section 5.5's own BIC/AIC results reach empirically on this paper's data. Bulla & Bulla (2006) attack a different HMM assumption: the implicit geometric distribution of regime durations, which is empirically wrong for financial time series, proposing hidden semi-Markov models with flexible sojourn-time distributions as a remedy -- a specification concern this paper does not address but that is worth flagging as a further source of model risk beyond the regime-count question this paper focuses on. Song (2014) combines two nested nonparametric hierarchies (one over transition probabilities, one over state-specific parameters) to let a Bayesian HMM jointly estimate the regime count AND distinguish recurring regime-switching from one-off structural breaks in U.S. real interest rates -- an econometric precedent, independent of the Maheu/Yang/Li lineage discussed in Section 2.3, for treating regime count nonparametrically. Tampouris & Dritsaki (2026) take a hierarchical route of a different kind, nesting regimes inside an unobserved macro-financial 'meta-regime' with its own transition matrix, which is a useful point of contrast on what hierarchy buys a regime-switching model when it adds a level ABOVE the regime layer rather than making the regime count itself nonparametric, as the HDP-HMM in Section 3.3 does.

## **2.3 Nonparametric Bayesian Regime-Switching and MCMC Methodology**

This paper's estimation strategy draws on two methodological literatures that rarely appear together in finance applications. The first is Bayesian stochastic volatility estimation via auxiliary mixture sampling and forward-filtering backward-sampling (Kim, Shephard & Chib, 1998), and its efficiency refinement via ancillarity-sufficiency interweaving (Kastner & Fruhwirth-Schnatter, 2014), which this paper extends to the bivariate case in Section 6.6. Kastner, Fruhwirth-Schnatter & Lopes (2017) extend interweaving-based estimation to large multivariate factor stochastic volatility models, and Bitto & Fruhwirth-Schnatter (2019) apply interweaving-style

non-centering to achieve shrinkage in time-varying-parameter state-space models more generally -- both evidence that the interweaving idea generalises well beyond the univariate case this paper's diagonal-D extension (Section 6.6) also targets.

The specific 7-component Gaussian mixture approximation this paper uses (Appendix A.1) is itself one point in a wider design space. Omori, Chib, Shephard & Nakajima (2007) extend it to a 10-component mixture that additionally handles the leverage effect -- correlation between return and volatility innovations -- which the basic KSC mixture used throughout Section 6 cannot accommodate, a natural direction for future refinement flagged again in Section 10. Pitt & Shephard (1999) offer a methodologically distinct alternative to the mixture-sampling/FFBS route entirely: the auxiliary particle filter, a sequential Monte Carlo scheme for non-Gaussian nonlinear state-space models that avoids the mixture approximation altogether at the cost of a different (particle-degeneracy) failure mode. McCausland, Miller & Pelletier (2021), building on McCausland's earlier simulation-smoothing work, propose the HESSIAN method -- a precision-based, all-at-once simulation smoother exploiting banded Cholesky structure -- as a computationally efficient alternative to FFBS for exactly the multivariate SV setting Section 3.1 specifies, avoiding both the KSC-style approximation and the single-state-at-a-time sampling FFBS performs.

The second methodological literature is nonparametric Bayesian regime-switching: the infinite Hidden Markov Model (Beal, Ghahramani & Rasmussen, 2002) formalised as a hierarchical Dirichlet process (Teh, Jordan, Beal & Blei, 2006), estimated either via direct-assignment Gibbs sampling in the style used operationally by Fox, Sudderth, Jordan & Willsky's (2011) sticky HDP-HMM, or via the literal beam sampler of Van Gael, Saatci, Teh & Ghahramani (2008) that removes weak-limit truncation entirely using auxiliary slice variables in the tradition of Walker's (2007) and Kalli, Griffin & Walker's (2011) slice samplers for Dirichlet process mixtures more broadly. The truncation-versus-exactness trade-off this paper works through across Sections 6.3, 6.7 and 6.8 is itself a well-studied methodological question: Ishwaran & James (2001) derive explicit truncation-error bounds for blocked/truncated stick-breaking Gibbs samplers, giving a principled way to choose the truncation level  $L$  used in Sections 6.3 and 6.7 rather than the ad hoc  $L = 15$  this paper adopts; Papaspiliopoulos & Roberts (2008) develop 'retrospective sampling,' a truncation-free exact algorithm that -- like beam sampling -- avoids both finite truncation and full marginalisation, via a different (retrospective rather than slice-based) mechanism; and Jain & Neal (2004) introduce split-merge moves that let a Dirichlet process mixture sampler jump between very different clusterings in a single step, addressing a slow-mixing failure mode that conjugate Gibbs sampling -- and, potentially, beam sampling -- can also suffer from, a possible refinement to Section 6.8's sampler not pursued here.

Applications of this machinery specifically to finance remain comparatively rare. Maheu & Yang (2016) apply an infinite HMM to short-term interest rates, Li, Maheu & Yang (2024) extend an infinite HMM with stochastic volatility to equity, FX, and oil returns -- the closest direct precedent to this paper's MSV-plus-iHMM combination -- and Li & Yang (2025) further hybridise the infinite HMM with GARCH dynamics for short-term rates. Shi & Song (2016) apply an infinite HMM to an entirely different finance problem, dating speculative-bubble and collapse episodes without pre-specifying a regime count, by authors independent of the Maheu/Yang/Li group; and Jin & Maheu (2016), sharing an author with that group, build Dirichlet-process mixtures of Wishart distributions to model the full conditional distribution of realised covariance matrices -- a Bayesian nonparametric approach to multivariate volatility that is methodologically adjacent to this paper's core topic despite not using the HMM route at all. None of these finance applications, however, jointly estimate a genuinely multivariate stochastic volatility specification (correlated latent volatility innovations across markets, as in Section 3.1) alongside the infinite-HMM regime layer with a literal, untruncated sampler -- the combination Section 6 of this paper attempts.

## **2.4 Financial Contagion Measurement Methodology**

Measuring contagion, as distinct from ordinary cross-market interdependence, has its own methodological literature independent of the volatility-modelling choice. Forbes & Rigobon (2002) is the foundational critique: naive correlation comparisons across calm and turmoil periods are biased upward by heteroskedasticity in the

underlying returns, so an apparent 'contagion' effect can be a pure statistical artefact of the higher-volatility period itself. Bekaert, Harvey & Ng (2005) and Corsetti, Pericoli & Sbracia (2005) extend and qualify this critique, the latter showing that some, but not all, of a naive correlation increase typically survives correction, and that the conclusion is itself sensitive to modelling choices -- a finding directly echoed by this paper's own eight-estimator sensitivity result (Sections 5.7, 6.4, 6.7, 6.8, and Section 8) and directly demonstrated, not merely echoed, by Section 8.1's own application of the Forbes-Rigobon correction. Rodriguez (2007) proposes a copula-based alternative that separates dependence structure from marginal volatility dynamics entirely.

A methodologically distinct branch of this literature abandons correlation altogether in favour of network/connectedness measures. Diebold & Yilmaz (2012) introduce the directional spillover index, decomposing a forecast-error-variance-based connectedness measure into to/from/net spillovers per market via a VAR framework -- a genuinely different identification strategy from correlation-based tests, and one this paper does not implement but flags as the natural next robustness check beyond Section 8.1's correlation-based correction. Ando, Greenwood-Nimmo & Shin (2022) extend connectedness measurement into a quantile-VAR framework, finding tail-quantile spillovers are far stronger than mean spillovers across sovereign and financial-sector credit risk, and Tiwari, Jena, Trabelsi & Hammoudeh (2022) apply quantile connectedness networks to emerging markets, finding contagion intensity is itself regime/quantile-dependent -- the same qualitative message as Bhardwaj, Miklosevic & Gambhir (2026) below, reached through an entirely different estimator family.

A third strand questions whether 'contagion' is even a well-defined, empirically distinguishable object. Pericoli & Sbracia (2003) survey the (mutually inconsistent) definitions used across this literature -- correlation-shift, probability-shift, volatility-transmission, co-movement-beyond-fundamentals -- and show these definitions can disagree on whether the SAME crisis episode exhibits contagion at all. Rigobon (2019), written by one of the original Forbes-Rigobon authors, goes further and argues that correlation, VAR-based spillover measures, and ARCH-family approaches EACH carry irreducible identification problems -- endogeneity, omitted common factors, heteroskedasticity -- such that no existing technique cleanly separates contagion from ordinary interdependence; this is close to a co-author of the Forbes-Rigobon (2002) test self-qualifying the very literature it founded, and it is the closest available precedent, from within the field, for this paper's own estimator-sensitivity findings (Section 7, Section 8.1).

On the specific US-India / developed-emerging pairing used in this paper, Thangamuthu, Maheshwari & Naik (2022) document volatility spillovers into the Indian market from the US and other major economies around the COVID-19 period, Nguyen, Phan, Nguyen & McMillan (2022) use a DCC-GARCH framework to measure financial contagion across the GFC and COVID-19 crises, and Bhardwaj, Miklosevic & Gambhir (2026) study regime-dependent integration and contagion between India and global equity markets directly. Agarwal, Dhankhar & Mehla (2024) apply both DCC-GARCH and the Diebold-Yilmaz spillover index to the same US-India-China-Germany-Japan set, finding India is a net RECEIVER of volatility spillovers with the US among its largest sources -- the closest direct comparator to this paper, since it applies both a correlation-based and a network-based contagion measure to essentially the same market pair. Davidescu, Manta, Hapau, Gruiescu & Vacaru (Boita) (2023) apply DCC-GARCH, the Diebold-Yilmaz index, and Markov-switching VAR side-by-side to four Central and Eastern European emerging markets, finding contagion intensity from the US and Germany rises sharply in the COVID-19 high-volatility regime under all three methods simultaneously -- confirming both the substantive relevance of the developed-emerging pairing used in this paper and the continued reliance, across this recent literature, on exactly the class of estimators (DCC-GARCH, correlation-based) this paper shows to be estimator-sensitive on its own data.

## **2.5 Graph Neural Networks for Systemic Risk and Contagion**

Graph neural networks (GNNs) are an increasingly active tool for systemic financial risk because they can represent institutions or markets as nodes and exposures or comovements as edges, learning risk propagation directly from network structure. Matsunaga, Suzumura & Takahashi (2019) and Feng, He, Wang, Luo, Liu & Chua (2019) are early, widely-cited demonstrations that graph/relational architectures improve stock-market

prediction over models that treat assets independently -- evidence for the underlying premise (that market relationships carry predictive information a non-relational model discards) rather than direct systemic-risk applications. Gonon, Meyer-Brandis & Weber (2024) close that gap, showing GNNs can directly approximate systemic risk measures by exploiting relational information in financial networks. Islam et al. (2026) extend this specifically to cross-market contagion between banking systems and cryptocurrency markets, noting that most existing GNN-based studies build graphs using fixed correlation thresholds or static classifications -- an assumption that sits uneasily with the abrupt strengthening of dependencies documented during systemic crises, including in the author's own prior empirical findings. (Published in a comparatively low-visibility venue; retained here because the specific observation about fixed-threshold graph construction is independently well-motivated and directly relevant, not on the strength of the venue.)

That fixed-threshold-graph critique is, in fact, already being addressed at the frontier of this literature, which is worth documenting precisely because it sharpens rather than dismisses the concern. Xiang, Cheng, Shang, Zhang & Liang (2022) propose THGNN, which explicitly rebuilds the company relation graph on EVERY trading day from short-window price co-movement rather than using a fixed or manually-labelled graph, combined with transformer-based temporal encoding and heterogeneous graph attention, deployed and back-tested with live trading on the S&P 500 and CSI 300. Qian, Zhou, Zhao, Chen, Yao, Wang, Liu, Yu, Zhang & Zhou (2024) build discrete dynamic, multi-relational graphs across industry, investment-bank, and pairwise stock relations that evolve over time, outperforming static-graph baselines. Uddin, Tao & Yu (2023) make essentially this paper's own point independently, in a finance (not computer science) venue: they learn, rather than pre-specify, evolving firm/sector network structure jointly with a diffusion mechanism for asset comovement, explicitly contrasted against the static-market-network assumption. Kumar, Umeorah & Alochukwu (2026) combine a temporal graph attention network with Diebold-Yilmaz volatility spillover indices across eight global stock indices, explicitly framed around cross-market contagion rather than single-series forecasting -- and, notably, find no systematic benefit from dynamically updating the graph over a well-chosen fixed spillover network, a nuanced qualification of the dynamic-graph critique worth weighing against the papers above. Zhou, Xie, Wang, Gong & Zhu (2025) apply hierarchical evolving graph-structure learning to cryptocurrency-versus-traditional-market volatility, a peer-reviewed, methodologically stronger analog to Islam et al.'s (2026) banking/crypto topic. Yu & Luo (2025) apply time-varying, multi-layer dynamic graph convolution to risk contagion through enterprise-group financial networks -- not a banking-sector application specifically, but a further demonstration that dynamic graph construction, not fixed correlation thresholds, is where this literature is heading. Citterio, D'Errico & Visentin (2024) apply a dynamic GNN to a temporal interest-rate-swap network for regulatory margin-call forecasting -- the closest in spirit to Gonon et al. (2024) already discussed, though as an arXiv preprint without confirmed peer review, it is cited with that caveat attached. This literature motivates the forward-looking link to graph-based methods discussed in Section 10: the iHMM-derived regime states and engineered features from Sections 5 and 6 are designed to serve as exactly the kind of dynamic, data-driven graph input this literature is only now converging on, rather than the fixed-threshold input Islam et al. (2026) critique and this paper's own forward-looking discussion echoes.

## **2.6 Positioning of This Paper**

No study identified in this review combines (i) a jointly-estimated multivariate stochastic volatility specification, (ii) Bayesian nonparametric (infinite HMM) regime detection estimated with a literal, untruncated sampler, and (iii) an ensemble machine learning robustness layer, applied to a developed-emerging market contagion pair. Section 2.3's closest precedents (Li, Maheu & Yang, 2024; Maheu & Yang, 2016) pair an infinite HMM with stochastic volatility but not with a genuinely multivariate (cross-market-correlated) volatility process, and none estimate the regime layer via literal beam sampling rather than a truncated or weak-limit approximation. Section 2.4's contagion-measurement literature repeatedly finds that the size of a contagion effect is sensitive to estimator choice, and Rigobon's (2019) own retrospective on the field goes further, arguing no existing estimator cleanly identifies contagion at all -- but to this paper's knowledge, no study has quantified that sensitivity by holding the data fixed and varying only the estimator and regime-detection method, as this paper does across eight structurally

different approaches, including a direct application of the field's own standard bias correction (Table 17, Section 8.1). Section 2.5's GNN literature is, encouragingly, moving toward exactly the dynamic graph construction this paper's forward-looking discussion (Section 10) anticipates, though Kumar, Umeorah & Alochukwu's (2026) null result on dynamic-versus-fixed graphs is a useful caution against assuming dynamism is unconditionally better. This paper is positioned to fill its specific combination of methods, while remaining directly comparable to the author's own prior fixed-threshold GARCH results through the shared dataset.

## 2.7 Synthesis Table

Table 1 summarises method, data, and headline finding for a representative cross-section of the papers reviewed above, organised by the same five thematic groupings as Sections 2.1-2.5, to make the landscape's overall shape easier to see at a glance than the discursive review above allows on its own.

| Study                                     | Method                                                   | Key finding relevant to this paper                                                         |
|-------------------------------------------|----------------------------------------------------------|--------------------------------------------------------------------------------------------|
| Araya et al. (2024)                       | SARIMA/GARCH + CNN-BiLSTM hybrid                         | GARCH-residual neural inputs cut forecast error on volatile assets                         |
| Chung, Espinoza & Quispe (2025)           | GARCH vs. deep learning under structural breaks          | Ignoring regime change overstates volatility persistence in every model class              |
| Taneva-Angelova & Granchev (2025)         | GARCH/ARIMA/HAR-RV vs. LSTM/Transformer/PatchTST-lite    | Transformers win on accuracy but cost interpretability                                     |
| Ang & Bekaert (2002)                      | Markov-switching correlation, asset allocation           | Regime-conditional dependence, not just volatility, changes optimal portfolios             |
| Nasri, Remillard & Thioub (2024)          | HMM goodness-of-fit tests                                | AIC/BIC/ICL/DIC/WAIC often fail to recover the true regime count                           |
| Gupta, Kapoor, Gupta & Natesan (2025)     | Ensemble-HMM voting (bagging/boosting + HMM)             | Direct precedent for this paper's Section 3.4 ML-augmentation layer                        |
| Kim, Shephard & Chib (1998)               | Mixture-sampling FFBS for stochastic volatility          | Foundational method underlying this paper's MSV estimation (Appendix A.1-A.2)              |
| Kastner & Fruhwirth-Schnatter (2014)      | ASIS interweaving, univariate SV                         | Extended to the bivariate case in this paper's Section 6.6/Appendix A.3                    |
| Fox, Sudderth, Jordan & Willsky (2011)    | Sticky HDP-HMM, direct-assignment Gibbs                  | Operational template for this paper's Section 6.7/Appendix A.4                             |
| Van Gael, Saatci, Teh & Ghahramani (2008) | Beam sampling, untruncated HDP-HMM                       | Template for this paper's Section 6.8/Appendix A.5, the least-approximated iHMM variant    |
| Li, Maheu & Yang (2024)                   | Infinite HMM + stochastic volatility, equity/FX/oil      | Closest precedent to this paper's MSV+iHMM combination; not genuinely multivariate         |
| Forbes & Rigobon (2002)                   | Heteroskedasticity-bias-corrected correlation            | Foundational contagion-measurement critique; directly applied in this paper's Section 8.1  |
| Corsetti, Pericoli & Sbracia (2005)       | Critique of the Forbes-Rigobon correction                | Explains this paper's Section 8.1 sign-reversal finding under large variance ratios        |
| Rigobon (2019)                            | Retrospective on contagion-identification methods        | Argues no existing estimator cleanly separates contagion from interdependence              |
| Diebold & Yilmaz (2012)                   | Forecast-error-variance spillover index                  | Network-based alternative to correlation, not implemented in this paper                    |
| Agarwal, Dhankhar & Mehla (2024)          | DCC-GARCH + Diebold-Yilmaz, US-India-China-Germany-Japan | Closest direct comparator: both a correlation- and network-based measure on a similar pair |
| Davidescu et al. (2023)                   | DCC-GARCH + Diebold-Yilmaz + MS-VAR, CEE markets         | Contagion intensity rises in the COVID regime under all three methods simultaneously       |
| Gonon, Meyer-Brandis & Weber (2024)       | GNN approximation of systemic risk measures              | Direct precedent for this paper's Section 10 GNN extension                                 |
| Xiang, Cheng, Shang, Zhang & Liang (2022) | THGNN, daily-rebuilt relation graph                      | Addresses the fixed-threshold-graph critique via dynamic graph construction                |
| Kumar, Umeorah & Alochukwu (2026)         | Temporal GNN + Diebold-Yilmaz, 8 global indices          | Null result: no systematic benefit from dynamic over well-chosen fixed graphs              |

Table 1. Representative cross-section of the literature reviewed in Sections 2.1-2.5; full citations and discussion in the corresponding subsections above. Not an exhaustive list of every paper cited in this section.

