Persistent Homology and Stochastic Dynamics: Stability in Financial Modeling By Hassan Shahid

Abstract

A rolling cross-asset correlation matrix can be embedded in a Euclidean metric space and analyzed through persistent homology. A topological summary may then enter the drift or diffusion of a stochastic price model as a lagged, observable input. This manuscript develops that construction in detail and asks two deliberately separate questions: how sensitive is the resulting model to errors in the observed dependence structure, and does topology add genuine forecasting information? We address the former analytically; the latter remains an empirical hypothesis.

For standardized return vectors, the chordal correlation distance is $d_{i□} = \sqrt{[2(1 − \rho_{i□})]}$. A Vietoris–Rips filtration yields a persistence diagram, whose longest finite H_1 lifetime is a simple scalar feature. On a common asset universe, bounded perturbations of the distance matrix imply bounded changes in the diagram and its longest lifetime. When this feature conditions predictable Lipschitz coefficients in a log-price Itô process, a synchronous-coupling estimate controls expected squared model-path deviations. We give explicit constants, discuss how correlation estimation errors propagate nonlinearly, and calculate a four-asset example under shrinkage exactly.

We additionally distinguish the topology of empirical asset geometry from the Riemannian geometry of stochastic-volatility diffusions such as SABR. The research design specifies null controls, as-of data rules, forecasting baselines, walk-forward tests, and criteria for rejecting the model. None of the propositions is presented as a newly discovered general theorem. No trading profits, fitted coefficients, statistical significance, or crash predictions are claimed.

Keywords: persistent homology; Vietoris–Rips complexes; correlation matrices; stochastic differential equations; model stability; probabilistic forecasting; SABR.

AUTHOR’S EVIDENCE NOTE

All calculated examples in this manuscript are analytic or proposed simulations. The manuscript is suitable for further technical review, not yet for claims of demonstrated financial performance.

1. Introduction

Financial time series have at least two kinds of structure. Individual returns exhibit volatility clustering, tail behavior, and temporal dependence. Collections of assets also have changing relations: their correlations, factor exposures, and conditional co-movements evolve. The first is the conventional domain of univariate stochastic dynamics; the second is naturally represented by matrices and networks. Persistent homology offers a multiscale way to compress some of the latter structure.

The motivating hypothesis is restricted: a summary of historical dependence geometry may help condition a forecast of future uncertainty. It need not forecast the sign of a return or identify a crisis. Financially interesting features can be useful for conditional variance, density calibration, or stress classification without having a profitable directional trading strategy. Conversely, a feature can accurately describe historical turmoil and have no incremental out-of-sample value.

The proposed research program has three layers. The observation layer maps returns observed by a forecast origin t to a positive-semidefinite correlation matrix R and then to a labeled finite distance space (V,d). The topological layer maps this metric to a Vietoris–Rips persistence diagram D_1, and a scalar feature Z. The stochastic layer uses the lagged and predictable process Z in model coefficients. This ordering is essential: no future return is allowed to influence a past coefficient.

Three contributions follow. First, the construction is reproducible and mathematically well-defined, including degenerate data and changing universes. Second, its robustness claim is expressed as a chain of inequalities rather than an intuition about noise tolerance. Third, the empirical hypothesis is subjected to comparison with simple and strong baselines, negative controls, and formal uncertainty assessment.

The word “topological” is not a synonym for “nonlinear” or “causal.” Homology measures features of a chosen filtration; a loop is not necessarily an economic cycle. Topological summaries depend on which objects constitute the point cloud, the distance definition, the observation window, and preprocessing. None is imposed by the market itself.

Throughout, propositions concern a mathematical model under stated assumptions. Statements about real financial data are research questions unless accompanied by independently reproducible evidence. This distinction is particularly important for a young research field in which attractive visual indicators can be mistaken for forecasts.

2. Prior Work and Novelty Boundary

The stability of persistence diagrams was established in foundational work by Cohen-Steiner, Edelsbrunner, and Harer [1]. Bubenik [2] later introduced persistence landscapes as functional summaries suitable for statistical analysis. These tools precede this manuscript and provide the theoretical background for its elementary estimates.

Finance applications also predate this paper. Gidea and Katz [3] studied persistence landscapes of major United States stock-market indices near historical crashes; Gidea and colleagues [4] examined critical transitions in cryptocurrency time series. De Jesus, Fernández-Navarro, and Carbonero-Ruz [5] incorporated persistent-homology features into an N-BEATS forecasting framework and reported favorable results in their evaluated instruments and periods. Those papers use different data objects, time scales, feature maps, and validation targets; their results cannot be transferred automatically to a cross-asset correlation-metric model.

More recent work illustrates why independent testing matters. Guritanu, Barbierato, and Gatti [6] reported a causal warning framework for selected equity-market crises. A 2026 working paper by Sajinesh [7] instead reported little dependable incremental out-of-sample benefit from topological indicators over elementary volatility and correlation measures in its tested exchange-traded-fund universe. The latter is a preprint, and its findings are not a universal refutation; likewise, favorable studies are not universal validation. Contradictory designs motivate a prespecified comparison.

This manuscript therefore does not claim to invent TDA in finance, stochastic volatility, diffusion geometry, or the Lipschitz stability of Itô equations. Its contribution is a carefully bounded synthesis: a transparent choice of asset-based filtration, a feature-to-SDE perturbation chain, an exactly solvable shrinkage illustration, and a falsifiable evaluation specification.