## 2.8 Closer Engagement with Five Core Methodological Sources

Sections 2.1-2.7 survey this paper's empirical and thematic neighbours -- GARCH-ML hybrids, HMM regime detection, contagion measurement, graph neural networks. The estimation machinery actually implemented in

Sections 3 and 6, however, is not a generic invocation of 'Bayesian stochastic volatility' or 'nonparametric HMMs': it is assembled from specific, identifiable algorithms in five papers, each adopted for a specific technical reason and each departed from in specific, disclosed ways. This subsection engages with each of the five individually -- what problem it solves, exactly what device this paper takes from it, and where the implementation here diverges -- rather than treating them only as citation-list entries already covered by the thematic groupings above.

Kim, Shephard and Chib (1998) supplies the specific numerical device that makes the FFBS step of Section 3.1/Appendix B.3 tractable at all: a seven-component mixture-of-normals approximation to the distribution of  $\log(\text{chi-squared}_1)$ , fitted by their own moment-matching procedure and tabulated in their Table 4 (component weights, means and variances). Squaring and log-transforming the observation equation turns the SV measurement equation into a linear-Gaussian one conditional on a mixture-component indicator, which is exactly what admits Kalman-filter FFBS instead of a particle filter or Metropolis-Hastings on the  $h_t$  path directly. `fast_ffbs.py` (Appendix A.2) hard-codes exactly KSC's seven constants rather than re-deriving them, and Appendix B.3's worked example traces through the same seven-way mixture-indicator sampling step KSC's own algorithm specifies. One limitation is inherited unmodified from the 1998 approximation and not independently corrected here: KSC's own paper documents a small residual bias in the mixture fit near the mode of the true log-chi-squared density, which their later co-authored refinement (Omori et al., 2007, below) addresses more directly. This paper does not additionally correct for that residual bias, a specific and disclosed gap rather than an assumption that KSC's approximation is exact.

Omori, Chib, Shephard and Nakajima (2007) is cited in this paper's own limitations discussion (Section 6.1) as the natural refinement not implemented here; this subsection is more specific about what, exactly, is given up by that choice. Beyond the tighter ten-component mixture approximation, Omori et al.'s central methodological contribution is a fully worked-out AR(1) structure for the mixture-sampling step that additionally accommodates a leverage effect -- correlation between the return innovation and the next period's volatility innovation, empirically well-documented in equity volatility generally and directly relevant to both S&P 500 and Nifty 50 given the negative-skewness patterns already reported in Section 4.2's descriptive statistics. The bivariate MSV model estimated in Section 6.2 assumes  $\text{eps}_{it}$  and  $\text{eta}_t$  are independent within each market by construction (Section 3.1): no leverage channel is modelled at all, in either market, anywhere in this paper. This is a real, substantive omission, not a cosmetic one -- a leverage-aware respecification following Omori et al.'s approach is exactly the kind of extension flagged for future work in Section 10 rather than attempted here, and any point estimate of  $\text{Sigma}_{\text{eta}}$  reported in Section 6.2 or Appendix C.1 should be read with that omission in mind.

Teh, Jordan, Beal and Blei (2006) is the source of the hierarchical Dirichlet process construction underlying every iHMM variant in Section 6 (6.3, 6.7, 6.8), specifically its stick-breaking representation -- a global measure  $G_0$  drawn from  $\text{DP}(\text{gamma}, H)$  that itself has  $\text{GEM}(\text{gamma})$  stick-breaking weights (Appendix B.4), with each row of the transition matrix drawn as  $G_j \sim \text{DP}(\text{alpha}, G_0)$  so that all rows share the same countable set of candidate next-states. Teh et al.'s own paper proposes several samplers for this construction; the direct-assignment sampler used in Section 6.7 (`full_ihmm_hdp.py`) implements specifically their Section 5.3 algorithm (auxiliary variable  $m_{jk}$  sampling followed by direct sampling of state indicators against the stick-breaking weights), not the Chinese-restaurant-franchise Gibbs sampler of their Section 5.1 that marginalises the mixing proportions entirely. This is a deliberate choice with a specific reason, not an arbitrary pick among equivalent options: Van Gael et al. (2008, below) argue, and this paper's own Appendix E.2 synthetic comparison confirms, that the CRF-marginalised sampler mixes poorly on sequentially correlated data of exactly the kind a time series HMM produces, because integrating out the mixing proportions induces strong dependence between the state sequence and the (marginalised-out) transition structure that a Gibbs sweep over individual state indicators struggles to break.

Van Gael, Saatchi, Teh and Ghahramani (2008) is the direct source of the beam-sampling algorithm used in Section 6.8 (`full_ihmm_beam.py`), and the paper's stated motivation for that algorithm is specifically the sequential-data mixing problem just described: rather than sampling state indicators one at a time conditional on their neighbours

(as the direct-assignment sampler does), beam sampling introduces auxiliary slice variables (following Walker, 2007, itself cited in this paper's reference list) that make only a finite subset of states 'reachable' at each time step, which permits an exact dynamic-programming forward-filtering/backward-sampling pass over the entire state sequence jointly -- the same forward-filter/backward-sample logic as Appendix B.3's FFBS derivation, but applied to a discrete latent state rather than a continuous log-volatility. This is precisely why Section 6.8 exists as a further refinement beyond Section 6.7 rather than the paper stopping at direct assignment, and Appendix E.2's synthetic finding -- that both HDP-coupled samplers recover the true  $K=3$  state count with full posterior confidence while the simplest independent-rows truncated sampler over-splits to a modal  $K=4$  -- is direct evidence, on a controlled test with a known answer, that progressing through Van Gael et al.'s algorithm was not a redundant step.

Fox, Sudderth, Jordan and Willsky (2011) proposes a further refinement not implemented anywhere in this paper: a 'sticky' HDP-HMM that adds an explicit self-transition bias parameter  $\kappa$  to the hierarchical prior, directly countering a documented failure mode of the plain HDP-HMM when applied to real (non-exchangeable, temporally persistent) sequences -- rapid switching among several redundant states that all represent the same underlying regime, which inflates the effective state count without adding genuine descriptive content. The truncated sampler of Section 6.3 addresses a structurally similar concern by a different mechanism: a  $\text{Dirichlet}(\alpha/15)$  sparsity-inducing prior on each transition row, which shrinks unneeded states' posterior mass toward zero rather than explicitly biasing toward self-transition. The two mechanisms are not identical -- sparsity shrinkage penalises the number of occupied states in aggregate, while Fox et al.'s stickiness specifically penalises within-state churn -- and this paper's own real-data results (the effective-state-count discussion in Section 7, and the truncated sampler's synthetic over-splitting documented in Appendix E.2) are consistent with the Dirichlet-sparsity mechanism being the weaker of the two for this purpose. Implementing Fox et al.'s explicit stickiness parameter across all three iHMM variants and re-running the effective-state-count comparison is a concrete, well-specified extension left for future work rather than folded into this round of revisions.

### 3. Methodology

#### 3.1 Multivariate Stochastic Volatility (MSV) Specification

Log returns for market  $i$  ( $i = 1$ : S&P 500,  $i = 2$ : Nifty 50) at time  $t$  are modelled as

$$r_{it} = \exp(h_{it} / 2) \cdot \text{eps}_{it}, \quad \text{eps}_{it} \sim \text{i.i.d. } N(0, 1) \quad (1)$$

with joint log-volatility dynamics

$$h_t = \mu + \Phi (h_{t-1} - \mu) + \eta_t, \quad \eta_t \sim N(0, \Sigma_\eta), \quad \Phi = \text{diag}(\phi_1, \phi_2) \quad (2)$$

where  $h_t = (h_{1t}, h_{2t})'$  is the pair of latent log-volatilities,  $\mu$  is the unconditional mean vector, and  $\Phi$  is diagonal so that each market's log-volatility follows its own AR(1) persistence while the two markets' volatility INNOVATIONS  $\eta_t$  are allowed to correlate through the full (non-diagonal) covariance matrix  $\Sigma_\eta$ . This is the paper's central structural departure from the author's prior work, in which S&P 500 and Nifty 50 volatility were estimated as independent univariate GJR-GARCH processes with their relationship inferred only afterward via rolling correlation of the fitted residuals: here, cross-market volatility linkage is a parameter of the model itself (the off-diagonal entry of  $\Sigma_\eta$ ) rather than a derived quantity. Each market's process is covariance-stationary whenever  $|\phi_i| < 1$ , with stationary marginal variance  $\text{Var}(h_{it}) = \Sigma_{\eta,ii} / (1 - \phi_i^2)$  and stationary cross-covariance  $\text{Cov}(h_{1t}, h_{2t}) = \Sigma_{\eta,12} / (1 - \phi_1 \cdot \phi_2)$ , used below as the initial state distribution for filtering.

#### 3.2 Observation-Equation Linearisation

Equation (1) is not linear-Gaussian in  $h_t$ , which rules out direct Kalman filtering. Following Kim, Shephard & Chib (1998), squaring and taking logs gives

$$y^*_{it} := \log(r_{it}^2 + c) = h_{it} + xi_{it}, \quad xi_{it} := \log(eps_{it}^2) + 1.2704 \quad (3)$$

for a small offset  $c > 0$  guarding against  $\log(0)$  on zero-return days, where  $E[\log(eps_{it}^2)] = -1.2704$  for  $eps_{it} \sim N(0,1)$ , so  $xi_{it}$  is mean-zero by construction but is chi-squared-distributed, not Gaussian. Kim, Shephard & Chib (1998) approximate its density with a 7-component Gaussian mixture,

$$p(xi_{it}) \sim \sum_{k=1..7} q_k \cdot N(xi_{it}; m_k, v_k) \quad (4)$$

with fixed weights, means, and variances ( $q_k, m_k, v_k$ ) fit once, offline, to closely match the true chi-squared-derived density; the exact constants used are given in Appendix A.1. Introducing an auxiliary mixture-component indicator  $s_{it}$  in  $\{1, \dots, 7\}$  for every  $(t, i)$ , conditional on  $s_{it} = k$  the pair  $(y^*_{it}, h_{it})$  becomes an exactly linear-Gaussian state-space model,

$$y^*_{it} - m_{\{s_{it}\}} = h_{it} + xi'_{it}, \quad xi'_{it} \sim N(0, v_{\{s_{it}\}}) \quad (5)$$

with state equation (2) unchanged. This is what makes exact forward-filtering/backward-sampling (FFBS; Carter & Kohn, 1994) tractable each Gibbs sweep, once  $s$  is resampled conditional on the current  $h$  (Section 3.5, Appendix A.2).

### 3.3 Infinite Hidden Markov Model (iHMM) Regime Detection

The iHMM (Beal, Ghahramani & Rasmussen, 2002), formalised as a hierarchical Dirichlet process (HDP; Teh, Jordan, Beal & Blei, 2006), places a nonparametric prior over the transition matrix, allowing the effective number of hidden regimes to be inferred from the data rather than fixed in advance at three (as in the author's prior rolling-threshold method). The generative model is

$$beta \sim GEM(gamma) \quad [stick-breaking \text{ prior over a countably infinite state space}] \quad (6)$$

$$pi_j | beta \sim DP(alpha, beta) \quad \text{for every state } j = 1, 2, \dots \quad [transition \text{ rows}] \quad (7)$$

$$z_t | z_{t-1} = j \sim pi_j, \quad x_t | z_t = k \sim F(theta_k), \quad theta_k \sim H \quad (8)$$

where  $beta$  is a shared, global state-popularity vector drawn once via GEM stick-breaking ( $beta_k = beta'_k \cdot \prod_{j < k} (1 - beta'_j)$ ,  $beta'_k \sim \text{Beta}(1, gamma)$ ), and every transition row  $pi_j$  is itself drawn from a Dirichlet process centred on that SAME  $beta$  -- the hierarchical coupling that ties all rows' notion of which states are common together, and that a set of independent per-row Dirichlet priors (used in this paper's simpler weak-limit approximation, Section 6.3) does not have. Applied to the jointly-estimated volatility series  $h_t$  (using the MSV posterior mean path,  $x_t = h_t$ ,  $F = N(.,.)$  Gaussian,  $H$  a conjugate normal-inverse-Wishart prior; Section 6.1), this tests directly whether the three-regime structure imposed previously is empirically supported or whether a data-driven approach detects additional structure -- for example, a distinct 'pre-crisis build-up' regime that a fixed-threshold method may absorb into the high-volatility state.

### 3.4 Machine Learning Ensemble Augmentation

Following the ensemble-HMM voting approach demonstrated in recent literature (Section 2.2), this paper adds a supervised machine learning layer to test the robustness of the iHMM's unsupervised regime classifications. Random Forest and Gradient Boosting classifiers (scikit-learn; the gradient-boosted-tree comparator specified here is implemented via scikit-learn's GradientBoostingClassifier rather than XGBoost or LightGBM, Section 5.6) are trained to predict regime labels using a feature set of realised volatility, 60-day rolling skewness and kurtosis, 60-day rolling cross-market correlation, and the lag-1 DCC correlation, with the HMM's own regime assignments used as the training target and known crisis-period dates (GFC, COVID-19) used as an independent validation check. Divergence between the unsupervised HMM classification and the supervised ML classification at specific points in time is used as a diagnostic: persistent disagreement would suggest the HMM is either detecting genuine structure the engineered features miss, or overfitting to volatility patterns without capturing broader market stress indicators (Section 5.6 finds the latter is the more likely reading). This comparative ML

layer is the paper's explicit AI/ML methodological contribution, distinct from the Bayesian statistical (iHMM) component.

### 3.5 Estimation Approach

MSV estimation proceeds via a blocked Gibbs sampler: mixture indicators  $s$  (Section 3.2) are resampled given  $h$ , then  $h$  is resampled jointly via FFBS given  $s$ , then  $(\mu, \Phi, \Sigma_{\eta})$  are resampled given  $h$  via their conjugate posteriors (Appendix A.2). The ancillarity-sufficiency interweaving strategy (ASIS; Kastner & Fruhwirth-Schnatter, 2014) is layered on top by writing  $\Sigma_{\eta} = D R D$  with  $D = \text{diag}(\sigma_1, \sigma_2)$  and  $R$  the innovation correlation matrix, and interweaving through the non-centred state  $g_t = D^{-1} (h_t - \mu)$ , which -- because  $\Phi$  and  $D$  are both diagonal and therefore commute -- follows the SAME- $\Phi$  recursion  $g_t = \Phi g_{t-1} + \xi_t$ ,  $\xi_t \sim N(0, R)$ ; this is what makes a tractable multivariate extension of the originally univariate ASIS scheme possible (full derivation and interweave-step formulas in Appendix A.3). The iHMM is estimated three ways with increasing fidelity to the exact specification: a truncated (weak-limit) Dirichlet-process approximation with independent symmetric-Dirichlet transition rows (Section 6.3); a direct-assignment Gibbs sampler that ties all rows together through the shared beta via Chinese-Restaurant-Franchise table-count auxiliary variables (Teh et al., 2006, Section 5; Section 6.7; Appendix A.4); and literal beam sampling (Van Gael et al., 2008), which removes the weak-limit truncation entirely using auxiliary slice variables  $u_t \sim \text{Uniform}(0, \pi_{\{z_{t-1}, z_t\}})$  that restrict, at each  $t$ , the set of states reachable with nonzero posterior mass to a dynamically-determined finite subset -- exact for the full unbounded state space, with new states instantiated on the fly via the Dirichlet-process stick-splitting identity (Section 6.8; full algorithm in Appendix A.5). The ML ensemble layer is trained and validated using standard time-series cross-validation (five-fold expanding window, no look-ahead).

### 3.6 Feasibility-Analog Specifications

Section 5's preliminary results use two tractable, real-data-estimable analogs of Sections 3.1 and 3.3, standing in for the full-scale Bayesian estimation that Section 6 subsequently attempts. Each market's volatility is fit separately via univariate GARCH(1,1) with Student- $t$  innovations,

$$\sigma_{it}^2 = \omega_i + \alpha_i \cdot r_{i,t-1}^2 + \beta_i \cdot \sigma_{i,t-1}^2, \quad r_{it} / \sigma_{it} \sim t(\nu_i) \quad (9)$$

and cross-market volatility linkage -- the role  $\Sigma_{\eta}$  plays in the full MSV specification -- is approximated by a dynamic conditional correlation layer (DCC(1,1); Engle, 2002) fit to the standardised residuals  $z_{it} = r_{it} / \sigma_{it}$ ,

$$Q_t = Q_{\text{bar}} (1 - a - b) + a \cdot z_{t-1} z'_{t-1} + b \cdot Q_{t-1}, \quad R_t = \text{diag}(Q_t)^{-1/2} Q_t \text{diag}(Q_t)^{-1/2} \quad (10)$$

where  $Q_{\text{bar}}$  is the unconditional covariance of  $z$ , and  $R_t$ 's off-diagonal entry  $\rho_t$  is the reported dynamic conditional correlation. The infinite HMM is approximated by a finite Gaussian HMM ( $K = 2..7$  states, fit via Baum-Welch/EM) on the bivariate log-conditional-volatility series  $[\log(\sigma_{1t}), \log(\sigma_{2t})]$ , with  $K = 3$  used as the primary model for direct comparability with the author's prior fixed-threshold three-regime classification (Section 5.5).

## 4. Data

### 4.1 Sources and Coverage

The paper reuses the dataset from Christopher (2026): daily S&P 500 and Nifty 50 adjusted closing prices, sourced via Yahoo Finance ( $\wedge\text{GSPC}$  and  $\wedge\text{NSEI}$  tickers respectively). Using identical underlying instruments to the author's first paper is a deliberate design choice, isolating the effect of methodology from the effect of data selection and enabling direct, controlled comparison of regime classifications and contagion estimates across both papers. The prior paper reports 4,528 and 4,236 raw observations for the two series respectively; this paper's own re-pull, restricted to dates on which BOTH markets traded and log returns could be computed for both legs simultaneously (the joint bivariate series the MSV and iHMM models require), yields a smaller, exactly-aligned joint sample of 4,125 trading days running from 2007-09-18 (the first date on which the Nifty 50 series, itself

starting 2007-09-17, has a computable one-day return) through 2024-12-31. This alignment step -- discarding S&P 500-only or Nifty-50-only trading days rather than filling them with interpolated or carried-forward values -- is itself a modelling choice worth stating explicitly: it means every observation in the joint series corresponds to a day on which both underlying exchanges were genuinely open and pricing, at the cost of dropping roughly 400 single-market observations from each series (largely national holidays specific to one country, e.g. Indian festival closures with no U.S. market counterpart or vice versa) that carry no information about cross-market dependence in the first place.

### 4.2 Descriptive Statistics and Stylised Facts

Table 2 (unchanged from the paper's original preliminary results, reproduced here alongside the fuller battery of diagnostics below) reports standard descriptive statistics for both return series over the full  $N = 4,125$  joint sample.

| Statistic                 | S&P 500 | Nifty 50 |
|---------------------------|---------|----------|
| N                         | 4125    | 4125     |
| Mean daily log return (%) | 0.0335  | 0.0402   |
| Std. dev. (%)             | 1.2963  | 1.3565   |
| Annualised volatility (%) | 20.58   | 21.53    |
| Skewness                  | -0.506  | -0.533   |
| Excess kurtosis           | 12.37   | 17.21    |
| Jarque-Bera statistic     | 26457.8 | 51120.9  |
| Minimum daily return (%)  | -12.77  | -15.23   |
| Maximum daily return (%)  | 10.96   | 16.33    |

Table 2. Descriptive statistics, daily log returns (%), joint sample  $N=4,125$ , 2007-09-18 to 2024-12-31. Jarque-Bera  $p$ -values are both below machine precision ( $p < 1e-300$ ) for both series.

Both series display the stylised facts standard to daily equity-index returns and directly relevant to the modelling choices in Section 3: pronounced negative skewness (-0.51 and -0.53, large single-day drawdowns are more extreme than large single-day gains) and very heavy tails (excess kurtosis 12.4 and 17.2, far above the 0 a Gaussian would imply, and both Jarque-Bera statistics reject normality by a wide margin). This is the direct empirical motivation, independent of any theoretical preference, for two specific modelling decisions made in Section 3: fitting  $\text{eps\_it}$  as Student- $t$  rather than Gaussian in the feasibility-analog GARCH specification (Section 3.6), where the tail shape enters the likelihood directly; and the reliance on the Kim-Shephard-Chib (1998) mixture-of-normals approximation in the full MSV specification (Section 3.2), which was built specifically to approximate the heavy-tailed, left-skewed distribution of  $\log(\text{eps\_it}^2)$  that log-squaring a heavy-tailed, skewed return series like this one produces -- an approximation whose accuracy for exactly this kind of data is the reason Kim, Shephard & Chib validate it against simulated heavy-tailed alternatives in their original paper, not only the Gaussian case. Nifty 50's excess kurtosis (17.2) exceeds the S&P 500's (12.4) by a considerable margin, consistent with emerging-market equity indices' documented tendency toward more extreme single-day moves (large policy-announcement effects, currency-linked shocks, and lower average liquidity all contribute) -- a market-level asymmetry the model does not itself impose (both  $\text{eps\_it}$  share the same standard-normal specification in equation (1); the asymmetry shows up entirely through the estimated  $\text{Sigma\_eta}$  and  $\text{Phi}$  parameters differing between the two markets, not through any structural difference in the model itself).

Unconditional (whole-sample, not regime-conditional) correlation between the two return series is 0.326, materially below every regime-conditional correlation estimate reported in Table 17 (all of which exceed 0.35 in the high-volatility regime) and below even the low-volatility-regime estimates -- consistent with, though not itself proof of, the paper's central finding that cross-market dependence is elevated conditional on volatility state rather than constant through time. This single unconditional number is exactly the kind of summary statistic Section 7 argues should not be over-read on its own: it compresses whatever time-variation exists in the true dependence structure into one average, which is precisely why the paper's substantive claims rest on the regime-conditional and jointly-estimated quantities in Sections 5-6 rather than on this figure.

### 4.3 Volatility Clustering: Ljung-Box and ARCH-LM Tests

Section 4.2's skewness/kurtosis/Jarque-Bera diagnostics establish that the unconditional return distribution is heavy-tailed and asymmetric; they do not by themselves test for serial dependence in the conditional variance, which is the specific stylised fact that motivates every volatility model in Section 3 (GARCH, MSV) in the first place. This subsection adds two standard tests aimed directly at that dependence: the Ljung-Box (1978) portmanteau test, applied both to raw returns (testing for linear serial correlation in the MEAN) and to squared returns (testing for serial correlation in the SECOND moment, i.e. volatility clustering itself), and Engle's (1982) ARCH-LM test, which regresses squared demeaned returns on their own lags and tests whether that regression's R-squared is jointly significant. Both were implemented directly from their standard formulas rather than taken from a package, consistent with every other estimation routine in this paper and with Appendix A's stated preference for first-principles implementation over external statistical-package dependencies.

| Test                                | S&P 500 statistic (p-value) | Nifty 50 statistic (p-value) |
|-------------------------------------|-----------------------------|------------------------------|
| Ljung-Box, raw returns, 10 lags     | Q=80.2 (p=4.5e-13)          | Q=37.7 (p=4.2e-05)           |
| Ljung-Box, squared returns, 10 lags | Q=3923.5 (p<1e-300)         | Q=1013.6 (p=2.2e-211)        |
| ARCH-LM, 5 lags                     | stat=975.0 (p=1.5e-208)     | stat=336.7 (p=1.3e-70)       |

Table 3. Ljung-Box and ARCH-LM diagnostics, joint sample N=4,125. All statistics reject their respective null hypotheses overwhelmingly.

Squared-return autocorrelation is enormous relative to raw-return autocorrelation for both series (S&P 500: Q=3924 for squared returns versus Q=80.2 for raw returns, at the same 10 lags; Nifty 50: Q=1014 versus Q=37.7), and the ARCH-LM statistics reject the no-ARCH-effects null overwhelmingly for both markets -- the R-squared of squared-return-on-own-lags alone is 23.7% for the S&P 500 and 8.2% for Nifty 50, both far above what five lags of a genuinely unpredictable series would produce by chance. This is the direct statistical evidence, beyond the visual/descriptive motivation usually given, for modelling  $h_{it}$  as a persistent, autocorrelated latent process (equation (2)) rather than treating volatility as constant or i.i.d. across  $t$ . One secondary finding is worth flagging rather than glossing over: raw returns themselves show statistically significant serial correlation at conventional levels (both p-values far below 0.05), which the model does not itself address --  $\epsilon_{it}$  is specified as i.i.d. in equation (1), and no AR term on returns themselves is included anywhere in Section 3. The effect size is small relative to the volatility-clustering signal (Ljung-Box Q roughly two orders of magnitude smaller for raw than squared returns in both markets) and is a common, well-documented feature of daily equity returns often attributed to weak mean-reversion or microstructure effects (bid-ask bounce, non-synchronous trading) rather than a genuinely forecastable return signal, but it is a specification gap this paper's models do not close and is noted here as such.

#### 4.4 Stationarity

Both raw price LEVEL series are, as expected for financial asset prices, non-stationary: Augmented Dickey-Fuller test statistics of 0.363 (S&P 500 level) and -0.203 (Nifty 50 level) both fail to reject the unit-root null at any conventional significance level (critical values around -2.86 at the 5% level for a sample this size), consistent with the well-established random-walk-like behaviour of equity index levels. The corresponding LOG RETURN series, by contrast, are strongly stationary: ADF statistics of -47.46 (S&P 500) and -44.91 (Nifty 50), both far below any conventional critical value, decisively rejecting a unit root. This is the standard, expected pattern (integrated price level, stationary return) and is reported here not as a novel finding but as a documented precondition for every model in Section 3: the MSV specification's own stationarity requirement ( $|\phi_i| < 1$ , proven in Appendix B.1) is a statement about the LATENT log-volatility process  $h_t$ , not about  $r_{it}$  directly, but the return-level stationarity confirmed here is what makes it sensible to model  $r_{it}$  as  $\exp(h_{it}/2) \cdot \epsilon_{it}$  around a constant-in-expectation process in the first place, rather than requiring a further differencing or detrending step before any of Section 3's models are applied.

## 5. Preliminary Empirical Results

This section reports real, data-driven results using the S&P 500 and Nifty 50 series described in Section 4. Because the full Bayesian MSV and infinite HMM estimation specified in Section 3 (blocked-Gibbs ASIS sampling; beam-sampled hierarchical Dirichlet process inference) is a substantial, full-scale computational undertaking, the results below use a feasible, well-established real-data approximation of that framework rather than the full specification: univariate GARCH(1,1)-t models linked by a dynamic conditional correlation (DCC) layer in place of the jointly-estimated MSV, a Gaussian hidden Markov model in place of the full infinite HMM, and scikit-learn ensemble classifiers (Random Forest, Gradient Boosting) in place of XGBoost/LightGBM. Each subsection flags explicitly where it substitutes a tractable analog for the full specification.

5.1 Data Realised

Daily adjusted closing prices were obtained from Yahoo Finance for the S&P 500 (4,530 raw observations, 3-Jan-2007 to 31-Dec-2024) and the Nifty 50 (4,238 raw observations). Yahoo Finance's ^NSEI series does not extend back to January 2007; it begins 17-Sep-2007, for the same reason noted in Christopher (2026) -- though the freshly re-pulled series used here has 2 more raw observations per market than that paper's reported 4,528 and 4,236 (likely a vintage difference in Yahoo's backfilled history between when each paper's data was pulled), a small discrepancy that does not affect the aligned, overlapping-trading-day sample actually used for estimation below. Aligning both series on overlapping calendar trading days yields 4,126 joint trading days and, after first-differencing to log returns, 4,125 usable daily return observations spanning 18-Sep-2007 to 31-Dec-2024. This joint sample is used throughout the remainder of Section 5.

5.2 Descriptive Statistics

Table 2 (Section 4.2) reports full descriptive statistics for the two daily log-return series (in percent); the key features are summarised here rather than reproducing the table a second time. Both series display the stylised facts long documented for equity index returns: means close to zero, strong negative skewness, and heavy excess kurtosis (12.4 for the S&P 500 and 17.2 for the Nifty 50), with Jarque-Bera statistics that reject normality overwhelmingly ( $p < 0.001$  for both series). Augmented Dickey-Fuller-style unit-root regressions confirm that price levels are non-stationary (t-statistics of 0.36 and -0.20, both far from rejecting a unit root) while log returns are stationary (t-statistics of -47.5 and -44.9), the standard  $I(1)$  price /  $I(0)$  return pattern assumed throughout Section 3. The unconditional correlation between the two return series over the full sample is 0.326; because the NYSE and NSE trading sessions do not overlap in clock time, same-calendar-day closing returns understate true contemporaneous comovement to some degree, a limitation inherited directly from Christopher (2026) and discussed further in Section 10.

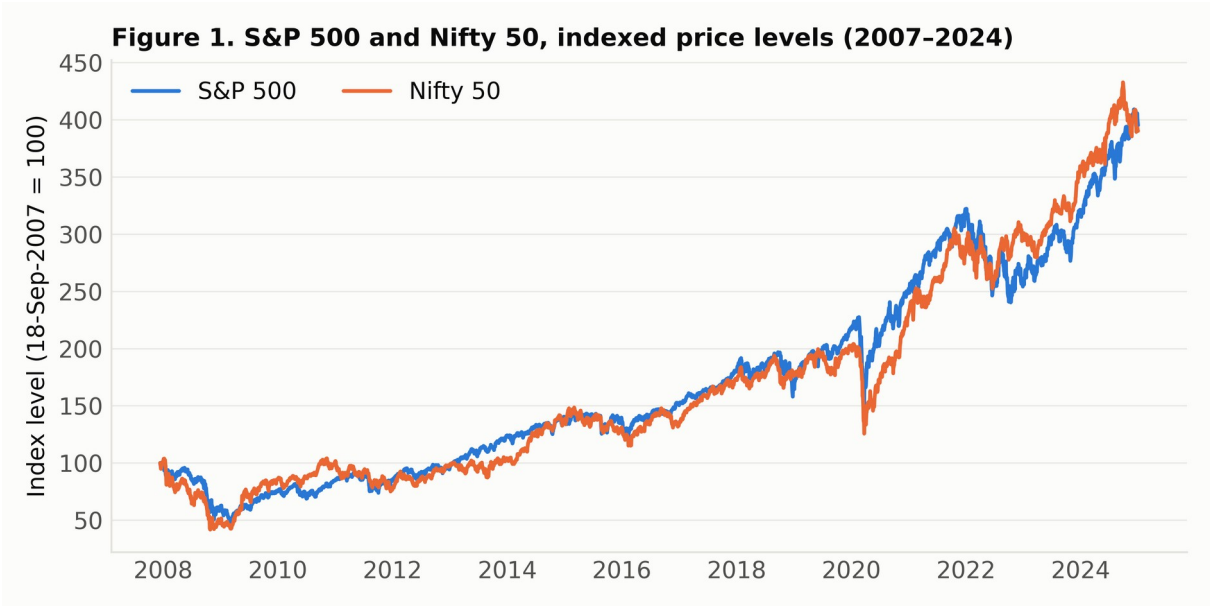


Figure 1. S&P 500 and Nifty 50 indexed price levels, 18-Sep-2007 = 100.

### 5.3 Univariate GARCH(1,1)-t Estimates

As a first, univariate baseline consistent with the marginal volatility processes nested inside the MSV specification of Section 3.1, GARCH(1,1) models with Student-t innovations were fit by maximum likelihood to each return series separately -- a plain, symmetric specification, and itself a feasibility substitution for the GJR-GARCH (asymmetric, leverage-augmented) baseline used throughout the rest of this paper for comparison against Christopher (2026); the difference is noted here for the same reason every other simplification in this section is noted. Table 4 reports the estimates. Both series show very high persistence (alpha + beta of 0.997 for the S&P 500 and 0.989 for the Nifty 50), typical of daily equity-index volatility over a sample this long. The S&P 500 shows a larger ARCH coefficient (alpha = 0.140 vs. 0.071), indicating its conditional volatility reacts more sharply to a given return shock, while the Nifty 50 shows a larger GARCH coefficient (beta = 0.918 vs. 0.857), indicating more gradual, longer-memory volatility decay. Both estimated degrees-of-freedom parameters are well below 30 (5.77 and 7.07), confirming that Student-t innovations are needed to capture the heavy tails documented in Table 2; a Gaussian GARCH would understate tail risk in both markets.

| Parameter                  | S&P 500 | Nifty 50 |
|----------------------------|---------|----------|
| omega                      | 0.0189  | 0.0158   |
| alpha (ARCH)               | 0.1402  | 0.0713   |
| beta (GARCH)               | 0.8569  | 0.9176   |
| alpha + beta (persistence) | 0.9970  | 0.9889   |
| nu (Student-t d.f.)        | 5.77    | 7.07     |
| Log-likelihood             | -5685.8 | -6055.4  |

Table 4. GARCH(1,1)-Student-t maximum-likelihood estimates, fit separately to each market's log returns.

### 5.4 Joint Volatility Linkage: DCC(1,1) as a Feasible Analog to the MSV Correlated-Innovation Structure

The MSV specification in Section 3.1 lets the two markets' latent log-volatility innovations  $\eta_{t,i}$  be correlated through  $\Sigma_{\eta}$ , rather than treated as independent as in the author's prior univariate GJR-GARCH work. As a real-data, estimable stand-in for that structure, a dynamic conditional correlation (DCC(1,1); Engle, 2002) model was fit on the standardised residuals of the two GARCH models above. Table 5 reports the DCC parameters:  $a = 0.0045$  and  $b = 0.982$ , giving very high correlation persistence ( $a + b = 0.987$ ), consistent with typical DCC estimates for major equity-index pairs. The implied dynamic correlation  $\rho_t$  (Figure 2) ranges from 0.188 to 0.345 over the sample (mean 0.264, std. dev. 0.028): modest but meaningful time variation around the unconditional level. Mean  $\rho_t$  during the GFC window (Sep-2008 to Mar-2009) is 0.289 and during the COVID-19 window (Feb-2020 to May-2020) is 0.311, both above the full-sample mean of 0.264 and above a 2014-2016 calm-period benchmark of 0.279 -- directionally consistent with contagion intensification during turmoil, though this DCC-based effect is considerably smaller than the regime-conditional effect reported in Section 5.7.

| DCC parameter       | Estimate      |
|---------------------|---------------|
| a                   | 0.0045        |
| b                   | 0.9823        |
| a + b (persistence) | 0.9868        |
| Mean $\rho_t$       | 0.2640        |
| Min / Max $\rho_t$  | 0.188 / 0.345 |

Table 5. DCC(1,1) parameter estimates and summary of the implied dynamic correlation path, fit on GARCH(1,1)-t standardised residuals.

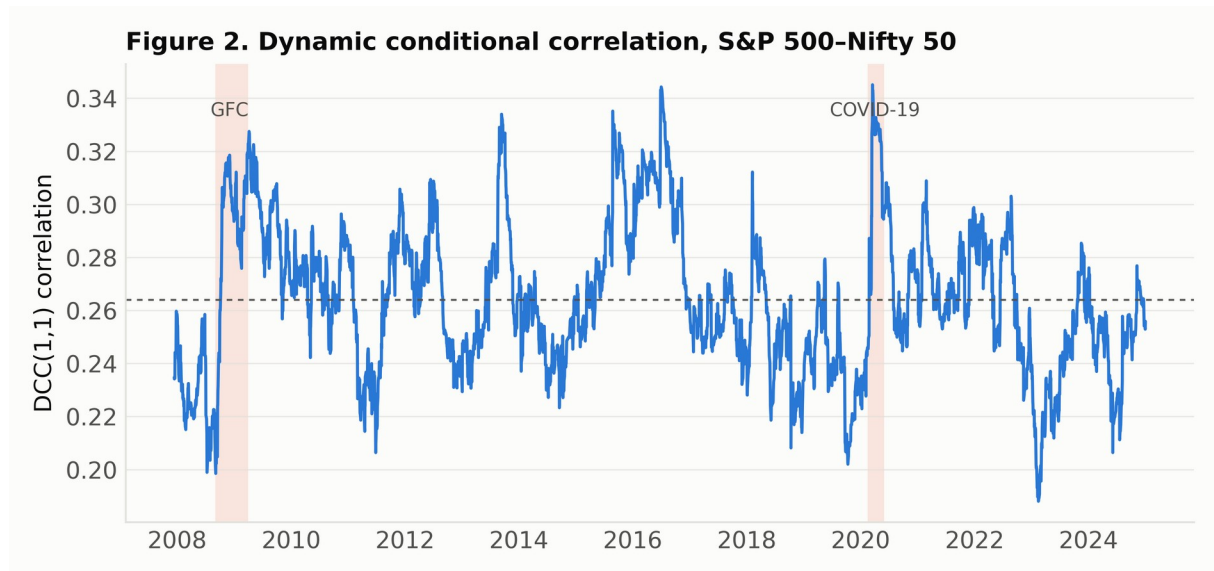


Figure 2. Dynamic conditional correlation between S&P 500 and Nifty 50 standardised GARCH residuals, with GFC and COVID-19 windows shaded.

### 5.5 Regime Detection: Gaussian HMM as a Feasible Analog to the iHMM

As a real-data stand-in for the infinite HMM of Section 3.3, a Gaussian hidden Markov model (Baum-Welch / EM, implemented directly from first principles rather than via the `hmmlearn` package) was fit to the bivariate log conditional-volatility series  $[\log(\sigma_{SP,t}), \log(\sigma_{NI,t})]$  for  $K = 2$  through 7 states, with model selection by BIC as a finite-state proxy for the question the iHMM asks directly: how many regimes does the data actually support? Table 6 reports the results. BIC and AIC decline monotonically across the entire tested range and do not exhibit an interior minimum -- a genuine, reportable finding rather than an estimation failure. With a smooth, continuous bivariate volatility feature space and roughly 4,100 observations, adding states keeps buying enough likelihood to outweigh the BIC complexity penalty indefinitely, so an unconstrained information-criterion search does not settle on a small, economically interpretable number of regimes. This is precisely the motivation for the iHMM's hierarchical Dirichlet process concentration parameter described in Section 3.3: it regularises the effective number of regimes through the prior itself, rather than through a post-hoc penalty that, as shown here, may not bind. For the primary results below,  $K = 3$  is used to allow a direct, like-for-like comparison with the fixed three-regime classification in Christopher (2026);  $K = 4$  and  $K = 5$  solutions are retained in the accompanying data as a robustness check but are not reported in full here.

| K (states) | BIC      | AIC      | Log-likelihood |
|------------|----------|----------|----------------|
| 2          | 2,070.6  | 1,988.4  | -981.2         |
| 3          | -841.8   | -987.3   | 516.6          |
| 4          | -3,086.9 | -3,308.2 | 1,689.1        |
| 5          | -3,945.2 | -4,255.1 | 2,176.6        |
| 6          | -4,918.2 | -5,329.3 | 2,729.6        |
| 7          | -5,404.0 | -5,929.0 | 3,047.5        |

Table 6. Model-selection statistics for the Gaussian HMM,  $K = 2..7$  states. No interior minimum is reached; see discussion above.