A genuine new theorem would require materially new assumptions, a sharp or nontrivial bound, a new estimator theory, or a coupled endogenous model not covered by standard arguments. A genuine empirical result would require data, executable methods, archived forecasts, and uncertainty estimates. Neither should be asserted merely because a derivation or architecture is technically elaborate.

The testable claim left open is conditional incremental information: after using realized volatility, mean dependence, and major factors, does Z_t improve a prespecified forecast loss on genuinely unseen time periods? The null answer is scientifically meaningful.

3. Data, Information Sets, and Timings

Let $P_{i,t} > 0$ be an adjusted price for asset i at trading timestamp t. Define $r_{i,t} = \log(P_{i,t}/P_{i,t-1})$. Prices, corporate actions, calendar alignments, exchange holidays, and vendor revisions must be associated with their actual availability timestamps. For a panel of n assets and w historical return observations, write $X_t \in \mathbb{R}^{w \times n}$ for the trailing matrix ending at the observable cutoff t_c.

All inputs to a decision made after t_c must be measurable with respect to the information set $\mathcal{F}_{t_c}$. A rolling statistic from observations through the close at t_c can support a post-close decision for the next interval, but cannot support an order purportedly executed earlier that afternoon. This includes preprocessing parameters: window length, shrinkage strength, imputation rules, and model parameters must be fixed or learned entirely within the available training history.

Define a forecast origin t_o, an information cutoff $c_o \le t_o$, and a target interval $(t_o, t_{o+\square}]$. The distinction allows explicit accounting for data publication delay and execution latency. Topological feature Z_o uses observations no later than c_o; the outcome $Y_{\square,\square}$ is measured only afterward. If targets overlap across examples, adjacent train and test records share future returns and should be separated or purged accordingly.

The asset universe V_o must also be chosen as of the date, not reconstructed using today's surviving constituents. A fixed ex ante basket simplifies mathematical comparisons but may introduce missing-price periods. A historical index universe is economically natural but requires point-in-time membership data, a clear rule for entries and exits, and a nontrivial correspondence when comparing filtrations on different vertices.

Missingness is not innocent. Pairwise-complete correlations computed from unequal samples need not form a positive-semidefinite matrix. A complete-case window, a documented positive-semidefinite projection, or an explicit latent model is preferable to silently treating arbitrary pairwise correlations as Euclidean inner products. Stale prices can also create artificial perfect correlations or zero returns.

A valid experiment archives the dataset version, symbols, identifiers, split calendar, preprocessing configuration, and a hash or immutable description of each generated feature panel. Reproducibility starts with knowing exactly what was observable when the forecast was made.

4. From Correlations to a Finite Metric Space

Within a trailing window, center the returns of asset i and normalize the centered vector to Euclidean length one. Denote the result by $u_i \in \mathbb{R}^w$. Exclude zero-variance vectors or process them under a declared rule. The empirical Pearson correlation is $\rho_{ij} = \langle u_i, u_j \rangle$. Since $R = U^T U$, a complete-window sample correlation matrix is positive semidefinite with unit diagonal.

Define the chordal distance.

$$d_{ij} = \|u_i - u_j\|_2 = \sqrt{[2(1 - \rho_{ij})]}. \qquad (1)$$

This formula is not an arbitrary graph transform. Its Euclidean representation guarantees symmetry and the triangle inequality. Distinct asset labels can have identical normalized return vectors, making $d_{ij} = 0$ even when $i \neq j$; the labeled object is then a pseudometric. Identifying such points yields a metric, or the finite pseudometric can be handled explicitly by the persistence software.

Do not substitute $d = 1 - \rho$ without checking what changes: that expression is not generally the Euclidean chordal metric and may violate the triangle inequality. Likewise, converting negative correlations to absolute values can erase economically meaningful sign differences. Different choices generate different filtrations.

Because u_i lies on a unit sphere in the centered-return subspace, $0 \leq d_{ij} \leq 2$. All pairwise distances can be calculated without a low-dimensional visualization. A two-dimensional projection may be useful for illustration but may distort distances and therefore should not silently replace the actual metric.

The sample covariance of n assets with w observations has rank at most $w - 1$. When n is large relative to w, this is an important numerical constraint even though the Gram construction remains positive semidefinite. Covariance shrinkage and other regularization can stabilize estimates [8], but they also alter distances and topological lifetimes; the choice must be evaluated as part of the model.

Alternative financial TDA pipelines treat observation dates as points in a delay-embedded return space rather than treating assets as points. Neither construction is inherently correct. The scientific interpretation changes: asset topology measures the geometry of co-movement patterns, whereas time-embedded topology describes recurrent shapes in a sequence of market states. This manuscript commits to the former for the main proofs.

5. Vietoris-Rips Filtration and Features

Let $V = \{1,\ldots,n\}$ be the fixed set of asset vertices at a forecast origin and d its correlation-derived pseudometric. For scale $\varepsilon \geq 0$, define the Vietoris–Rips complex.

$$VR\varepsilon(V,d) = \{\sigma \subseteq V : \max_{i,\square \in \sigma} d_{i\square} \leq \varepsilon\}. \qquad (2)$$

Every vertex enters at scale zero; an edge enters when its distance is no larger than ε; a triangle enters when all three edges have entered. As ε increases, the complexes form a nested filtration. Over a chosen coefficient field, such as $\mathbb{F}_2$, we compute homology in dimensions zero and one. H_0 records the mergers of connected components, while H_1 records independent one-dimensional cycles prior to their filling by two-simplices.