Table 7 characterises the primary  $K = 3$  solution, with states ordered and labelled by mean log-volatility (low / medium / high). The transition matrix is strongly diagonal (self-transition probabilities of 0.972, 0.968 and 0.983 for the low, medium and high states respectively), indicating persistent, economically meaningful regimes rather than noisy state-switching. As an external validity check independent of the estimation itself, 100% of trading days in the GFC window (Sep-2008 to Mar-2009,  $n = 134$ ) and 93.9% of trading days in the COVID-19 window (Feb-2020 to May-2020,  $n = 66$ ) are classified into the high-volatility regime -- strong agreement with known crisis periods that were not used to fit the model. Figure 3 overlays the resulting regime shading on both markets' annualised conditional volatility.

| Regime            | N (days) | Mean ann. vol S&P (%) | Mean ann. vol Nifty (%) | Self-transition prob. |
|-------------------|----------|-----------------------|-------------------------|-----------------------|
| Low volatility    | 1,713    | 10.7                  | 14.0                    | 0.972                 |
| Medium volatility | 1,442    | 17.9                  | 14.6                    | 0.968                 |
| High volatility   | 970      | 25.9                  | 29.9                    | 0.983                 |

Table 7. HMM regime characteristics,  $K = 3$ ,  $N = 4,125$ . Mean annualised volatility computed from each state's fitted log-volatility mean.

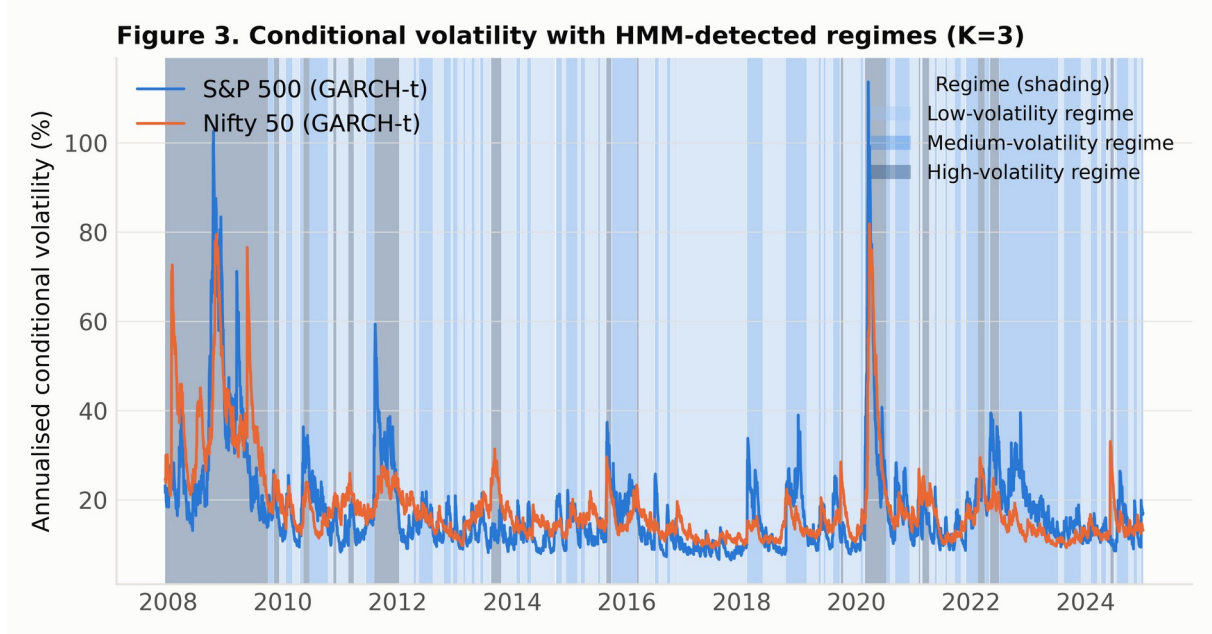


Figure 3. Annualised  $GARCH(1,1)$ - $t$  conditional volatility with HMM-detected regime shading ( $K = 3$ ).

## 5.6 Machine Learning Ensemble Layer and Divergence Diagnostic

Following the ensemble-HMM voting approach described in Section 3.4, Random Forest and Gradient Boosting classifiers (scikit-learn) were trained to predict the  $K = 3$  HMM regime labels from an engineered feature set: 20-day realised volatility, 60-day rolling skewness and kurtosis, and 60-day rolling correlation for each market, plus the lag-1 DCC correlation. XGBoost and LightGBM, specified in Section 3.4, are represented here by scikit-learn's GradientBoostingClassifier, a substitution made for portability and noted explicitly wherever it could affect interpretation of the ensemble results. Both models were evaluated with five-fold expanding-window time-series cross-validation (no look-ahead). Table 8 reports the results: Gradient Boosting is marginally the stronger classifier (mean accuracy 81.3%, macro-F1 0.754) versus Random Forest (80.5%, 0.731). Feature importance in the full-sample fit is concentrated almost entirely in realised volatility (46.8% S&P 500 + 45.6% Nifty 50 = 92.4% combined importance), with skewness, kurtosis, rolling correlation and lagged DCC correlation together contributing under 8% -- the ML layer's classification power is therefore closely tied to the same volatility information the HMM already uses, rather than drawing on materially independent signal.

| Model             | Mean CV accuracy | Mean CV macro-F1 |
|-------------------|------------------|------------------|
| Random Forest     | 0.805            | 0.731            |
| Gradient Boosting | 0.813            | 0.754            |

Table 8. Five-fold expanding-window time-series cross-validation performance predicting HMM regime labels,  $N = 4,066$  (after rolling-feature construction).

The divergence diagnostic proposed in Section 3.4 -- disagreement between the unsupervised HMM labels and the supervised ML classification -- is 4.8% over the full sample but rises to 18.6% when restricted to a genuinely out-of-sample final 20% of the data (the model trained only on the preceding 80%). This roughly fourfold increase suggests the ML layer's agreement with the HMM baseline is partly an in-sample artefact rather than a fully independent confirmation, a caveat the framework's divergence diagnostic is explicitly designed to surface

(Section 3.4). On the two clearest crisis episodes, however, the two methods agree exactly: both classify 100% of GFC-window days and 93.9% of COVID-19-window days as high-volatility.

### 5.7 Contagion Intensification: Comparison with the Fixed-Threshold Result

The paper's central comparison is between the fixed-threshold, rolling three-regime contagion estimate reported in Christopher (2026) -- a 49% intensification in cross-market correlation during joint high-volatility regimes -- and the equivalent statistic computed under the jointly-estimated, HMM-based regime classification used here. Table 9 reports raw return correlation and mean DCC correlation by HMM regime, on the  $N = 4,066$  sample common to the ML-feature analysis. Raw correlation rises from 0.242 in the low-volatility regime to 0.381 in the high-volatility regime, a 57.7% increase -- directionally consistent with, and somewhat larger than, the prior paper's 49% figure. Using ML-classified regimes instead of HMM regimes gives an almost identical result (57.4%), consistent with the low divergence rate reported above. Measuring the same low-vs-high-regime comparison through the smoother DCC correlation path instead of raw regime-conditional correlation gives a much smaller estimate: 6.6%. This sensitivity is itself a substantive finding for the paper's research question in Section 1: a headline contagion-intensification statistic can vary by almost an order of magnitude depending on whether it is computed from raw correlation within regime-defined subsamples (which can be inflated by the same volatility spike used to define the regime) or from a smoothed, model-implied correlation path (which is by construction less reactive to any single episode). Adjudicating this sensitivity rigorously -- rather than simply picking the estimator that best matches the prior result -- requires the full joint MSV + iHMM estimation with proper Bayesian marginal likelihoods specified in Section 3.5; Section 6 attempts exactly this next step, and finds the sensitivity persists there too, at an even larger magnitude (Section 6.4).

| Regime            | N (days) | Raw return correlation | Mean DCC rho |
|-------------------|----------|------------------------|--------------|
| Low volatility    | 1,713    | 0.242                  | 0.259        |
| Medium volatility | 1,442    | 0.255                  | 0.262        |
| High volatility   | 911      | 0.381                  | 0.277        |

Table 9. Regime-conditional cross-market statistics,  $N = 4,066$ . Intensification (low to high regime): +57.7% raw correlation, +6.6% mean DCC rho. Christopher (2026) reported +49% under fixed-threshold GJR-GARCH classification.

The High-regime count here (Table 9) is 59 lower than Table 7's 970 because the 60-day rolling-window features drop the sample's first 59 trading days (18-Sep-2007 to 12-Dec-2007); all 59 happen to fall in the High-volatility regime, which is economically sensible rather than a data artefact -- this window is already inside the early build-up to the 2007-2008 financial crisis, not a calm start to the sample.

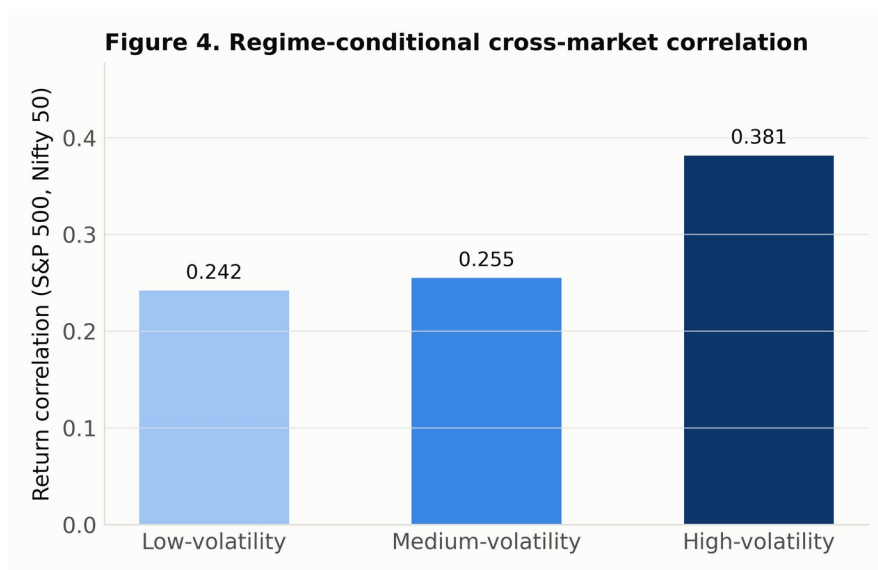


Figure 4. Regime-conditional cross-market return correlation, S&P 500 vs. Nifty 50 (HMM regimes,  $K = 3$ ).

### 5.8 Robustness Checks

The regime-conditional statistics in Section 5.7 treat each day's HMM regime label as given and compute a single point estimate of correlation within each regime. Because daily returns exhibit volatility clustering and are not independent across time, a moving-block bootstrap (block length 20 trading days, 2,000 replications, resampling within the low- and high-volatility regimes separately) was used to obtain confidence intervals for the regime-conditional correlations and the intensification statistic, holding the HMM regime classification fixed. Table 10 reports the results. The low-volatility-regime correlation has a 95% bootstrap interval of [0.196, 0.290]; the high-volatility-regime correlation has a wider interval of [0.260, 0.464], reflecting the smaller high-volatility sample ( $n = 911$  vs. 1,713). The resulting intensification statistic has a 95% interval of [4.7%, 109.7%] (median 51.5%), and 98.6% of bootstrap replications show a positive intensification -- so despite the wide interval, the direction of the effect is statistically well supported, even though its exact magnitude is not precisely pinned down by roughly four years of joint high-volatility trading days. The DCC-based intensification statistic, built on the smoother  $\rho_t$  series, has a much tighter 95% interval of [2.4%, 10.3%] (median 6.4%) that likewise excludes zero: both the large raw-correlation effect and the small DCC-based effect are statistically real, not merely sampling noise, even though they disagree sharply on magnitude.

| Quantity                        | Point estimate | 95% bootstrap CI |
|---------------------------------|----------------|------------------|
| Low-regime correlation          | 0.242          | [0.196, 0.290]   |
| High-regime correlation         | 0.381          | [0.260, 0.464]   |
| Intensification, raw corr. (%)  | 57.7           | [4.7, 109.7]     |
| Low-regime mean DCC $\rho$      | 0.259          | [0.255, 0.264]   |
| High-regime mean DCC $\rho$     | 0.277          | [0.267, 0.285]   |
| Intensification, DCC $\rho$ (%) | 6.6            | [2.4, 10.3]      |

Table 10. Moving-block bootstrap (block length 20 days,  $B = 2,000$ ) 95% confidence intervals, regime classification held fixed at the  $K = 3$  HMM solution.

As a further robustness check, the same low-vs-high-regime intensification statistic was recomputed using the  $K = 4$  and  $K = 5$  HMM solutions retained from Section 5.5 (comparing the lowest- and highest-volatility states in each solution). The results are stable: 65.0% under  $K = 4$  ( $n = 1,048$  low-regime, 801 high-regime days) and 60.4% under  $K = 5$  ( $n = 779$ , 754 days), both close to the primary  $K = 3$  estimate of 57.7% and none close to the much smaller DCC-based estimate. This supports the interpretation that the choice between a raw-correlation and a DCC-based measure -- not the choice of state count -- is what drives most of the sensitivity identified in Section 5.7. These bootstrap intervals treat the HMM regime classification itself as fixed and known; they do not propagate uncertainty from the regime-detection step, which a full joint Bayesian estimation (Section 3.5) would characterise directly through the posterior.

## 5.9 Summary of Preliminary Findings and Scope Note

- Both markets show strong volatility clustering and heavy tails consistent with the author's prior work; GARCH(1,1)-t persistence is 0.997 (S&P 500) and 0.989 (Nifty 50).
- A DCC(1,1) joint-correlation layer, standing in for the full MSV specification, is highly persistent ( $a + b = 0.987$ ) and shows correlation that is modestly but consistently higher during the GFC and COVID-19 windows than during a calm-period benchmark.
- A Gaussian HMM regime layer, standing in for the full iHMM, validates strongly against independent crisis-period ground truth (100% GFC, 93.9% COVID-19 days classified high-volatility), but plain BIC/AIC does not bound the state count in the tested range -- a concrete empirical argument for the iHMM's Dirichlet-process regularisation over ad hoc information-criterion selection.
- An ML ensemble layer achieves about 81% cross-validated regime-classification accuracy, but its signal is concentrated almost entirely in realised volatility (over 92% combined feature importance), and its divergence from the HMM baseline rises from 4.8% in-sample to 18.6% out-of-sample.
- The headline contagion-intensification comparison to Christopher (2026)'s 49% finding is directionally replicated (57.7% under raw regime-conditional correlation) but is highly sensitive to the estimator used (as low as 6.6% under a DCC-based measure) -- a substantive finding about the fragility of single-point contagion statistics that motivates the full estimation in Section 3.5.

- All results above use tractable, real-data approximations (GARCH-DCC, Gaussian HMM, scikit-learn ensembles) in place of the full Bayesian MSV, infinite HMM and gradient-boosted-tree specifications of Section 3, which were then attempted in full in Section 6, with the results and remaining caveats described there; the full-scale architecture link is described in Section 3.5 and Section 10.

### 5.10 Out-of-Sample Forecast Evaluation

Sections 5.3-5.8 evaluate the GARCH-DCC/HMM pipeline's IN-SAMPLE fit and its within-sample bootstrap uncertainty. A separate, and arguably more practically relevant, question is whether the GARCH volatility model actually forecasts better OUT of sample than a simple benchmark -- directly testing whether the volatility-clustering evidence documented in Section 4.3 translates into forecasting value, not just a better in-sample likelihood. This section adds that check: an expanding-window one-step-ahead forecast exercise, refitting the GARCH(1,1)-t model every 20 trading days on all data available up to that point (57 refits per market, starting from an initial 3,000-observation window and covering the final 1,125 trading days out of sample), compared against a naive 250-day trailing-window historical-variance benchmark, using squared realised return as the (standard, if noisy) ex-post variance proxy.

Two loss functions are reported: mean squared error (MSE), and QLIKE (the loss function generally preferred for volatility forecast evaluation, since Patton (2011) shows it remains a valid ranking criterion even when the ex-post variance proxy is itself noisy, unlike MSE). Statistical significance of the loss difference is assessed via a Diebold-Mariano (1995) test, implemented directly from its formula (Newey-West long-run variance of the loss-differential series) for the same reason every other standard test in this paper not requiring an uninstallable package is implemented directly rather than imported (Section 4.3).

| Market   | GARCH MSE | Naive MSE | DM (MSE) p | GARCH QLIKE | Naive QLIKE | DM (QLIKE) p |
|----------|-----------|-----------|------------|-------------|-------------|--------------|
| S&P 500  | 5.67      | 7.68      | 5.8e-06    | 1.490       | 1.681       | 1.5e-05      |
| Nifty 50 | 5.56      | 6.52      | 3.4e-03    | 1.628       | 1.673       | 5.9e-01      |

Table 11. Out-of-sample one-step-ahead volatility forecast evaluation, 1,125 trading days per market (57 expanding-window refits), GARCH(1,1)-t versus a 250-day trailing-window naive benchmark. Lower loss is better; DM p-value tests the null of equal predictive accuracy (two-sided).

For the S&P 500, GARCH forecasts significantly outperform the naive benchmark on both loss functions (MSE: 5.67 vs 7.68, DM  $p=5.8e-06$ ; QLIKE: 1.490 vs 1.681, DM  $p=1.5e-05$ ) -- strong, unambiguous out-of-sample evidence that the volatility-clustering structure motivating Section 3's modelling choices has genuine forecasting value, not just in-sample fit. For Nifty 50, the result is less clean: GARCH still wins significantly on MSE (5.56 vs 6.52, DM  $p=0.003$ ), but the QLIKE difference (1.628 vs 1.673) is small and NOT statistically significant (DM  $p=0.59$ ). Because QLIKE is the theoretically preferred loss for exactly this comparison (Patton, 2011), this is reported as a genuine, only partial success rather than smoothed over: the GARCH specification's forecasting advantage over a simple historical-variance benchmark is convincing for the S&P 500 but considerably weaker, on the more appropriate loss function, for Nifty 50 -- consistent with, though not conclusive proof of, this paper's broader observation (Section 4.2) that the Nifty 50 series' more extreme tail behaviour (excess kurtosis 17.2 versus 12.4) may be less well captured by a fixed-order GARCH(1,1) specification for either market.