A persistence diagram $D_1(d)$ is a multiset of birth–death pairs (b,d) for finite H_1 intervals. Its lifetime is $\ell = d - b$. Define the longest finite H_1 lifetime by

$$L_1(D) = \max\{d - b : (b,d) \in D, b < d\}, \text{ with } L_1(\varnothing)=0. \qquad (3)$$

The feature is interpretable and admits a simple stability estimate, but it is lossy. Two radically different networks may share the same longest bar. A diagram with no nonzero H_1 bar receives zero even if its H_0 behavior is economically unusual. Ties and zero-length intervals are harmless if conventions are fixed.

Complementary summaries include: the number of components at selected scales; the area under a Betti curve; quantiles of persistence lifetimes; total persistence after a specified truncation; and Bubenik's persistence landscapes [2]. Not all summaries have the same Lipschitz constants or statistical behavior. Unbounded counts can jump under arbitrarily small perturbations, especially near the diagonal. Each proposed replacement needs its own mathematical analysis and multiple-testing control.

The numerical filtration must declare whether ε denotes edge diameter or a radius parameter. A factor-of-two discrepancy between conventions can propagate into every reported birth and death. Essential H_0 classes require a separate convention; we do not put their infinite lifetimes into L_1.

For n vertices, the full abstract clique complex can contain $2^n - 1$ nonempty simplices, so the direct calculation quickly becomes impractical. A practical implementation should specify the largest homology dimension, maximal filtration scale, numerical precision, and any approximation. Approximate diagrams should carry their own error tolerance before the theoretical bounds are applied.

6. Two Elementary Stability Estimates

For persistence diagrams D and D′, let d_B(D, D′) denote bottleneck distance under the L∞ norm, allowing points to match the diagonal. The diagonal match cost of (b,d) is (d − b)/2. Consider finite diagrams without infinite off-diagonal bars.

Proposition 1 (longest-lifetime stability). Under the definition (3),

$$|L_1(D) - L_1(D')| \le 2d_B(D,D'). \qquad (4)$$

Proof. Choose $\eta > d_B(D, D')$ and a matching of cost below η. A point (b,d) matched to (b′,d′) has lifetime difference no larger than $|b-b'|+|d-d'| < 2\eta$. A point matched to the diagonal has lifetime below 2η. Therefore $L_1(D) \le L_1(D')+2\eta$. Reverse the diagrams and let η decrease to d_B to obtain (4). ∎

Proposition 2 (fixed-vertex Rips stability). If two finite labeled distance systems d and d′ on the same V satisfy $\delta = \max_{i,\square} |d_{i\square} - d'_{i\square}|$, then

$$d_B(D_1(d),D_1(d')) \le \delta. \qquad (5)$$

Proof. If every edge of a simplex has d-length at most ε, the corresponding edges under d′ have length at most ε+δ. Thus $VR_\varepsilon(d) \subseteq VR_{\varepsilon+\delta}(d')$, and the reverse inclusion holds after the same shift. The resulting δ-interleaving of finite persistence modules yields (5), using standard algebraic persistence stability [1]. ∎

Corollary 1. With $Z=L_1(D_1(d))$ and $Z'=L_1(D_1(d'))$,

$$|Z - Z'| \le 2\delta. \qquad (6)$$

This is a deterministic bound, not a probability statement. It protects the numerical feature from limited perturbations of a fixed labeled distance matrix. It does not say that either matrix estimates the true population dependence well, that Z has an economic interpretation, or that it predicts future uncertainty. If universes differ, one needs a correspondence-based argument, often involving a Gromov–Hausdorff or related distance rather than entrywise comparison.

The factor two in Proposition 1 arises because both birth and death can move. These estimates should be attributed to ordinary stability theory rather than advertised as new theorems.

7. Correlation Error and its Nonlinear Propagation

The distance $d_{i\square} = \sqrt{[2(1-\rho_{i\square})]}$ is continuous in each correlation coefficient, but its slope is unbounded near ρ=1. Therefore, it is misleading to convert a correlation estimation tolerance directly into the same numerical tolerance for distances.

Let R and R′ be two valid correlation matrices on the same aligned assets. Write $e = \max_{i \neq j} |\rho_{ij} - \rho'_{ij}|$. For nonnegative a,b, $|\sqrt{a} - \sqrt{b}| \leq \sqrt{|a-b|}$. Applying this inequality yields

$$\max_{i,j} |d_{ij} - d'_{ij}| \leq \sqrt{(2e)}, \text{ and } |L_1(D_1(d)) - L_1(D_1(d'))| \leq 2\sqrt{(2e)}. \quad (7)$$

This is a Hölder-type bound: a small correlation error does imply a small persistence-feature error, but not generally at a uniform linear rate. It applies to any two valid matrices that admit the chordal construction, including regularized Gram correlations.

A stronger bound is possible away from exact duplication. Suppose $1 - \rho_{ij} \geq \kappa$ and $1 - \rho'_{ij} \geq \kappa$ for all distinct vertices, with $\kappa > 0$. Since the derivative of $f(r) = \sqrt{[2(1-r)]}$ has magnitude $1/\sqrt{[2(1-r)]}$, the mean-value theorem gives

$$\max |d_{ij} - d'_{ij}| \leq e/\sqrt{(2\kappa)}, \text{ and } |Z - Z'| \leq \sqrt{(2/\kappa)}e. \quad (8)$$

The κ-dependent constant reveals the failure mode: near-perfectly correlated assets can magnify error in the distance scale. That does not invalidate persistence stability; it identifies an upstream estimation sensitivity.