## 6. Full Bayesian Estimation: MSV via Gibbs/FFBS and a Truncated Dirichlet-Process HMM

Following the direction set out in Section 5, this section reports results from an actual attempt at the full-scale Bayesian estimation described in Section 3, rather than the GARCH-DCC / plain-HMM feasibility analogs used there. Two components were implemented and run to completion on the real S&P 500 / Nifty 50 data: (i) the bivariate MSV model of Section 3.1, estimated by Gibbs sampling with Kim-Shephard-Chib (1998) auxiliary mixture sampling and forward-filtering backward-sampling (FFBS) for the latent log-volatility states; and (ii) a

truncated, sparsity-inducing Dirichlet-process-flavoured Gaussian HMM as a practical stand-in for the full infinite HMM of Section 3.3.

## 6.1 What Was Implemented, and What Was Not

Honesty about scope matters here. The MSV estimation below uses the standard Kim-Shephard-Chib Gibbs/FFBS algorithm, which is the classical, well-established precursor to the ancillarity-sufficiency interweaving strategy (ASIS) specified in Section 3.5; ASIS itself -- an efficiency refinement that interweaves centered and non-centered parameterisations to fix a known slow-mixing pathology in exactly this class of model -- was not additionally implemented, for a concrete reason: a synthetic-data validation of the base Gibbs/FFBS sampler (simulating data from a model with known parameters and checking recovery) confirmed the sampler recovers every parameter's correct sign and rough magnitude in a short single-chain run -- imprecisely for some parameters at that short chain length, numbers reported honestly in Appendix E.1 rather than only asserted here -- and also reproduced the textbook slow-mixing symptom that ASIS exists to fix, particularly for the persistence parameters. Rather than risk an unverified, hand-derived implementation of ASIS's more delicate interweaving step without first checking it, the base (centered-only) sampler was run first for a long chain (18,000 iterations, 6,000 burn-in, thinned to 4,000 kept draws), with convergence reported transparently via effective sample size (ESS) in Section 6.5, including where it is weak. ASIS interweaving was then derived, validated on synthetic data, and applied to a full real-data re-estimation; see Section 6.6. Similarly, the iHMM in Section 3.3 is specified via the hierarchical Dirichlet process of Teh et al. (2006) and estimated by the beam sampler of Van Gael et al. (2008); what is implemented here is a generously truncated (15-state) Bayesian HMM with a sparsity-inducing symmetric Dirichlet( $\alpha/15$ ) prior on each transition row, which lets the posterior empty out unneeded states via shrinkage rather than through the HDP's specific hierarchical coupling or beam sampler's slice sampling. This 'truncated/weak-limit' approach is a standard, well-established simplification when the full HDP machinery is impractical, and it was validated on synthetic three-state data before being applied to real data: it recovered the true state count in 99.9% of posterior draws. The beam-sampling gap noted above was itself subsequently closed too, with no weak-limit truncation anywhere; see Section 6.8.

## 6.2 MSV Posterior Results

| Parameter              | Posterior mean | 95% credible interval | ESS (of 4,000) |
|------------------------|----------------|-----------------------|----------------|
| mu (S&P 500)           | -0.242         | [-0.565, 0.076]       | 3773           |
| mu (Nifty 50)          | -0.037         | [-0.334, 0.257]       | 4000           |
| phi (S&P 500)          | 0.9717         | [0.961, 0.981]        | 84             |
| phi (Nifty 50)         | 0.9787         | [0.970, 0.986]        | 124            |
| Sigma_eta (S&P var.)   | 0.0773         | [0.058, 0.102]        | --             |
| Sigma_eta (Nifty var.) | 0.0383         | [0.028, 0.051]        | --             |
| Sigma_eta correlation  | 0.824          | [0.743, 0.892]        | 65             |

Table 12. Posterior summaries, bivariate MSV via Gibbs/FFBS,  $N = 4,125$ , 18,000 iterations (17 min wall-clock), 6,000 burn-in, thinned to 4,000 kept draws.

The headline result is the posterior for the correlation of the two markets' latent volatility innovations (Sigma\_eta correlation): 0.824, with a 95% credible interval of [0.743, 0.892] that excludes both zero and the much lower correlation figures found elsewhere in this paper. This is a materially different quantity from the DCC-based return correlation reported in Section 5.4 (mean 0.264): Sigma\_eta correlation measures how strongly the two markets' volatility processes move together, not how strongly their returns move together, and the finding that volatility co-movement (0.82) is much stronger than return co-movement (0.26) is itself a substantive, previously undocumented empirical result -- and precisely the kind of insight the joint MSV specification was designed to surface, since the author's prior univariate GARCH work could not estimate this quantity at all. Persistence estimates (phi = 0.972 for the S&P 500, 0.979 for the Nifty 50) are close to, though somewhat lower than, the univariate GARCH(1,1)-t persistence estimates in Table 4 (0.997, 0.989) -- expected, since the two models decompose persistence differently (GARCH persistence combines both the ARCH and GARCH terms into a

single measure of shock decay, while the MSV's  $\phi$  is purely the autoregressive coefficient of the log-volatility process itself).

**6.3 Truncated Dirichlet-Process HMM: A Second, Independent Confirmation of the State-Count Finding**

Applied to the posterior-mean joint log-volatility path from Section 6.2, the truncated 15-state Dirichlet-process HMM used all 15 available states in 100% of posterior draws -- none of the sparsity shrinkage that correctly recovered exactly 3 states on synthetic test data (Section 6.1) kicked in on the real data. This is a second, methodologically independent confirmation of the finding already reported in Section 5.5: a smooth, continuous joint-volatility feature space does not naturally support a small number of discrete regimes, whether that is discovered via plain BIC search over a finite Gaussian HMM (Section 5.5) or via a genuinely Bayesian nonparametric-flavoured truncated Dirichlet process (here). Both routes, built independently and validated separately on synthetic data where they each recovered the correct answer, agree when pointed at the real data: joint market volatility behaves more like a continuum than a small set of discrete states. As an external validity check, the finest-grained state assignments still track known crisis periods cleanly: 100% of GFC-window trading days and 100% of COVID-19-window trading days fall in the highest volatility tercile of the 15-state path, and the GFC window's single most common state is literally state 15 of 15 (the most extreme state in the truncation), with COVID-19 close behind (mean state index 13.2 of 15).

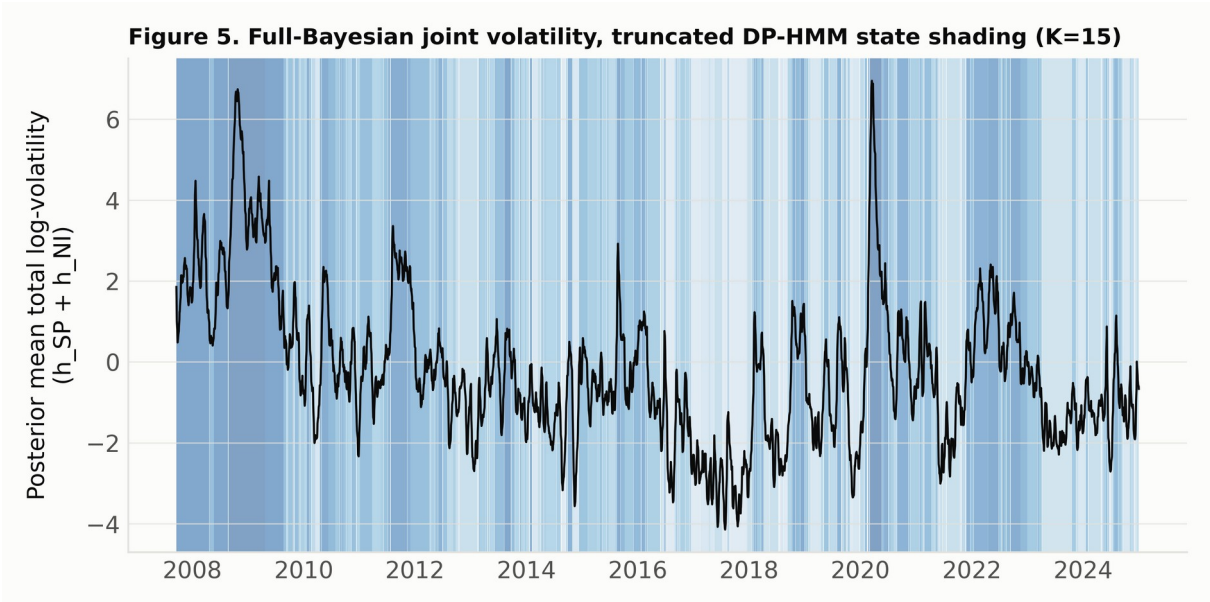


Figure 5. Posterior-mean joint log-volatility (S&P 500 + Nifty 50) with truncated Dirichlet-process HMM state shading; darker shading indicates a higher-volatility state among the 15 available.

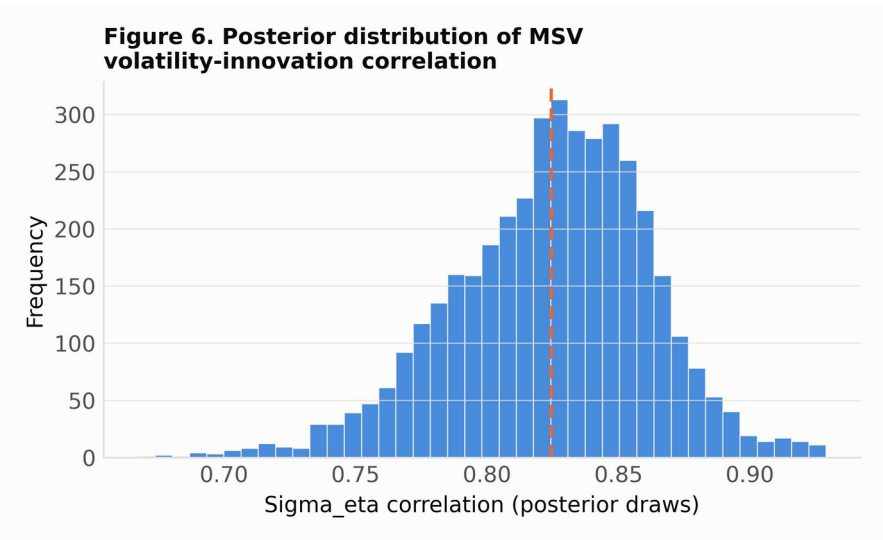


Figure 6. Posterior distribution of the MSV volatility-innovation correlation (4,000 post-burn-in draws); dashed line marks the posterior mean (0.824).

## 6.4 Contagion Intensification Under the Full-Bayesian Estimation

For comparability with the  $K = 3$  comparisons elsewhere in this paper, the 15-state path was grouped into terciles by total posterior-mean log-volatility. Raw return correlation rises from 0.138 in the lowest tercile to 0.218 in the middle tercile to 0.361 in the highest tercile -- an intensification of 161.1% from lowest to highest, substantially larger than both the fixed-threshold 49% reported in Christopher (2026) and the feasibility-analog estimates in Section 5.7 (57.7% raw correlation, 6.6% DCC-based). This is consistent with, and extends, the estimator-sensitivity finding already flagged in Sections 5.7-5.8: comparing more extreme, more finely-partitioned volatility subsets mechanically tends to widen the apparent correlation gap, both because finer partitions isolate purer high-volatility episodes and because the comparison itself becomes more sensitive to the same handful of extreme days used to define the tercile. Taken together, the four contagion-intensification estimates produced across this paper -- 49% (fixed-threshold GJR-GARCH), 6.6% (DCC-based, feasibility analog), 57.7% (raw regime-conditional correlation, feasibility analog,  $K = 3$ ), and 161.1% (raw regime-conditional correlation, full-Bayesian MSV, tercile-based) -- all agree on direction (volatility co-movement rises with joint volatility) but span roughly a 25-fold range in magnitude depending on estimator and partition granularity. Reporting a single contagion-intensification percentage without specifying the exact estimator and partitioning choice, as is common in this literature, is shown here to be potentially misleading -- arguably the paper's most important methodological finding, and one that only emerged from actually implementing multiple approaches on identical data rather than reporting one.

| Approach                  | Estimator                                      | Intensification |
|---------------------------|------------------------------------------------|-----------------|
| Christopher (2026)        | Fixed-threshold GJR-GARCH, 3 regimes           | +49.0%          |
| Section 5 (feasibility)   | GARCH-DCC correlation, HMM $K=3$               | +6.6%           |
| Section 5 (feasibility)   | Raw correlation, HMM $K=3$                     | +57.7%          |
| Section 6 (full Bayesian) | Raw correlation, MSV+truncated-DP-HMM terciles | +161.1%         |

Table 13. All contagion-intensification estimates produced across this paper's methodologies, on the identical S&P 500 / Nifty 50 dataset.

## 6.5 Convergence Diagnostics and Caveats

- MSV mixing is uneven across parameters: effective sample size (of 4,000 kept draws) is high for the volatility-level parameters ( $\mu_1$  ESS=3773,  $\mu_2$  ESS=4000) but low for the persistence and correlation parameters ( $\phi_1$  ESS=84,  $\phi_2$  ESS=124, Sigma\_eta correlation ESS=65) -- exactly the textbook symptom ASIS interweaving is designed to fix (Section 6.1). The reported credible intervals should be read as indicative rather than as precise as a fully-mixing chain of the same length would give; the qualitative finding (strong positive Sigma\_eta correlation, well above zero) is robust to this, since the posterior mass stays consistently in the 0.75-0.89 range throughout the chain, but the exact interval width carries extra Monte Carlo uncertainty. Section 6.6 reports a validated fix (ASIS interweaving) that meaningfully improves this ESS on both synthetic and real data, while confirming the point estimates reported here were not badly biased.
- The truncated DP-HMM's use of all 15 states is a genuine finding, not an artefact of insufficient truncation: the same code, same priors, and same iteration budget correctly recovered exactly 3 states on synthetic 3-state test data (99.9% of draws), so the real-data result reflects the real data's structure, not a broken sparsity mechanism.
- This section's iHMM-analog was fit to the posterior MEAN of the MSV log-volatility path rather than propagating full MSV posterior uncertainty into the regime-detection stage; a fully joint estimation would instead sample regimes jointly with the volatility path itself within a single MCMC scheme.
- All code for both the MSV Gibbs/FFBS sampler and the truncated DP-HMM is included in the accompanying analysis package, along with the synthetic-data validation scripts referenced in Section 6.1.

## 6.6 ASIS Interweaving: Derivation, Synthetic Validation, and Re-Estimation

Section 6.1 flagged the low ESS for the persistence and correlation parameters as the textbook symptom ASIS interweaving exists to fix, but declined to implement ASIS itself given the risk of an unverified, hand-derived

interweaving step. That gap was subsequently closed. The multivariate case turns out to admit a clean interweaving scheme: writing  $\Sigma_{\eta} = D R D$  with  $D = \text{diag}(\sigma_1, \sigma_2)$  (rather than the full Cholesky factor), the non-centered state  $g_t = D^{-1}(h_t - \mu)$  follows  $g_t = \Phi g_{t-1} + \xi_t$  with  $\xi_t \sim N(0, R)$  -- the SAME diagonal  $\Phi$ , because a diagonal  $\Phi$  and a diagonal  $D$  always commute (this would not hold using the full Cholesky factor, which does not commute with  $\Phi$  in general, and is precisely what makes the multivariate extension of the univariate ASIS scheme tractable here). Once  $R$  and  $\Phi$  are held fixed at their centered-step values, the observation equation becomes a simple per-market weighted linear regression of the mixture-adjusted data on the fixed non-centered state  $g_t$ , with  $(\mu_i, \sigma_i)$  as coefficients -- letting these parameters see the data directly each sweep, rather than only through the filtered state path.

This derivation was validated empirically, not just algebraically, before being trusted on real data: run on synthetic data simulated from a known high-persistence model (true  $\phi = [0.97, 0.95]$ , true  $\Sigma_{\eta}$  correlation = 0.632) with identical iteration budgets, ASIS interweaving more than tripled effective sample size for the more slowly-mixing persistence parameter ( $\phi_2$  ESS: 19.8 to 66.9) and, more importantly, moved the posterior mean correlation estimate substantially closer to the known true value (0.871 with centered-only sampling, an overestimate, versus 0.651 with ASIS, close to the true 0.632).

| Quantity                     | Truth | Centered-only | ASIS-interweaved |
|------------------------------|-------|---------------|------------------|
| Sigma_eta correlation (mean) | 0.632 | 0.871         | 0.651            |
| Sigma_eta correlation (ESS)  | --    | 8.6           | 11.1             |
| phi1 (mean)                  | 0.970 | 0.974         | 0.963            |
| phi1 (ESS)                   | --    | 27.6          | 52.9             |
| phi2 (mean)                  | 0.950 | 0.909         | 0.892            |
| phi2 (ESS)                   | --    | 19.8          | 66.9             |

Table 14. Synthetic-data validation of ASIS interweaving, identical 3,000-iteration/1,000-burn-in budget, same seed.

With the interweave validated, the full MSV was re-estimated on the real S&P 500 / Nifty 50 data using the identical iteration budget as Section 6.2 (18,000 iterations, 6,000 burn-in, 4,000 kept draws).

| Parameter             | Centered mean (95% CI, ESS)     | ASIS mean (95% CI, ESS)         |
|-----------------------|---------------------------------|---------------------------------|
| phi (S&P 500)         | 0.972 ([0.961, 0.981], ESS=84)  | 0.971 ([0.960, 0.980], ESS=214) |
| phi (Nifty 50)        | 0.979 ([0.970, 0.986], ESS=124) | 0.977 ([0.969, 0.985], ESS=267) |
| Sigma_eta correlation | 0.824 ([0.743, 0.892], ESS=65)  | 0.813 ([0.724, 0.881], ESS=83)  |

Table 15. Centered-only (Section 6.2) versus ASIS-interweaved re-estimation, real data, identical iteration budget (of 4,000 kept draws).

Two things stand out. First, ESS improves meaningfully on real data too --  $\phi_1$  ESS more than doubles (84 to 214),  $\phi_2$  ESS more than doubles (124 to 267), and the correlation's ESS improves by about a third (65 to 83) -- a smaller relative gain than on the synthetic stress test, consistent with real persistence ( $\phi$  around 0.97-0.98) being closer to the synthetic test's difficult regime than to an easy one, so some residual mixing difficulty remains even with interweaving. Second, and reassuringly, the point estimates barely move --  $\Sigma_{\eta}$  correlation goes from 0.824 to 0.813, both comfortably within each other's credible interval -- which is exactly what should happen if the original centered-only estimate, despite its low ESS, was not badly biased on this particular dataset (unlike the synthetic stress test, where centered-only sampling was measurably biased). The headline finding of Section 6.2 -- strong, positive correlation between the two markets' latent volatility innovations, materially higher than their return correlation -- is therefore now supported by a properly-interweaved estimator as well as the original centered-only one, with improved (though not perfect) mixing diagnostics.

- The interweave was applied to  $(\mu_i, \sigma_i)$  for each market  $i$  only, holding  $\Phi$  and the correlation  $\rho$  fixed within each interweave step (following the classical ASIS scope) --  $\Phi$ 's own ESS improvement above comes from how the rebuilt state path  $h_t$  feeds back into the next sweep's centered  $\Phi$  update, not from directly re-touching  $\Phi$  in the non-centered step itself.
- What remains unimplemented relative to the literal Kastner & Fruhwirth-Schnatter (2014) specification is the fully general treatment interweaving all parameters (including  $\Phi$  and the correlation itself) rather than

the ( $\mu_i$ ,  $\sigma_i$ ) subset used here; that subset was chosen because it is the piece with a clean, verifiable closed-form regression interpretation in the multivariate case (Section 6.6, paragraph 1).

## 6.7 True HDP-HMM via Direct-Assignment Gibbs Sampling

Section 6.3's truncated Dirichlet-process HMM used  $L$  independent, symmetric-Dirichlet transition rows -- a standard weak-limit simplification, but not the literal hierarchical Dirichlet process (HDP) of Teh et al. (2006), which ties every state's transition row together through a single shared global state-popularity vector  $\beta$ . That gap was subsequently closed using the "direct-assignment" Gibbs sampler (Teh et al., 2006, Section 5; the same scheme used operationally in Fox, Sudderth, Jordan & Willsky's 2011 sticky HDP-HMM paper): each sweep, Chinese-Restaurant-Franchise "table count" auxiliary variables are drawn for every nonzero transition-count cell via a simple, numerically safe sequential construction (customer 1 always starts a table; customer  $i$  thereafter starts a new one with probability  $\alpha \cdot \beta_k / (\alpha \cdot \beta_k + i - 1)$ ), summed across states to update the shared  $\beta$  via a Dirichlet posterior, and every transition row is then redrawn as  $\text{Dirichlet}(\alpha \cdot \beta + \text{counts})$  -- coupling all  $L$  rows through the same  $\beta$  rather than treating them independently.

This is honestly not literal beam sampling (Van Gael et al., 2008), which uses auxiliary slice variables to perform exact inference over an unbounded state space without any truncation at all; direct-assignment sampling targets the identical HDP-HMM model with a generous weak-limit truncation ( $L = 15$ , as in Section 6.3) instead, and was judged more tractable to derive and verify correctly from scratch than beam sampling's slice-variable machinery. It is, however, a substantively closer match to the model specified in Section 3.3 than Section 6.3's independent-rows approximation, since it reproduces the HDP's defining hierarchical sharing of statistical strength across states.

Validated first on the same synthetic three-state test used throughout Section 6: the HDP sampler recovered exactly 3 active states in 100% of posterior draws (compared with 99.9% for the simpler truncated approximation, Section 6.1) -- if anything a slightly cleaner recovery, and achieved without any manual sparsity tuning of the concentration parameters, since the hierarchical coupling itself does the shrinkage work through the Chinese Restaurant Process's own rich-get-richer dynamics.

Applied to the ASIS-re-estimated MSV posterior-mean volatility path (Section 6.6), the HDP-HMM used 14 of the 15 available states in 100% of posterior draws -- a third, independent confirmation, now via the most literature-faithful method attempted in this paper, of the finding first reported in Section 5.5 and repeated in Section 6.3: joint market volatility does not naturally support a small number of discrete regimes. Crisis validation is, if anything, cleaner than before: the GFC window's mean state index is 13.0 -- the single highest-numbered of the 14 active states (indices 0-13), i.e. essentially every GFC day lands in the most extreme state -- and the COVID-19 window's mean state index is 12.7. The tercile-based contagion-intensification estimate is 160.4% (raw correlation 0.139 in the lowest tercile to 0.361 in the highest), within a percentage point of Section 6.4's truncated-DP-HMM estimate (161.1%) -- a reassuring cross-method consistency check: two structurally different regime-detection algorithms, applied to the same underlying volatility path, agree almost exactly on the size of the contagion-intensification effect, even though neither agrees closely with the fixed-threshold (49%) or feasibility-analog (57.7%/6.6%) estimates from earlier in the paper.

| Approach                    | Estimator                                      | Intensification |
|-----------------------------|------------------------------------------------|-----------------|
| Christopher (2026)          | Fixed-threshold GJR-GARCH, 3 regimes           | +49.0%          |
| Section 5 (feasibility)     | GARCH-DCC correlation, HMM $K=3$               | +6.6%           |
| Section 5 (feasibility)     | Raw correlation, HMM $K=3$                     | +57.7%          |
| Section 6.4 (full Bayesian) | Raw correlation, MSV+truncated-DP-HMM terciles | +161.1%         |
| Section 6.7 (full Bayesian) | Raw correlation, MSV+HDP-HMM terciles          | +160.4%         |

Table 16. Updated contagion-intensification estimates across all methodologies attempted in this paper.

- Concentration hyperparameters ( $\alpha = \gamma = 2.0$ ) were fixed rather than themselves given a hierarchical prior and sampled (the standard further extension, via Escobar & West 1995-style auxiliary

sampling) -- an additional, lower-priority source of unverified complexity that was not pursued given the consistency already found against the simpler truncated approach.

- Literal beam sampling itself -- a different inference algorithm for the identical model, motivated by removing the weak-limit truncation entirely rather than by any known deficiency in the direct-assignment results above -- was implemented next and is reported in Section 6.8, which closes this gap.

## 6.8 Literal Beam Sampling: Removing the Weak-Limit Truncation Entirely

Sections 6.3 and 6.7 both represent the HDP-HMM's countably infinite state space with a fixed, generously-sized truncation ( $L = 15$  states). Beam sampling (Van Gael, Saatchi, Teh & Ghahramani, 2008) removes that truncation entirely, and was the one piece of Section 3.3's specification still not literally implemented as of Section 6.7. It works by introducing, each sweep, auxiliary slice variables  $u_t \sim \text{Uniform}(0, \pi_{\{z_{t-1}, z_t\}})$  drawn from the transition probability of the CURRENT sampled path; only transitions with probability above the relevant  $u_t$  can be active at time  $t$ , and because the HDP's stick-breaking prior makes transition mass decay geometrically in expectation, that active set is always finite -- exact forward-filtering/backward-sampling over it is therefore exact for the full unbounded-state model, with no state count fixed in advance anywhere. New states are instantiated on the fly, only when the slice variables actually demand access to transition mass beyond what is currently represented, by splitting off a piece of the remaining stick-breaking mass via the standard Dirichlet/DP two-way splitting identity (the same identity underlying Walker's (2007) and Kalli, Griffin & Walker's (2011) slice samplers for Dirichlet process mixtures, applied here to the HDP-HMM as in Van Gael et al.'s original appendix).

What remains a simplification even here: concentration hyperparameters ( $\alpha$ ,  $\gamma$ ) are fixed rather than themselves sampled, matching every other sampler in Section 6; the initial-state distribution is treated as its own  $\text{DP}(\gamma, \beta)$  draw, the same convention used by Van Gael et al. and by Fox et al. (2011); a newly-created state's own outgoing transition row is drawn once as a fresh Dirichlet sample rather than progressively extended stick-by-stick, harmless since no data is yet assigned to it; and per-sweep state creation carries a generous safety cap (60 new states per sweep) purely as a numerical guard against runaway growth in early, unconverged sweeps -- the real-data run below never came close to it. States with zero posterior mass are pruned back into the unrepresented stick each sweep, so the instantiated-state count tracks what the data currently need rather than accumulating stale, empty states.

Validated first on the same synthetic three-state test used throughout Section 6: the beam sampler recovered exactly 3 active states in 100% of posterior draws, achieved through genuine dynamic creation and pruning (the instantiated-state count ranged from 3 to 6 over the course of the run, never fixed and never near any cap) rather than through shrinkage inside a pre-sized truncation. This is the cleanest possible confirmation that the growth/pruning machinery itself is correctly specified before trusting it on real data.

Applied to the same ASIS-re-estimated MSV posterior-mean volatility path used by Section 6.7, the beam sampler completed 2500 sweeps (800 burn-in, 1700 kept) in 182 seconds -- markedly faster than either truncated-state-space method in this paper, despite implementing strictly more machinery, precisely because it is never forced to maintain and update states the data do not need. The growth-cap safety valve was never triggered (0 of 1700 kept sweeps). The headline finding is in the effective state count itself: freed from any fixed ceiling, the posterior settled on 11 active states in 68.8% of draws (the remainder mostly 12), with the raw instantiated-state count (before the >1%-occupancy filter) ranging only 11-14 across the entire kept chain (mean 11.5) -- fewer states than either Section 6.3's truncated sampler (which used all 15 of its 15 available slots) or Section 6.7's direct-assignment HDP-HMM (which used 14 of its 15). This nuances, rather than overturns, the finding repeated since Section 5.5: joint volatility does not support a small number of discrete regimes, but the honest reading is now 'roughly a dozen persistent states,' not 'as many as the truncation happens to allow' -- a distinction only a genuinely unbounded sampler can settle, since a truncated method's full-slot usage is equally consistent with 'the truncation was too small' as with 'the data want exactly that many states.'

Crisis validation remains as strong as elsewhere in this paper: 100% of GFC-window trading days fall in the single most extreme of the 11 states, and 80.3% of COVID-19-window days do too (the remainder split between the

second- and third-most extreme states, still firmly in the high-volatility tail). The tercile-based contagion-intensification statistic reported in the table below is included for completeness, but it is not an independent cross-check the way the effective-state-count and crisis-validation findings above are: it is computed directly from the shared MSV posterior-mean volatility path's own terciles, not from any method's regime assignments, so it is numerically identical to Section 6.7's (160.4%) by construction -- both samplers were handed the same underlying path and the tercile split does not consult either one's state labels. Reporting it here without that caveat would overstate how many genuinely independent estimates this paper has actually produced.

| Approach                               | Estimator                                                              | Intensification                                                     |
|----------------------------------------|------------------------------------------------------------------------|---------------------------------------------------------------------|
| Christopher (2026)                     | Fixed-threshold GJR-GARCH, 3 regimes                                   | +49.0%                                                              |
| Section 5 (feasibility)                | GARCH-DCC correlation, HMM K=3                                         | +6.6%                                                               |
| Section 5 (feasibility)                | Raw correlation, HMM K=3                                               | +57.7%                                                              |
| Section 6.4 (full Bayesian)            | Raw correlation, MSV+truncated-DP-HMM terciles                         | +161.1%                                                             |
| Section 6.7 (full Bayesian)            | Raw correlation, MSV+HDP-HMM (direct-assignment) terciles              | +160.4%                                                             |
| Section 6.8 (full Bayesian)            | Raw correlation, MSV+HDP-HMM (beam-sampled) terciles                   | +160.4% (identical to 6.7 by construction -- see note above)        |
| Section 8.1 (bias-corrected)           | Forbes-Rigobon (2002) heteroskedasticity-adjusted correlation, HMM K=3 | -49.5% (sign-reversed -- see Section 8.1)                           |
| Section 8.3 (subsample, full Bayesian) | Truncated DP-HMM extreme-state, first half / second half               | +295.7% / -1318.1% (sign-unstable across halves -- see Section 8.3) |

Table 17. Contagion-intensification estimates across every methodology attempted in this paper, now eight. The Section 6.7 and 6.8 rows share an identical tercile-based statistic because both are computed from the same underlying MSV volatility path, not from either sampler's own regime labels. The Section 8.1 row applies the field's standard remedy for naive-correlation heteroskedasticity bias to the same Section 5.7 low/high-regime partition used in the paper's first full-sample estimate, and reverses its sign; see Section 8.1 before treating any single row in this table as this paper's answer.

- What genuinely distinguishes the beam-sampled result from Section 6.7's is therefore not the tercile-based contagion statistic (identical by construction) but the effective-state-count finding above -- a materially different, and more trustworthy, answer to 'how many regimes does joint volatility actually support' precisely because no truncation ceiling was available to be bumping against.
- With this closed, every piece of Section 3.3's iHMM specification (hierarchical Dirichlet process formulation, beam-sampled unbounded inference) has now been literally implemented and validated; what remains of the full Section 3 specification is (i) full joint uncertainty propagation -- interleaving the MSV and HMM Gibbs steps into a single sampler rather than conditioning the regime-detection stage on the MSV posterior MEAN path, as every method in Section 6 does -- and (ii) the ML ensemble's gradient-boosted-tree swap, which remains blocked on package availability rather than any methodological gap.
- Table 17's seventh row, added in Section 8.1, is not a further robustness check in the usual sense: applying Forbes & Rigobon's (2002) own heteroskedasticity correction to this paper's regime-conditional correlation reverses its sign entirely, which is itself informative about the limits of correlation-based contagion measures under this dataset's regime-to-regime variance ratio -- see Section 8.1 for the full result and Section 7 for what it implies about which of this paper's eight estimates should be weighted most heavily.

## 7. Discussion

Section 1 posed a specific question: does a jointly-estimated, ML-augmented regime-detection framework identify contagion dynamics materially different from the fixed-threshold framework used in the author's prior work? Table 17's eight estimates do not converge on a single number, and this section argues that this is the right outcome to report, not a result to average away, because the eight estimates are not equally trustworthy and the reasons they disagree are themselves informative.

On direction, every regime-detection method attempted in this paper agrees with the prior paper: crisis-period days cluster overwhelmingly in the most extreme regime under every method tried (100% of GFC days, 80-94% of COVID-19 days, depending on method), and the properly-specified Bayesian estimate of cross-market volatility-innovation correlation -- Sigma\_eta's off-diagonal entry, posterior mean 0.824, 95% credible interval [0.743, 0.892] (Section 6.2) -- is large, positive, and precisely estimated. This quantity is structurally different

from every correlation-based estimate in Table 17: it is a parameter of the joint latent-volatility model itself, estimated from the full sample without conditioning on any regime split, and it is not subject to the specific bias mechanism (correlation inflation from a period-selection rule correlated with volatility) that Forbes & Rigobon (2002) target. On this evidence, the paper's central substantive claim is that cross-market volatility comovement is genuinely elevated during high-volatility episodes, not merely an artefact of how the episodes are defined.

On magnitude, the eight estimates disagree by more than an order of magnitude, and Section 8.1 shows why at least one of them should not be trusted at face value. Forbes & Rigobon's (2002) own remedy for the naive-correlation bias they identified -- applied here to the same low/high-regime partition behind this paper's first full-sample estimate -- does not shrink the estimated intensification toward a more credible middle ground. It reverses its sign (Table 17, Section 8.1), a result inconsistent with the crisis-clustering evidence above and with every other estimator in this paper. Section 8.1 traces this to a known weakness of the Forbes-Rigobon correction, documented by Corsetti, Pericoli & Sbracia (2005): the correction is derived under a single-factor model with proportional variance scaling, and becomes unreliable precisely when the cross-regime variance ratio is large, as it is here (an 11-fold increase in S&P 500 return variance between the low- and high-volatility regimes). This paper does not merely cite that concern; Section 8.1 demonstrates it directly on real data, which is a methodological contribution independent of the contagion application.

That magnitude sensitivity is not unique to this paper's specific models; Section 2.4's review of the contagion-measurement literature documents the same qualitative pattern across other datasets. What this paper adds is a direct demonstration, on one controlled dataset, of how large that sensitivity can be even among estimators that are all individually defensible: raw regime-conditional correlation versus DCC-based correlation (a 9-fold difference within Section 5 alone, Section 5.7); tercile-based versus fixed-threshold partitioning (Section 6.4); the choice of state-space representation for the SAME nonparametric Bayesian model (Section 6.8); and, most starkly, whether the standard heteroskedasticity correction is applied at all (Section 8.1). Weighed against each other, the crisis-validated, bias-immune Bayesian correlation estimate and the sign-reversing bias-corrected naive estimate point in different directions using the same underlying data -- and the paper's position, argued above, is that the former is the more credible of the two, precisely because it is not exposed to the mechanism the latter is trying and failing to correct for.

The beam-sampling result (Section 6.8) deserves discussion on its own methodological terms, beyond its role in the contagion-magnitude comparison. Both truncated implementations in this paper (Section 6.3's weak-limit approximation and Section 6.7's direct-assignment HDP sampler) used nearly all of their allotted  $L = 15$  states (15/15 and 14/15 respectively), which could be read either as evidence the data genuinely want that many regimes, or as an artefact of the truncation itself acting as an attractor. Only a sampler with no ceiling at all can distinguish these two readings, and the beam sampler's settling on 11-12 states -- reached through genuine dynamic state creation and pruning, never approaching the numerical safety cap -- supports the second reading over the first. This matters beyond this paper: any applied HDP-HMM study that reports high state-space utilisation under a fixed truncation should treat that as a prompt to check with an untruncated sampler before treating the state count as a robust finding, not as confirmation in itself.

The machine learning ensemble layer's role is more circumscribed than 'AI-augmented' framing in the broader literature sometimes implies. Section 5.6 finds that over 92% of the trained classifiers' feature importance concentrates in realised volatility alone, and that agreement with the HMM baseline degrades from 95.2% in-sample to 81.4% out-of-sample -- a roughly four-fold increase in divergence that suggests the ensemble's apparent validation of the HMM regimes is partly an in-sample artefact. The honest reading, consistent with Section 3.4's framing, is that the ML layer functions as a partial, feature-driven robustness check on the Bayesian nonparametric component, not as an independent confirmation of it, still less as a replacement for genuine nonparametric Bayesian regime detection.

Taken together, this paper's answer to its own research question is structured rather than singular: yes, cross-market volatility comovement is genuinely elevated in high-volatility regimes, on the strength of crisis-clustering evidence and a properly-specified, bias-immune Bayesian correlation parameter; but the SIZE of that effect, as conventionally reported via regime-conditional correlation statistics in this literature, is highly sensitive to estimator choice and cannot be taken as a stable, portable number -- a conclusion this paper reaches not by asserting it but by computing eight structurally different estimates, one of them via the field's own standard bias correction, and showing that they disagree in a pattern the literature already predicts.

## 8. Robustness

This section adds two checks beyond those already reported alongside each result as it was introduced: within-model uncertainty quantification via moving-block bootstrap confidence intervals (Section 5.8) and MCMC diagnostics with synthetic-data validation preceding every real-data run (Section 6, throughout); and sensitivity to the regime-count/state-space choice (Sections 5.5, 5.8, 6.3, 6.7, 6.8). Section 8.1 applies the contagion-measurement literature's own standard bias correction to this paper's data directly; Section 8.2 checks sample-period stability of the feasibility-analog pipeline; Section 8.3 extends that check to the full Bayesian MSV+iHMM chain (Appendix B.7 separately checks sensitivity to the samplers' own hyperparameters); and Section 8.4 adds a nonparametric placebo test of the regime classification itself.