An analyst should therefore report perturbations in both correlation and actual distance units. If distances are computed after a positive-semidefinite projection or shrinkage, the total estimation error must include the effect of that transformation. A bootstrap distribution of Z should not be called a confidence interval without defending the resampling scheme, especially when return observations are serially dependent.

Equations (7) and (8) explain why geometric robustness is conditional and why seemingly tiny changes in a dependence estimator may visibly alter a topology plot. They also guide an experiment: record the empirical discrepancy $\|R - R'\|_\infty$ and verify that the observed feature difference respects the analytic bound.

8. An Exact Four-Asset Topological Example

Consider centered orthonormal vectors u and v in a return-observation space of dimension at least three. Let four standardized asset histories be $x_1 = u$, $x_2 = v$, $x_3 = -u$, $x_4 = -v$. Their Gram correlation matrix in cyclic order is

$$R = [[1,0,-1,0],[0,1,0,-1],[-1,0,1,0],[0,-1,0,1]]. \quad (9)$$

This matrix is positive semidefinite with eigenvalues 2,2,0,0. Adjacent vertices in the cycle have zero correlation and chordal distance √2. Opposite vertices have correlation −1 and distance 2. At scale $\varepsilon = \sqrt{2}$, four edges form a square, but neither diagonal is present; H_1 has one class. At ε=,2 both diagonals and the triangles enter, filling the class. Hence the single positive-length interval is (√2,2), giving.

$$L_1 = 2 - \sqrt{2} \approx 0.585786. \quad (10)$$

Now regularize the correlation matrix toward the identity:

$$R\lambda = (1-\lambda)R + \lambda I, \qquad 0 \le \lambda \le 1. \tag{11}$$

The regularized matrix is positive semidefinite with unit diagonal. Adjacent correlations remain zero, while opposite correlations become $-(1-\lambda)$. The square edges retain distance $\sqrt{2}$; the diagonal distance becomes $\sqrt{(4-2\lambda)}$. The resulting lifetime is exactly.

$$L_1(\lambda) = \sqrt{(4-2\lambda)} - \sqrt{2} \quad \text{for } 0 \le \lambda \le 1. \tag{12}$$

The function declines from $2-\sqrt{2}$ to zero. For $\lambda<1$, its derivative is $-1/\sqrt{(4-2\lambda)}$. At $\lambda=1$,1 all edges enter simultaneously, ly and there is no positive-length H_1 interval. Shrinkage has erased a feature even though the underlying standardized vectors in the original construction have not changed.

This demonstration serves three purposes. First, the filtration and lifetime can be checked without numerical software. Second, it verifies that a strong one-dimensional persistent feature is compatible with an entirely synthetic Gram configuration. Third, it shows a concrete confounder: changes in estimator regularization can masquerade as changing market geometry.

The example is not a simulated trading result. Its vectors were constructed to produce a loop; nothing about this calculation establishes predictive information in real financial returns.

9. Statistical Uncertainty and Confounding

A finite persistence diagram is a deterministic function of its chosen data. The data, however, are statistical estimates. Sampling noise, time-varying observation quality, outliers, common factors, and window overlap all affect the correlation matrix before topology is computed. The inequality in Section 6 is not a substitute for a finite-sample error analysis of those inputs.

Consider a simple one-factor return model $r_{i,t} = \beta_i f_t + \varepsilon_{i,t}$, with idiosyncratic errors independent across assets conditional on the factor. Changes in factor variance can alter correlations across many pairs simultaneously. The resulting topology might respond to a change in market-wide dependence rather than to a new independent cycle or hidden mechanism. In that case, average correlation, the leading eigenvalue of R, or the first principal-component variance share may contain essentially the same information as L_1.

Another failure arises when w is small relative to n. Sample R has low rank; different configurations of nearly collinear standardized return vectors can be highly unstable under new observations. Pairwise deletion can produce a non-Gram matrix; aggressive shrinkage can suppress bars; nonstationarity can make a trailing window an inconsistent summary of the coming interval. The model must compare these alternatives rather than describe every moving diagram as economic discovery.

Robustness analyses should include at least three estimators: ordinary complete-panel sample correlation, a prespecified shrinkage estimator, and a robust or rank-based dependence estimate when justified. Each produces a distinct geometry. Results should be displayed as sensitivity to the estimator, rather than averaging incompatible outputs without disclosure.

Useful perturbation experiments are: a controlled single-entry correlation change followed by valid positive-semidefinite reconstruction; block-bootstrap resampling of trailing returns; removal of one asset at a time; and injection of missing observations or stale prices. The last two generally violate the fixed-vertex hypothesis and must be interpreted as separate stress tests, not violations of a theorem that never applied.

The appropriate statistical question is conditional: does a topology feature retain variation and forecast usefulness after the information already captured by common factors, volatility, correlation concentration, and sampling uncertainty is accounted for? If not, the added algebraic machinery is scientifically interesting but not justified as a forecasting component.

10. Causal Topological Features in Continuous Time

Let $(\Omega, F,(F\square)\square\geq_0, P)$ be a filtered probability space satisfying the usual conditions and supporting a Brownian motion W. Fix observation times $0=t_0<t_1<\cdots$. At $t\square$, calculate $Z\square=L_1(D_1(d\square))$ using only the completed return window and information available by that time. Define the predictable step input.

$$Z\square = Z\square \text{ for } t\in(t\square,t\square_{+1}], \text{ with an appropriate initial value.} \qquad (13)$$

At every positive interval start, $Z\square$ is $F\square\square$-measurable. This left-continuous step convention admits a predictable version. It captures the actual operational sequence: observe, compute, freeze the feature, produce a forecast for the next interval, then update after additional data arrive.