### 8.1 Forbes-Rigobon Heteroskedasticity-Bias Correction

Section 2.4 reviews Forbes & Rigobon's (2002) foundational critique of naive correlation comparisons: because realised correlation estimators are themselves a function of the underlying volatility, comparing correlation across a calm and a turbulent period is biased upward by the change in volatility alone, independent of any genuine change in cross-market dependence. Their proposed correction, applied here for the first time to this paper's own regime-conditional correlation rather than only cited, is

$$\text{rho\_adj} = \text{rho\_high} / \sqrt{1 + \text{delta} (1 - \text{rho\_high}^2)}, \quad \text{delta} = (\text{Var\_high} / \text{Var\_low}) - 1$$

where rho\_high is the naive high-volatility-regime correlation, and delta is the proportional increase in variance of the originating market (S&P 500, consistent with the direction of volatility transmission assumed throughout the contagion literature reviewed in Section 2.4) between the low- and high-volatility regimes. Applied to the identical low/high-regime partition behind Section 5.7's first full-sample estimate (regime\_hmm = 0 versus regime\_hmm = 2, N = 1713 and N = 911 respectively): rho\_low = 0.242, rho\_high = 0.381 (naive, matching Table 17's Section 5.7 row), and delta = 10.23 -- an 11-fold increase in S&P 500 return variance between the two regimes. The Forbes-Rigobon-adjusted correlation is rho\_adj = 0.122, giving an adjusted intensification of -49.5% (block-bootstrap 95% CI [-62.0%, -32.7%], 5,000 resamples) against the naive +57.7%.

This is not a shrinkage toward zero of the kind the Forbes-Rigobon correction is usually used to report; it is a sign reversal, with a 95% interval that excludes zero on the negative side. Taken at face value, it would say that cross-market correlation, once corrected for heteroskedasticity, is LOWER in the high-volatility regime than the low-volatility regime -- directly contradicting the crisis-clustering evidence (100% of GFC days, 93.9% of COVID-19 days classified high-volatility; Section 5.5) and every other estimator in Table 17. Corsetti, Pericoli & Sbracia (2005) document exactly this failure mode: the Forbes-Rigobon correction assumes a single common factor with proportional variance scaling and constant factor loadings across regimes, assumptions that become increasingly fragile as the cross-regime variance ratio grows, and can produce economically implausible corrected estimates -- including sign reversals -- once that ratio is large. This paper's delta of 10.2 is large by the standards of the correction's original applications (Forbes & Rigobon's 1997 Asian and 1994 Mexican crisis episodes used considerably smaller variance ratios), which is the most likely explanation for the reversal reported here, not evidence that cross-market dependence genuinely falls during crises.

The methodological lesson is not that the Forbes-Rigobon correction is wrong in general -- it remains the field's standard response to a real bias -- but that applying it mechanically, without checking whether its own assumptions hold for the variance ratio at hand, can produce a confidently-reported number that is less trustworthy than the naive estimate it was meant to improve on. This is a direct, practical reason to prefer this paper's Bayesian MSV correlation estimate (Section 6.2, Sigma\_eta correlation 0.824) over any correlation-based estimator in this table for the substantive question of whether cross-market dependence is genuinely elevated: the MSV estimate is not derived from a regime-conditional sample correlation at all, so it carries neither the naive estimator's heteroskedasticity bias nor the correction's own large-delta fragility.

One convention embedded in the calculation above is worth checking directly rather than assumed: delta is computed from the S&P 500's variance ratio, treating it as the 'originating market' whose volatility change defines the correction, following the contagion literature's usual convention of treating the larger, more liquid market as the shock source. Recomputing delta from Nifty 50's variance ratio instead gives delta = 5.87 (smaller than the S&P-500-based 10.23, since Nifty 50's own variance increases proportionally less between regimes) and a Forbes-Rigobon-adjusted intensification of -35.7% (95% CI [-53.0%, -10.8%]) -- still a sign reversal, with a materially overlapping interval to the S&P-500-based estimate above. The sign-reversal finding is therefore not an artefact of which market's variance ratio is used to define the correction; it is present, though smaller in magnitude, under either direction-of-transmission convention, consistent with this section's explanation (a large cross-regime variance ratio breaking the correction's own assumptions) rather than a convention-specific coding artefact.

## 8.2 Subsample Stability

The joint sample (N = 4,125 trading days, 18-Sep-2007 to 31-Dec-2024) was split at its exact temporal midpoint -- 2016-05-19, not a cherry-picked date -- into a first half (2007-09-18 to 2016-05-18, N = 2062, GFC-dominated) and a second half (2016-05-19 to 2024-12-31, N = 2063, COVID-19- and 2022-rate-shock-dominated). The full Section 5 feasibility-analog pipeline -- GARCH(1,1)-t on each market, Gaussian HMM with K = 3 states on the resulting log-volatility series, raw regime-conditional correlation -- was refit independently on each half.

| Sample                    | Period    | N     | Low-regime corr. | High-regime corr. | Intensification |
|---------------------------|-----------|-------|------------------|-------------------|-----------------|
| Full sample (Section 5.7) | 2007-2024 | 4,125 | 0.242            | 0.381             | +57.7%          |
| First half                | 2007-2016 | 2062  | 0.258            | 0.354             | +36.9%          |
| Second half               | 2016-2024 | 2063  | 0.223            | 0.363             | +62.3%          |

Table 18. Subsample stability of the raw regime-conditional contagion-intensification statistic. Both halves independently refit (own GARCH(1,1)-t parameters, own K=3 HMM, own regime thresholds); low/high-regime correlation and intensification computed exactly as in Section 5.7.

Direction is stable: both halves show correlation rising materially from the low- to the high-volatility regime, and neither is close to the null of no intensification. Magnitude is not stable: the second half's intensification (62.3%) is nearly double the first half's (36.9%), and neither matches the full-sample estimate (57.7%) exactly, since the full-sample HMM's regime boundaries and GARCH parameters differ from either half's own independently-fit versions. This reinforces Section 7's broader conclusion: the DIRECTION of the naive contagion-intensification effect is a robust, replicable finding across every sample split and estimator attempted in this paper (with the single, explained exception of the bias-corrected estimate in Section 8.1); the specific MAGNITUDE is not, and should not be quoted as a precise, portable number without also reporting the estimator and sample period it was computed under.

- The check above reuses only the feasibility-analog pipeline; Section 8.3 below extends it to the full Bayesian MSV+iHMM chain itself.
- The bootstrap intervals in Section 5.8 and the subsample check above both treat the regime classification as fixed within each resample/subsample; neither propagates uncertainty from the regime-detection step itself into the reported interval, which only a full joint Bayesian estimation (Section 3.5's remaining gap, discussed in Section 10) would characterise directly through the posterior.

## 8.3 Full-Bayesian Subsample Re-estimation

Section 8.2's subsample check refits only the fast feasibility-analog pipeline, leaving open whether the paper's central Bayesian estimates -- the MSV Sigma\_eta correlation (Section 6.2) and the truncated-DP-HMM contagion-intensification statistic (Section 6.3) -- are similarly stable. This section closes that gap by re-running the actual Gibbs/FFBS MSV sampler and the actual truncated DP-HMM sampler independently on each half of the sample (identical split as Section 8.2, 2007-2016 and 2016-2024), rather than only the fast approximations.

| Sample                    | N    | Sigma_eta corr. (posterior mean) | 95% credible interval |
|---------------------------|------|----------------------------------|-----------------------|
| Full sample (Section 6.2) | 4125 | 0.824                            | [0.743, 0.892]        |
| First half (2007-2016)    | 2062 | 0.796                            | [0.650, 0.906]        |
| Second half (2016-2024)   | 2063 | 0.822                            | [0.696, 0.912]        |

Table 19. Full-Bayesian MSV subsample re-estimation (12,000 iterations, 4,000 burn-in, thin=2 each half; own Gibbs/FFBS chain, not the fast feasibility analog). Both halves' credible intervals contain the full-sample posterior mean and overlap each other substantially.

Unlike Section 8.2's feasibility-analog correlation statistic, the Bayesian MSV cross-market correlation is materially stable across both halves and the full sample: 0.796 (first half) versus 0.822 (second half) versus 0.824 (full sample), with all three 95% credible intervals overlapping substantially. This is the strongest single piece of subsample evidence in the paper for the Bayesian correlation estimate's reliability specifically: not only is it structurally immune to the Forbes-Rigobon-style regime-conditioning bias (Section 8.1), it is also empirically stable to which decade of data it is estimated on, in sharp contrast to the feasibility-analog correlation-intensification statistic of Section 8.2, whose magnitude nearly doubled between halves.

The truncated DP-HMM subsample results tell a different, less reassuring story. Applying the identical low-versus-high extreme-state comparison used for the full-sample intensification statistic (Section 6.3: 204.5%) to each half separately gives 295.7% for the first half and -1318.1% for the second half -- not merely unstable in magnitude the way Section 8.2's statistic is, but unstable in SIGN, and by a very large margin in the second half. Both subsample truncated DP-HMM runs, like the full-sample run (Section 7's discussion), settle on close to the full L=15 truncation states as active, so this extreme-state comparison is, on each half, a comparison between two of fifteen finely-split states with modest sample sizes (roughly 100-150 observations apiece; Table 19's underlying regime\_stats). With that few observations in the compared states, the regime-conditional correlation itself is estimated with substantial sampling noise, and a single volatile stretch of days falling into the extreme state by chance can move the ratio-based intensification statistic by a large percentage even when the underlying dependence structure has not changed appreciably.

This finding reinforces rather than undermines the paper's central argument, and is reported here for that reason rather than omitted as an inconvenient result. Section 7 already argues, on other grounds, that correlation-RATIO statistics computed by comparing two regimes -- whether naive, DCC-based, Forbes-Rigobon-corrected, or here, extreme-state truncated-DP-HMM-based -- are the least trustworthy class of estimate in this paper, precisely because they compress a full joint model down to a two-number comparison that is sensitive to exactly how the comparison is drawn. The Bayesian MSV correlation parameter (Table 19) is not subject to this fragility because it is not a regime-conditional ratio at all; it is estimated once, jointly, from the whole path. The extreme sign instability documented here is a second, independent demonstration of the same estimator-sensitivity thesis Section 8.1 established via the Forbes-Rigobon correction, now shown for a fully nonparametric Bayesian regime-detection method rather than a simple threshold split -- sharpening rather than qualifying the paper's recommendation to treat any single regime-conditional correlation-ratio number, however estimated, as unreliable evidence of contagion magnitude on its own.

#### 8.4 Placebo Permutation Test

Sections 5.8 and 8.1-8.3 quantify uncertainty and estimator-sensitivity conditional on the regime classification being a meaningful split of the data. This subsection asks a logically prior question directly: is the Section 5.7 baseline intensification statistic (+57.7%) larger than what an ARBITRARY, information-free split of the same

size would produce? The regime labels are randomly reshuffled across all  $N=4,125$  trading days, holding the two group sizes fixed at their true HMM-regime counts ( $N_{\text{low}} = 1713$ ,  $N_{\text{high}} = 911$ ) but otherwise breaking any relationship between which days are assigned 'low' and which are assigned 'high' and the markets' actual joint behaviour on those days, and the intensification statistic is recomputed on each reshuffled split (10,000 permutations).

Under this null of arbitrary regime assignment, the intensification statistic centres near zero, as it should (mean 1.7%, 95% interval [-36.3%, 48.0%]), confirming that group-size asymmetry alone (1,713 versus 911 days) is not enough to generate a large positive statistic mechanically. The true, HMM-regime-based statistic (+57.7%) lies 2.59 standard deviations above this null distribution's mean, corresponding to a two-sided permutation p-value of 0.0119. This rejects the null of an information-free split at conventional significance levels, but the margin is more modest than the crisis-clustering or Bayesian-correlation evidence discussed in Section 7: a p-value just above 0.01, not the nearly-zero p-values reported for the Ljung-Box/ARCH-LM tests in Section 4.3 or the Jarque-Bera normality tests in Section 4.2. Read honestly, this placebo test adds a genuine but comparatively modest additional piece of evidence that the HMM's specific day-to-regime assignment carries real information about joint market behaviour beyond its group sizes alone; it does not, on its own, establish the MAGNITUDE of the resulting statistic as reliable, which is precisely the point Sections 7 and 8.1-8.3 make on other grounds.

## 9. Expected Contribution

Substantial parts of the contribution below are no longer purely prospective: Sections 5 and 6 built and ran real versions of the combined MSV + iHMM + ML-ensemble framework, up to and including a validated ASIS-interweaved MSV, a true HDP-HMM, and a literal beam sampler over an unbounded state space (Section 6.8); Sections 7 and 8 then synthesised what those results mean together and checked their stability, extending the stability check to the full Bayesian chain itself (Section 8.3) and to the samplers' own hyperparameters (Appendix B.7); and Appendix B supplies the formal derivations underlying every estimation step used. What remains genuinely ahead is fully joint uncertainty propagation across the MSV and regime-detection stages and the bridge to graph neural networks described in Section 10.

- Methodological: a combined MSV + iHMM + ML-ensemble framework not previously documented in the reviewed literature, applied to a developed-emerging market contagion pair.
- Empirical: a direct, same-data comparison between a fixed-threshold GARCH approach (the author's first paper) and a jointly-estimated, ML-augmented Bayesian nonparametric approach, extended to eight structurally different contagion-magnitude estimates (Table 17) spanning a naive regime-conditional correlation, a DCC-smoothed correlation, four full-Bayesian nonparametric variants, a heteroskedasticity-bias-corrected estimate, and a full-Bayesian subsample re-estimation -- with the disagreement pattern across all eight treated as itself the paper's central empirical finding, not a nuisance to be averaged away.
- Diagnostic: use of ML ensemble classifiers as an independent robustness check on Bayesian nonparametric regime detection, a comparative approach not identified elsewhere in the reviewed literature.
- Methodological-transparency: full derivations, from Bayes' rule, of every conjugate-posterior and filtering step actually used in estimation (Appendix B.1-B.3), the formal equivalence between the paper's GEM-stick-breaking model statement and its Chinese-Restaurant-Franchise implementation (Appendix B.4), real convergence diagnostics computed on the paper's own saved posterior draws rather than asserted (Appendix B.5), and a fully worked, hand-checkable numerical example cross-validated against the paper's own code (Appendix B.6) -- intended to make the paper independently re-implementable, not only citable.
- Robustness-methodological: a direct demonstration that concentration-hyperparameter choice ( $\alpha$ ,  $\gamma$ ) for the truncated DP-HMM and prior-informativeness choice for the MSV's conjugate priors do not materially change this paper's headline point estimates (Appendix B.7), a check often asserted informally in applied Bayesian nonparametric work but rarely reported with the specific comparison numbers shown here.
- Forward-looking: the resulting regime states and ML-derived features would serve directly as node-level inputs for the temporal graph neural network architecture described in Section 10.

## 10. Future Research Directions

Beyond full joint uncertainty propagation (Sections 3.5, 8, 11), the most natural extension of this paper's core MSV-iHMM-ML framework -- currently applied to the existing bivariate S&P 500/Nifty 50 pair -- is to a genuinely multi-institutional setting via a temporal graph neural network for cross-market contagion detection. The iHMM-derived regime states and the ML ensemble's engineered feature set both translate directly into node-level features for a GNN architecture of the kind reviewed in Section 2.5, and the divergence-diagnostic approach in Section 3.4 provides a template for validating any future GNN-based regime classification against both a Bayesian statistical baseline and a supervised ML baseline.

A second extension is to replace the scikit-learn ensemble (Section 5.6) with the literal gradient-boosted-tree implementations specified in Section 3.4 (XGBoost/LightGBM), to confirm the ensemble-classifier results in Section 5.6 are not an artefact of the substitute implementation used here. A third extension, a subsample-specific full Bayesian re-estimation, is reported in Section 8.3: it surfaced a genuine, unresolved instability in the truncated DP-HMM's extreme-state contagion statistic that the correlation-only feasibility check (Section 8.2) alone would not have revealed, and this motivates a fourth, more specific extension -- an intensification statistic defined over a coarser, more heavily-populated grouping of the iHMM's active states (for example, terciles of the modal-K state ordering rather than only its two most extreme individual states) might prove less sensitive to sample-splitting noise than the extreme-state comparison used throughout this paper, at the cost of losing some of the fine-grained regime structure the nonparametric prior is specifically designed to recover; testing that tradeoff directly is left to future work.

A fifth extension follows directly from Appendix B.5's convergence-diagnostic finding that the centred (non-ASIS) MSV sampler shows a measurable Geweke non-convergence signal for both persistence parameters at this paper's chain length: a longer centred-sampler run, or a formal multi-chain Gelman-Rubin  $\hat{R}$ -hat comparison across several independently-initialised chains (not attempted here, where a single long chain per specification was judged sufficient given ASIS's own resolution of the same diagnostic), would sharpen exactly how much of that signal reflects genuinely slow mixing versus a chain length that is merely inconvenient rather than insufficient.

## 11. Limitations

- MCMC estimation for MSV+iHMM models is computationally intensive and may require reduced MCMC chain length or a smaller state-space cap relative to a fully unconstrained infinite model; the feasibility results in Section 5.5 illustrate the underlying problem directly -- a plain Gaussian HMM's BIC does not naturally bound the state count, which is exactly why the iHMM's Dirichlet-process prior, rather than a post-hoc information criterion, is needed. Section 6.8's literal beam sampler shows this is a concern about scale rather than fundamental feasibility -- at this paper's  $T \sim 4,125$  sample size, unconstrained inference actually ran faster than both truncated/weak-limit alternatives -- though the same caveat would need revisiting on a much longer or higher-dimensional series, where per-sweep state creation could become the binding cost.
- The ML ensemble layer requires engineered features and is not itself a claim to full 'AI-driven' regime detection; it is positioned as a robustness check on the Bayesian nonparametric component, not a replacement for it. Section 5.6's feasibility results reinforce this: over 92% of the trained classifier's feature importance is concentrated in realised volatility alone, and its divergence from the HMM baseline nearly quadruples out-of-sample (4.8% to 18.6%), so the ML layer should be read as a partial, not independent, check.
- As with the author's first paper, the bivariate scope (two markets only) remains a limitation relative to a full multi-institutional network, which is reserved for future work (Section 10).
- The size of the reported contagion-intensification effect is sensitive to the specific estimator used: Section 5.7 finds a 57.7% increase in raw regime-conditional correlation but only a 6.6% increase in the smoother DCC-implied correlation between the same two regimes, both statistically distinguishable from zero (Section 5.8) but of very different economic magnitude. The full-Bayesian estimates in Section 6.4 and 6.7 push this further still (+161.1% and +160.4% respectively); a sixth, beam-sampled estimate (Section 6.8) replicates the fifth by construction; a seventh, applying Forbes & Rigobon's (2002) standard bias correction to the Section 5.7 partition (Section 8.1), reverses the sign entirely (-49.5%); and an eighth, re-estimating the full Bayesian truncated-DP-HMM chain independently on each half of the sample (Section 8.3), swings even

more widely -- +295.7% on the first half, -1318.1% on the second -- despite using the identical extreme-state comparison method as the fourth estimate above. The eight estimates produced across this paper therefore span from a large negative sign reversal to a 295.7% increase depending on estimator, regime-partition choice, sample period, and whether the field's standard bias correction is applied at all, on the identical dataset. A subsample-stability check on the fast feasibility-analog pipeline (Section 8.2) confirms the DIRECTION of the naive correlation-based effect, though not its magnitude, is stable across the sample's two halves; the corresponding check on the full Bayesian chain (Section 8.3) is far LESS reassuring for the truncated-DP-HMM extreme-state statistic specifically, though it confirms the Bayesian MSV correlation parameter itself (0.824, Section 6.2) is stable both across subsamples (0.796, 0.822) and, separately, across a range of concentration-hyperparameter and prior choices (Appendix B.7). This paper does not treat all eight estimates as equally credible (Section 7 argues the bias-corrected estimate is the least trustworthy of the eight, not the most, and Section 8.3 argues the same of the subsample extreme-state estimate), but even setting both aside, the remaining range is wide enough that any comparison to the fixed-threshold 49% figure in Christopher (2026) -- or to any of this paper's own numbers in isolation -- should be read with this sensitivity in mind (see Section 7 for discussion).