An important distinction separates three ideas commonly called a “regime.” A descriptive regime is a label derived from the past geometry; a latent statistical regime is an unobserved state inferred with estimation uncertainty; a stochastic regime is a specified random process influencing coefficients. Equation (13) defines an observable covariate process, not proof that the market is governed by discrete topological states.

A continuous threshold model $Z\to1\{Z>c\}$ creates a new mathematical issue. Even when Z changes smoothly under perturbations, the hard threshold is discontinuous, and an arbitrarily small feature error near c can change the selected volatility coefficient. One can use a bounded Lipschitz link, for example $g(z)=\tanh(az)$, or analyze threshold switching using separate assumptions. We use Lipschitz links in the primary theorem.

The random process $Z\square$ is not necessarily statistically independent of W or historical returns. Predictability is the relevant condition for an Itô integral. A feature constructed from the same price process within its current future interval, by contrast, can become anticipative or create feedback that is not covered by the prescribed-coefficient argument.

For a discrete production process, timestamp every input and log when a feature becomes available. The exact boundary between data ingestion and model invocation is part of the mathematical definition, not merely an engineering convenience. An elegant filtration-based proof cannot repair a pipeline that leaks information from a future observation.

11. A Well-Posed Topology-Conditioned Log-price Model

For an illustrative scalar price $S_t>0$, write $X_t=\log S_t$ and consider

$$dX_t = m(t,Z_t)dt + \sigma(t,Z_t)dW_t, \qquad X_0=x_0. \tag{14}$$

Here m and σ are measurable functions of calendar time and the predictable topological input. Assume their compositions with Z are predictable and square-integrable on each finite horizon T:

$$E\int_0^T(|m(t,Z_t)|^2+|\sigma(t,Z_t)|^2)dt < \infty. \tag{15}$$

Then the solution is explicitly

$$X_t = x_0 + \int_0^t m(s,Z_s)ds + \int_0^t \sigma(s,Z_s)dW_s. \tag{16}$$

It is continuous and adapted. The price $S_t=\exp(X_t)$ is strictly positive for every finite t for which X_t is finite. This statement is about a constructed process, not an assertion that actual markets cannot default, gap, or experience trading interruptions.

For an interpretable initial specification, set $m(t,z)=m_0(t)+u(z)$ and $\log \sigma(t,z)=v_0(t)+v(z)$, where u and v are fitted functions. To guarantee global Lipschitz coefficients in z, however, a bounded volatility link or an explicit cap may be needed: exponentiation of an unrestricted linear function is not globally Lipschitz. A simple alternative is σ(t,z)=σ_min+(σ_max−σ_min)/(1+e^(−az−b)), with fixed finite positive bounds.

We deliberately avoid inferring drift from an elaborate feature before showing robust volatility information. Expected returns are difficult to estimate precisely from daily data, while next-period conditional variance and predictive distribution calibration offer more tractable first targets. A research model can hold m constant and allow only the diffusion amplitude to depend on Z.

If state dependence is desired, the broader equation dX=b(t, X, Z)dt+a(t, X, Z)dW has a unique strong solution under the usual progressively measurable, global state-Lipschitz and linear-growth assumptions, with appropriate integrable Z. Its proof uses Picard iteration and stochastic maximal inequalities [9]. That theorem is classical; our explicit calculations below concern (14), where the coefficient processes are already specified after Z is given.

Every equation is under a chosen probability measure P unless otherwise stated. A historical forecasting fit and an arbitrage-free derivative-pricing model are not automatically the same thing.

12. The Topology-to-Stochastic-Path Bound

Assume $m(t,z)$ and $\sigma(t,z)$ are uniformly Lipschitz in z, with constants $L\square$ and $L\sigma$:

$$|m(t,z)-m(t,z')| \le L\square|z-z'|,\ |\sigma(t,z)-\sigma(t,z')| \le L\sigma|z-z'|. \tag{17}$$

Compare admissible predictable features Z and Z' constructed on the same filtered probability space. Let X and X' solve (14) with the same initial value and the same Brownian motion W. This is synchronous coupling: we isolate feature perturbation rather than unrelated simulation noise.

Proposition 3 (explicit log-path sensitivity). For every finite T,

$$E[\sup_{0\le u\le T}|X_u-X'_u|^2] \le C(T)E\int_0^T|Z\square-Z'\square|^2dt,\ C(T)=2TL\square^2+8L\sigma^2. \tag{18}$$

Proof. Subtract the integral solutions (16). The difference is the sum of a drift integral and an Itô integral. Apply $|a+b|^2\le 2a^2+2b^2$. Cauchy–Schwarz gives $\sup_u|\int_0^u\Delta m\,dt|^2 \le T\int_0^T|\Delta m|^2dt$. The L^2 Doob maximal inequality and Itô isometry give $E\sup_u|\int_0^u\Delta\sigma\,dW|^2 \le 4E\int_0^T|\Delta\sigma|^2dt$. Apply the Lipschitz bounds (17) and collect coefficients. ∎

Corollary 2 (distance-to-path stability). Suppose for each interval the feature is the longest finite H_1 lifetime of an aligned asset filtration, and let $\delta\square=\max_{i,\square}|d_i\square,t-d'_i\square, t\}- d\ '_\ \{i, j, t\}\ |$. Then (6) and (18) yield

$$E\sup_{0\le u\le T}|X_u-X'_u|^2 \le 4C(T)\ E\int_0^T\delta\square^2dt. \tag{19}$$

If the corresponding correlation-matrix errors satisfy $e\square=\max_{i\ne\square}|\rho_i\square,t-\rho'_i\square,t|$, then $\delta\square^2\le 2e\square$ and

$$E\sup_{0\le u\le T}|X_u-X'_u|^2 \le 8C(T)\ E\int_0^T e\square dt. \tag{20}$$

All estimates concern two models evaluated with the same Brownian driving noise. They do not bound differences between independently sampled paths or between a model and the real market. The constants quantify sensitivity under assumptions, not predictive accuracy.

13. What the Bound Does and Does Not Imply

The composite estimate has a clear interpretation: a small perturbation of an aligned return-correlation matrix leads to a controlled perturbation of the chordal metric, then the persistence diagram, then a chosen scalar feature, then the log-price model path when the same Brownian sample is used. Each arrow has a specified condition and quantitative cost.

The chain is directional, not reversible. A stable feature may discard most of the information in R. A stable model may be systematically misspecified. Two processes can be close under synchronous coupling without their common distribution resembling the market's actual conditional law. None of these inequalities implies a better forecast, a reliable leading indicator, or a positive trading return.

There is a second important caveat about prices. The estimate controls X=log S. Exponentiation is not globally Lipschitz: | e^x- e^y | can be large when x and y are both large, even if |x−y| is small. To obtain an analogous L^2 bound for $S-S'$, one needs additional exponential-moment or localization assumptions. Claiming a price-path bound directly from (18) without them would be mathematically incorrect.

A finite-time estimate also does not imply uniform robustness as $T\to\infty$. The constant C(T) grows with the horizon, and errors can accumulate. State-dependent models often introduce Grönwall factors that can grow exponentially with T. For deployment, the forecast horizon belongs in the specification and must match the horizon of empirical validation.

Itô's formula illustrates an additional modeling distinction. If S=exp X and (14) holds, then

$$dS_t/S_t = [m(t,Z_t)+\tfrac{1}{2}\sigma(t,Z_t)^2]dt + \sigma(t,Z_t)dW_t. \qquad (21)$$

Under a risk-neutral measure for a tradable asset with suitable regularity, pricing restrictions require the discounted asset to be a martingale (subject to payouts and other conventions). You cannot simply insert the estimated physical drift into derivative valuation for forecasting. A valid change of measure requires explicit assumptions, including integrability of any market price of risk.

A careful paper should therefore reserve the phrase "stability guarantee" for the particular mathematical mapping proved here. It should not use that phrase to describe statistical parameter reliability, market completeness, no-arbitrage pricing, operational trading safety, or economic profitability. Those properties demand different evidence.

14. SABR and the Distinction Between Two Geometries

Stochastic volatility provides an intellectually useful comparison. In the familiar driftless SABR specification, with forward $F_t>0$ and volatility factor $\alpha_t>0$,

$$dF_t = \alpha_t F_t^{\beta} dW_t^1, \quad d\alpha_t = \nu\alpha_t dW_t^2, \ d\langle W^1,W^2\rangle_t = \rho dt, \quad \nu>0, \ |\rho|<1. \qquad (22)$$

The parameter β controls the elasticity of the forward diffusion, ν controls volatility of volatility, and ρ is the instantaneous Brownian correlation. Hagan and coauthors [10] developed important smile approximations in this setting. Boundaries and degenerate parameter cases require separate analysis.

On a positive, nondegenerate state region, the instantaneous diffusion covariance in coordinates (F,α) is

$$A(F,\alpha)=\alpha^2[[F^{2\beta},\rho\nu F^{\beta}],[\rho\nu F^{\beta},\nu^2]]. \qquad (23)$$

Its inverse is a Riemannian metric tensor $g=A^{-1}$ for the principal diffusion covariance. This is a metric on continuous stochastic state variables, not a distance between asset-return histories. The inverse is well defined only when the covariance is positive definite.

The geometry becomes transparent after $q(F)=\int dF/F^{\beta}$, with q=log F for β=1, and coordinates $x=\nu q-\rho\alpha$, $y=\sqrt{(1-\rho^2)}\alpha$. Under the convention $g=A^{-1}$, direct substitution gives

$$ds^2=(dx^2+dy^2)/(\nu^2 y^2). \quad (24)$$

This is a rescaled Poincaré half-plane metric. Its Gaussian curvature is $K=-\nu^2$ and its two-dimensional scalar curvature is $R_{scalar}=-2\nu^2$. In heat-kernel conventions that define the metric as $2A^{-1}$, the curvature scales accordingly; naming the convention prevents a factor-of-two mistake. Henry-Labordère [11] develops the relation between SABR-type diffusions and hyperbolic geometry.

This diffusion-state geometry is mathematically distinct from the finite correlation-metric topology of Sections 4–8. The fact that both use the word "geometry" does not create an identity between them. An actual bridge requires a specified statistical or dynamical map from past persistence features to the conditional distribution of α, ν, ρ, or calibrated implied volatility.

That bridge is a future research direction. We do not claim to have fitted SABR parameters, validated an option-pricing model, or derived an option-implied market signal from persistence.

15. Forecast Identification and Baseline Design

The primary empirical target should be future conditional uncertainty, not an unconstrained attempt to predict the next return sign. For horizon h and forecast origin k, define the future realized variance target $V_{□,□}=\sum_{□=1}^{h} r^2_{□+□}$, with a documented sampling frequency and treatment of overnight returns. The object is observable only after the forecast interval. Its square root is a realized-volatility proxy; it is not identical to the latent instantaneous variance of a continuous-time diffusion.