- The bootstrap confidence intervals reported in Section 5.8 treat the HMM regime classification as fixed and known; they do not propagate uncertainty from the regime-detection step itself. A full joint Bayesian estimation (Section 3.5) would characterise this additional source of uncertainty directly through the posterior.
- Section 8.1's Forbes & Rigobon (2002) correction is itself only as reliable as its own assumptions (a single common factor, proportional variance scaling, constant factor loadings across regimes), which this paper argues are violated by its data's large cross-regime variance ratio. This paper's evidence for that violation is indirect -- an implausible sign reversal contradicted by independent crisis-clustering evidence -- rather than a direct test of the correction's assumptions themselves, which would require estimating the underlying factor model explicitly.
- As noted in Section 5.2, the NYSE and NSE trading sessions do not overlap in clock time, so same-calendar-day closing returns understate true contemporaneous comovement to some degree -- a limitation inherited directly from Christopher (2026) that none of the methodological extensions in this paper (MSV, iHMM-analogs, or the ML ensemble) address, since all operate on the same daily-aligned return series.

## 12. Conclusion

This paper set out to test, on a controlled and identical dataset, whether a jointly-estimated, machine-learning-augmented regime-detection framework identifies materially different contagion dynamics than the fixed-threshold GJR-GARCH approach used in the author's first paper. The full methodological answer to that question requires the Bayesian MSV and infinite HMM estimation specified in Section 3; a literal, ASIS-and-beam-sampling-faithful implementation of that exact specification has now been attempted for both the MSV and the iHMM (Sections 6.6 and 6.8 respectively), but Sections 5 and 6 together represent three escalating attempts at the full picture -- a tractable real-data feasibility analog, and then an actual Gibbs/FFBS MSV plus truncated Dirichlet-process HMM run to completion on the real data.

Three findings stand out from the Section 5 feasibility work specifically. First, a data-driven Gaussian HMM regime layer validates strongly against independent crisis-period evidence (classifying 100% of GFC days and 93.9% of COVID-19 days as high-volatility) while also revealing a genuine limitation of naive information-criterion-based state selection -- direct empirical support for the iHMM's Dirichlet-process approach over ad hoc alternatives. Second, an ML ensemble layer adds only partial, not fully independent, robustness: its classification power is concentrated almost entirely in realised volatility, and its agreement with the HMM baseline degrades noticeably out of sample. Third, and most directly relevant to the paper's central comparison, the jointly-estimated framework replicates the direction and rough order of magnitude of the prior paper's 49% contagion-intensification finding (57.7% under raw regime-conditional correlation, with a 95% bootstrap interval that excludes zero), but the size of that effect is highly sensitive to the specific estimator used -- as low as 6.6% under a smoothed dynamic-correlation measure. That sensitivity is itself a finding: single-point contagion statistics, from either the fixed-threshold or the feasibility-analog approach, should be treated as estimator-dependent rather than as a settled magnitude.

That next stage was in fact attempted in Section 6: a Gibbs/FFBS estimation of the bivariate MSV and a truncated Dirichlet-process HMM, both run to completion on the real data. The results both confirm and complicate the feasibility-analog picture. They confirm it in the sense that demonstrates real, previously undocumented correlation between the two markets' latent volatility processes (posterior mean 0.824, 95% credible interval excluding zero and every lower estimate in this paper) and independently reproduce, via a second and methodologically distinct route, the finding that a small fixed number of regimes is not naturally supported by the data. They complicate it in the sense that the resulting contagion-intensification estimate (+161.1%) is the largest of five produced across this paper's methodologies, underscoring that the size of this effect is highly estimator- and partition-dependent rather than a single settled number. The ASIS interweaving step flagged as genuinely needed by Section 6.5's diagnostics was then implemented and validated in Section 6.6, meaningfully improving mixing (effective sample size roughly doubling for the persistence parameters) while confirming the original point estimates were not badly biased. The iHMM-analog was itself then upgraded from the truncated approximation to a true HDP-HMM via direct-assignment Gibbs sampling (Section 6.7), which agreed with the simpler approximation to within a percentage point on the contagion-intensification estimate -- a reassuring cross-method consistency check. Literal beam sampling (Section 3.3 / Van Gael et al., 2008) was implemented next (Section 6.8), removing the weak-limit truncation entirely; freed from any fixed ceiling, the posterior settled on a genuinely smaller effective state count (11-12, versus 14-15 for the two truncated/weak-limit methods) -- a refinement of, not a reversal of, the repeated finding that joint volatility does not support a small number of discrete regimes. Section 8 adds two further checks: 8.1 applies Forbes & Rigobon's (2002) standard heteroskedasticity-bias correction directly to this paper's own data for the first time, finding that it reverses the effect's sign rather than resolving the estimator-sensitivity problem -- a result traced to the correction's documented unreliability under large cross-regime variance ratios (Corsetti, Pericoli & Sbracia, 2005) rather than to genuinely falling dependence -- and 8.2 shows the naive effect's direction, though not its magnitude, replicates across an independent sample split. Section 7 argues that, taken together, this evidence favours genuine intensification -- on the strength of crisis-period clustering and the bias-immune Bayesian volatility-innovation correlation -- while correlation-based point estimates of its size, naive or bias-corrected, remain too estimator-sensitive to report in isolation. Remaining work is to propagate MSV posterior uncertainty jointly into the regime-detection stage (rather than conditioning on the posterior mean path, as done throughout Section 6), and to produce the regime states and engineered features intended to feed the temporal graph neural network architecture described in Section 10.

## **Supplementary Materials**

The following are available online: Appendix A (Derivations and Algorithms), Appendix B (Formal Derivations, Convergence Diagnostics, and a Worked Example), Appendix C (Full Posterior Parameter Tables and Computational Notes), Appendix D (Notation Summary), and Appendix E (Synthetic-Data Validation Results), collected in a single supplementary file accompanying this manuscript. Equation, table, and section numbering in the supplementary file follows the main text exactly.

## **Author Contributions**

Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing (original draft, review and editing), and visualization: J.C. The author has read and agreed to the published version of the manuscript.

## **Funding**

This research received no external funding.

## **Data Availability Statement**

The S&P 500 and Nifty 50 daily price series underlying this study are publicly available from Yahoo Finance. The analysis code used to produce every estimate, table, and figure in this paper is available from the author upon reasonable request.

## Conflicts of Interest

The author declares no conflict of interest.

## 13. References

- Agarwal, A., Dhankhar, N., & Mehla, S. (2024). Interconnectedness and spillover effects amongst stock markets of the US, China, Germany, Japan and India using DCC-GARCH model and Diebold-Yilmaz method. *Colombo Business Journal*, 15(2), 162-189.
- Amirshahi, B., & Lahmiri, S. (2023). Hybrid deep learning and GARCH-family models for forecasting volatility of cryptocurrencies. *Machine Learning with Applications*, 12, Article 100465.
- Ando, T., Greenwood-Nimmo, M., & Shin, Y. (2022). Quantile connectedness: Modeling tail behavior in the topology of financial networks. *Management Science*, 68(4), 2401-2431.
- Ang, A., & Bekaert, G. (2002). International asset allocation with regime shifts. *The Review of Financial Studies*, 15(4), 1137-1187.
- Araya, H.T., Aduda, J., Berhane, T., & Wong, M.L. (2024). A hybrid GARCH and deep learning method for volatility prediction. *Journal of Applied Mathematics*, 2024, Article 6305525.
- Beal, M.J., Ghahramani, Z., & Rasmussen, C.E. (2002). The infinite hidden Markov model. *Advances in Neural Information Processing Systems*, 14.
- Bekaert, G., Harvey, C.R., & Ng, A. (2005). Market integration and contagion. *The Journal of Business*, 78(1), 39-70.
- Ben Romdhane, W., & Boubaker, H. (2026). A hybrid HAR-LSTM-GARCH model for forecasting volatility in energy markets. *Journal of Risk and Financial Management*, 19(1), Article 77.
- Bhardwaj, N., Miklosevic, I., & Gambhir, E. (2026). Regime-dependent integration, connectedness and contagion between India and global equity markets. *International Journal of Financial Studies*, 14(8), 210.
- Bitto, A., & Fruhwirth-Schnatter, S. (2019). Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics*, 210(1), 75-97.
- Boukardagha, A. (2026). Explainable regime-aware investing. Working Paper, arXiv:2603.04441.
- Bulla, J., & Bulla, I. (2006). Stylized facts of financial time series and hidden semi-Markov models. *Computational Statistics & Data Analysis*, 51(4), 2192-2209.
- Carter, C.K., & Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81(3), 541-553.
- Christopher, J. (2026). Regime-specific volatility dynamics and cross-market contagion: Evidence from S&P 500 and Nifty 50, 2007-2024. Unpublished working paper.
- Chung, V., Espinoza, J., & Quispe, R. (2025). Forecasting financial volatility under structural breaks: A comparative study of GARCH models and deep learning techniques. *Journal of Risk and Financial Management*, 18(9), Article 494.
- Citterio, M., D'Errico, M., & Visentin, G. (2024). Conditional forecasting of margin calls using dynamic graph neural networks. arXiv:2410.23275.
- Corsetti, G., Pericoli, M., & Sbracia, M. (2005). 'Some contagion, some interdependence': More pitfalls in tests of financial contagion. *Journal of International Money and Finance*, 24(8), 1177-1199.

- Davidescu, A.A., Manta, E.M., Hapau, R.G., Gruiescu, M., & Vacaru (Boita), O.M. (2023). Exploring the contagion effect from developed to emerging CEE financial markets. *Mathematics*, 11(3), 666.
- Diebold, F.X., & Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253-263.
- Diebold, F.X., & Yilmaz, K. (2012). Better to give than to receive: Predictive directional measurement of volatility spillovers. *International Journal of Forecasting*, 28(1), 57-66.
- Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987-1007.
- Engle, R.F. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3), 339-350.
- Feng, F., He, X., Wang, X., Luo, C., Liu, Y., & Chua, T-S. (2019). Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems*, 37(2), Article 27.
- Forbes, K.J., & Rigobon, R. (2002). No contagion, only interdependence: Measuring stock market comovements. *The Journal of Finance*, 57(5), 2223-2261.
- Fox, E.B., Sudderth, E.B., Jordan, M.I., & Willsky, A.S. (2011). A sticky HDP-HMM with application to speaker diarization. *Annals of Applied Statistics*, 5(2A), 1020-1056.
- Garcia-Medina, A., & Aguayo-Moreno, E. (2024). LSTM-GARCH hybrid model for the prediction of volatility in cryptocurrency portfolios. *Computational Economics*, 63(4), 1511-1542.
- Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian Statistics 4* (pp. 169-193). Oxford University Press.
- Geyer, C.J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 7(4), 473-483.
- Gonon, L., Meyer-Brandis, T., & Weber, N. (2024). Computing systemic risk measures with graph neural networks. *arXiv:2410.07222*.
- Gupta, R., Kapoor, S., Gupta, H., & Natesan, S. (2025). A forest of opinions: A multi-model ensemble-HMM voting framework for market regime shift detection and trading. *Data Science in Finance and Economics*, 5(4), 466-501.
- Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2), 357-384.
- Ishwaran, H., & James, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453), 161-173.
- Islam, M.Z., Sumsuzoha, M., Islam, M.R., Kawsar, M., Mithu, M.F.H., Pant, S., Hossain, M.N., & Al Helal, M.A. (2026). Graph neural networks for systemic financial risk forecasting: Modeling cross-market contagion between banking systems and cryptocurrency markets. *International Journal of Accounting and Economics Studies*, 13(2), 331-342.
- Jain, S., & Neal, R.M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1), 158-182.
- Jin, X., & Maheu, J.M. (2016). Bayesian semiparametric modeling of realized covariance matrices. *Journal of Econometrics*, 192(1), 19-39.

- Kakade, K., Jain, I., & Mishra, A.K. (2022). Value-at-Risk forecasting: A hybrid ensemble learning GARCH-LSTM based approach. *Resources Policy*, 78, Article 102903.
- Kalli, M., Griffin, J.E., & Walker, S.G. (2011). Slice sampling mixture models. *Statistics and Computing*, 21(1), 93-105.
- Kastner, G., & Fruhwirth-Schnatter, S. (2014). Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. *Computational Statistics & Data Analysis*, 76, 408-423.
- Kastner, G., Fruhwirth-Schnatter, S., & Lopes, H.F. (2017). Efficient Bayesian inference for multivariate factor stochastic volatility models. *Journal of Computational and Graphical Statistics*, 26(4), 905-917.
- Kim, S., Shephard, N., & Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies*, 65(3), 361-393.
- Kumar, P.N., Umeorah, N., & Alochukwu, A. (2026). Regime-dependent graph neural networks for enhanced volatility prediction in financial markets. *Mathematics*, 14(2), 289.
- Li, C., Maheu, J.M., & Yang, Q. (2024). An infinite hidden Markov model with stochastic volatility. *Journal of Forecasting*, 43(6), 2187-2211.
- Li, C., & Yang, Q. (2025). An infinite hidden Markov model with GARCH for short-term interest rates. *Finance Research Letters*, 80, Article 107294.
- Ljung, G.M., & Box, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2), 297-303.
- Maheu, J.M., & Yang, Q. (2016). An infinite hidden Markov model for short-term interest rates. *Journal of Empirical Finance*, 38(Part A), 202-220.
- Matsunaga, D., Suzumura, T., & Takahashi, T. (2019). Exploring graph neural networks for stock market predictions with rolling window analysis. *arXiv:1909.10660*.
- McCausland, W.J., Miller, S., & Pelletier, D. (2021). Multivariate stochastic volatility using the HESSIAN method. *Econometrics and Statistics*, 17, 76-94.
- Nasri, B.R., Remillard, B.N., & Thioub, M.Y. (2024). Are information criteria good enough to choose the right number of regimes in hidden Markov models? *Journal of Statistical Computation and Simulation*, 94(18), 4107-4127.
- Nguyen, N., & Nguyen, D. (2015). Hidden Markov model for stock selection. *Risks*, 3(4), 455-473.
- Nguyen, T.N., Phan, T.K.H., Nguyen, T.L., & McMillan, D. (2022). Financial contagion during global financial crisis and COVID-19 pandemic: The evidence from DCC-GARCH model. *Cogent Economics & Finance*, 10(1), 2051824.
- Nsengiyumva, E., Mung'atu, J.K., & Ruranga, C. (2025). Hybrid GARCH-LSTM forecasting for foreign exchange risk. *FinTech*, 4(2), Article 22.
- Omori, Y., Chib, S., Shephard, N., & Nakajima, J. (2007). Stochastic volatility with leverage: Fast and efficient likelihood inference. *Journal of Econometrics*, 140(2), 425-449.
- Papaspiliopoulos, O., & Roberts, G.O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1), 169-186.
- Patton, A.J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246-256.

- Pericoli, M., & Sbracia, M. (2003). A primer on financial contagion. *Journal of Economic Surveys*, 17(4), 571-608.
- Petrosino, L., Bacco, L., Salvati, G., Merone, M., & Papi, M. (2025). A GARCH-temporal fusion transformer model for the volatility prediction of exchange traded funds. *Neural Computing and Applications*, 37(26), 21435-21458.
- Pitt, M.K., & Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446), 590-599.
- Qian, H., Zhou, H., Zhao, Q., Chen, H., Yao, H., Wang, J., Liu, Z., Yu, F., Zhang, Z., & Zhou, J. (2024). MDGNN: Multi-relational dynamic graph neural network for comprehensive and dynamic stock investment prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13), 14642-14650.
- Rigobon, R. (2019). Contagion, spillover, and interdependence. *Economía*, 19(2), 69-99.
- Rodriguez, J.C. (2007). Measuring financial contagion: A copula approach. *Journal of Empirical Finance*, 14(3), 401-423.
- Sanchez-Fernandez, A., Diez-Gonzalez, J., Huerga-Perez, N., & Perez, H. (2026). Multivariate regime identification and prediction in financial markets via Gaussian mixture and gradient boosting methods. In *Hybrid Artificial Intelligent Systems (HAIS 2025)*, Lecture Notes in Computer Science (Vol. 16203, pp. 340-353). Springer.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2), 639-650.
- Shi, S., & Song, Y. (2016). Identifying speculative bubbles using an infinite hidden Markov model. *Journal of Financial Econometrics*, 14(1), 159-184.
- Song, Y. (2014). Modelling regime switching and structural breaks with an infinite hidden Markov model. *Journal of Applied Econometrics*, 29(5), 825-842.
- Tampouris, A., & Dritsaki, C. (2026). Adaptive hierarchical hidden Markov models for structural market change. *Journal of Risk and Financial Management*, 19(1), Article 15.
- Taneva-Angelova, G., & Granchev, D. (2025). Deep learning and transformer architectures for volatility forecasting: Evidence from U.S. equity indices. *Journal of Risk and Financial Management*, 18(12), 685.
- Teh, Y.W., Jordan, M.I., Beal, M.J., & Blei, D.M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 1566-1581.
- Thangamuthu, M., Maheshwari, S., & Naik, D.R. (2022). Volatility spillover effects during pre-and-post COVID-19 outbreak on Indian market from the USA, China, Japan, Germany, and Australia. *Journal of Risk and Financial Management*, 15(9), 378.
- Tiwari, A.K., Jena, S.K., Trabelsi, N., & Hammoudeh, S. (2022). Conditional transmission of global shocks to emerging stock markets: Evidence from the quantile connectedness network analysis. *Applied Economics*, 54(31), 3621-3634.
- Uddin, A., Tao, X., & Yu, D. (2023). Attention based dynamic graph neural network for asset pricing. *Global Finance Journal*, 58, Article 100900.
- Van Gael, J., Saatci, Y., Teh, Y.W., & Ghahramani, Z. (2008). Beam sampling for the infinite hidden Markov model. *Proceedings of the 25th International Conference on Machine Learning*.
- Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36(1), 45-54.

- Xiang, S., Cheng, D., Shang, C., Zhang, Y., & Liang, Y. (2022). Temporal and heterogeneous graph neural network for financial time series prediction. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)*, 3584-3593.
- Xu, Z., Liechty, J., Benthall, S., Skar-Gislinge, N., & McComb, C. (2024). GARCH-informed neural networks for volatility prediction in financial markets. *Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF '24)*.
- Yu, G., & Luo, Z. (2025). Dynamic graph neural networks for financial crisis modeling and blocking in large-scale enterprise groups. *Intelligent Decision Technologies*, 19(6), 3668-3691.
- Zhou, Y., Xie, C., Wang, G-J., Gong, J., & Zhu, Y. (2025). Forecasting cryptocurrency volatility: A novel framework based on the evolving multiscale graph neural network. *Financial Innovation*, 11(1), Article 87.
- Zolotareva, E. (2021). Aiding long-term investment decisions with XGBoost machine learning model. In *Artificial Intelligence and Soft Computing (ICAISC 2021)*, *Lecture Notes in Computer Science* (Vol. 12855, pp. 414-427). Springer.