A simple baseline might forecast $\log(V_{□,□}+\varepsilon)$ using lagged realized variances at daily, weekly, and monthly scales, together with a constant and possibly calendar effects. Stronger prespecified comparisons can include EWMA, GARCH-type volatility models, and the leading principal-component variance share of a rolling correlation matrix. The augmented version receives the same historical inputs plus the lagged $Z_{□}$. Keeping model class and tuning budget comparable isolates the incremental contribution of topology.

Let $\mathcal{L}(F,y)$ be a prespecified forecast loss. For predictive densities, the negative log score or continuous ranked probability score are examples of proper scoring approaches [12]. For positive variance forecasts, a suitable variance loss such as QLIKE may be considered if its assumptions and handling of realized-variance noise are explained. A favorable in-sample coefficient t-statistic is not the primary success criterion.

Define the paired out-of-sample loss improvement for each test origin by $\Delta_{□}=\mathcal{L}(F_{base,k}, Y_{□})-\mathcal{L}(F_{TDA,k}, Y_{□})$. A positive Δ means the augmented model did better on that observation. The principal estimand is the mean improvement over untouched test origins, accompanied by uncertainty accounting for serial dependence and overlapping forecast horizons. A Diebold–Mariano-style comparison [13] may be appropriate under its assumptions; it is not a substitute for a sound experimental split.

An ablation must test whether Z adds anything after mean absolute correlation, the largest eigenvalue, realized volatility, and a simple correlation-clustering summary. If those controls explain its forecast contribution, the interpretation is ordinary dependence concentration, not special topological information.

The null hypothesis is a respectable outcome: Z provides no incremental predictive value under this universe, horizon, data provider, and estimator. Rejection of the null requires more than attractive charts; it requires robust, prespecified loss improvement.

16. Synthetic Experiments and Negative Controls

Before historical evaluation, test the complete pipeline on synthetic data whose structure is known. This is an engineering validation exercise, not a substitute for market evidence. A first simulation uses n assets with returns $r_{i,t}=\beta_i f_t+\varepsilon_{i,t}$. The common factor and idiosyncratic innovations may be generated by Gaussian laws with fixed covariance, and next-period variance is specified independently of Z conditional on the baseline variables. Correctly implemented topology should not consistently deliver incremental out-of-sample value in this negative-control world.
A second simulation varies common-factor variance over time but allows the future target to depend only on past realized volatility and factor concentration. Topology will move, yet a properly controlled augmented forecast should have little or no remaining advantage. This directly tests whether the model mistakes a repackaged covariance statistic for novel information.

A third, artificial positive-control simulation can define σ^2_{t+1} as a bounded smooth function of an explicitly lagged persistence lifetime, plus stochastic noise. Under this constructed data-generating process, the augmented model should recover some information that the baseline deliberately lacks. Success would show the pipeline can detect a designed signal, not that such a signal exists in real markets.

For the deterministic four-asset configuration from Section 8, unit tests should reproduce the birth scale $\sqrt{2}$, death scale 2, lifetime $2-\sqrt{2}$, and shrinkage curve (12) to numerical tolerance. Additional tests should verify symmetry and positive semidefiniteness of R, the triangle inequality for chordal distances, the Rips stability inequality, and numerical convergence of the integrator for a solvable SDE.

Each simulation must declare sample size, asset count, random-number seed, lookback and horizon, regularization, topology library version, and fitting protocol. Repeat across many seeds and report the distribution of results instead of selecting a lucky run. Simulated regimes with jumps, heavy tails, and missing observations should be treated as stress cases where Brownian and complete-panel assumptions may fail.

Leave an explicit results table blank until the executable tests have run. Its columns should include experiment, prespecified metric, seed count, baseline result, augmented result, uncertainty interval, and pass/fail criterion. An empty table is more credible than numbers invented to make the theory look successful.

17. Historical Walk-Forward Evaluation

A complete historical study should predefine the market universe, data source, exchange calendars, adjustment policy, observation window, forecast horizon, training calendar, model classes, and final untouched holdout. Candidate asset universes might include liquid sector exchange-traded funds or a point-in-time equity index panel, but those are design choices rather than claims that this paper has already analyzed either.
At each forecast origin, reconstruct only the returns then available. Estimate R using the active rolling window, transform it to chordal distances, compute D_1 and Z, and save the feature with its cutoff timestamp. Fit the model using past examples alone. Predict the next realized variance or density and archive the forecast before the actual target is revealed. Repeat without retroactive changes to predictions.

Use expanding-window or rolling-origin validation, not random shuffled folds. Reserve a final chronological period that is never used for feature screening or hyperparameter selection. If $Y_{\square}$ uses returns from k+1 through k+h, training and testing samples near a split can overlap in outcome information; enforce a purge or embargo tailored to the target definition.

Track forecast losses and calibration by fold, crisis/noncrisis period, and market regime. A single favorable aggregate can conceal disastrous periods. Report confidence intervals with a method appropriate to temporal dependence, such as a justified block bootstrap or HAC-based estimate. Do not rely on ordinary independent-observation standard errors when the loss series is autocorrelated.

Sensitivity experiments must include window length, n/w ratio, covariance shrinkage, topology feature family, persistence scale truncation, asset universe, missingness treatment, and transaction-time assumptions. All alternatives investigated should be logged. Searching a large model zoo and publishing only the winner demands a data-snooping correction or a completely fresh validation sample; White's reality-check framework [14] is a relevant methodological precedent.

For each final forecast save: model version; preprocessing configuration; feature cutoff; data-provenance identifier; Z; non-topological baseline inputs; fitted parameters; predictive distribution; realized outcome; and loss. A reviewer should be able to reconstruct every evaluation row without querying a live brokerage account.

Only after robust statistical gains have been demonstrated should a separate economic study address a trading rule, position sizing, transaction costs, liquidity, capital requirements, and risk constraints. Forecast improvement is neither equivalent to nor sufficient for profitability.

18. Limitations, Interpretation, and Conclusion

The mathematical analysis has a narrow but useful strength: its assumptions are visible. It concerns finite aligned asset sets, well-defined distance matrices, persistent-homology summaries, predictable feature processes, and Lipschitz stochastic coefficients. If any assumption is dropped, a different theorem or explicit approximation argument is necessary.

Several important limitations remain. Correlations are estimated from finite, often nonstationary samples. Asset-universe turnover breaks simple matrix alignment. A single longest H_1 lifetime may be zero or uninformative in many windows. A continuous Brownian log-price model omits jumps and microstructure. Correlation geometry compresses dependence and does not identify economic causation. Diagram stability measures representation continuity rather than forecast accuracy. Hard regime thresholds can destroy coefficient continuity. Arbitrage-free option pricing requires assumptions absent from an ordinary physical-measure forecast.

Published evidence should be interpreted as setting-specific. Favorable forecasting or warning results [3–6] and cautionary out-of-sample evidence [7] can coexist because the studies differ in object construction, outcome, horizon, and control variables. A sound result for this manuscript requires direct evaluation of this exact pipeline against matched baselines. No claimed percentage accuracy or trading performance is warranted before that exercise.

The synthesis contributes a precise, auditable pathway from data to model. Standardized historical asset returns define a chordal correlation metric. A Vietoris–Rips filtration provides a longest finite H_1 lifetime. The lifetime obeys a bottleneck stability estimate, the diagrams obey a fixed-vertex filtration stability estimate, and the resulting predictable coefficient inputs generate a log-price path with an explicit finite-horizon sensitivity bound. The four-asset example demonstrates the mathematics and the role of shrinkage without implying a market signal.

The decisive scientific question remains unanswered: after conventional volatility and dependence information is used, does persistence improve the predictive distribution of future financial uncertainty? An experiment that finds no additional information would not invalidate the mathematics. It would correctly delimit the model's empirical usefulness.

REFERENCES AND SOURCE VERIFICATION

[1] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. "Stability of Persistence Diagrams." Discrete & Computational Geometry 37 (2007), 103–120. https://doi.org/10.1007/s00454-006-1276-5

[2] P. Bubenik. "Statistical Topological Data Analysis Using Persistence Landscapes." Journal of Machine Learning Research 16(3) (2015), 77–102. https://www.jmlr.org/papers/v16/bubenik15a.html

[3] M. Gidea and Y. Katz. "Topological Data Analysis of Financial Time Series: Landscapes of Crashes." Physica A 491 (2018), 820–834. https://doi.org/10.1016/j.physa.2017.09.028

[4] M. Gidea, D. Goldsmith, Y. Katz, P. Roldan, and Y. Shmalo. "Topological Recognition of Critical Transitions in Time Series of Cryptocurrencies." Physica A 548 (2020), 123843. https://doi.org/10.1016/j.physa.2019.123843

[5] L. C. de Jesus Jr., F. Fernández-Navarro, and M. Carbonero-Ruz. “Enhancing Financial Time Series Forecasting through Topological Data Analysis.” Neural Computing and Applications 37 (2025), 6527–6545. https://doi.org/10.1007/s00521-024-10787-x

[6] E. Guritanu, E. Barbierato, and A. Gatti. “Topological Machine Learning for Financial Crisis Detection.” Computers 14(10) (2025), 408. https://doi.org/10.3390/computers14100408

[7] H. Sajinesh. “Redundant Geometry: Topological Crisis Signals Add No Incremental Out-of-Sample Value …” Working paper, 2026 (not peer-reviewed). https://doi.org/10.2139/ssrn.6966980

[8] O. Ledoit and M. Wolf. “A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices.” Journal of Multivariate Analysis 88(2) (2004), 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4

[9] I. Karatzas and S. E. Shreve. Brownian Motion and Stochastic Calculus, 2nd ed. Springer, 1991. https://doi.org/10.1007/978-1-4612-0949-2

[10] P. S. Hagan, D. Kumar, A. S. Lesniewski, and D. E. Woodward. “Managing Smile Risk.” Wilmott Magazine (September 2002), 84–108.

[11] P. Henry-Labordère. “A General Asymptotic Implied Volatility for Stochastic Volatility Models.” Research paper, 2005. https://ssrn.com/abstract=698601

[12] T. Gneiting and A. E. Raftery. “Strictly Proper Scoring Rules, Prediction, and Estimation.” Journal of the American Statistical Association 102(477) (2007), 359–378. https://doi.org/10.1198/016214506000001437

[13] F. X. Diebold and R. S. Mariano. “Comparing Predictive Accuracy.” Journal of Business & Economic Statistics 13(3) (1995), 253–263. https://doi.org/10.1080/07350015.1995.10524599

[14] H. White. “A Reality Check for Data Snooping.” Econometrica 68(5) (2000), 1097–1126. https://doi.org/10.1111/1468-0262.00152

Source note: References are primary papers, original research manuscripts, or author/publisher records checked for this draft. The 2026 working paper is identified as a preprint; empirical findings from any source are not treated as universal laws.