High Fidelity Synthetic Financial Universes Through Regime Switching With Stochastic Volatility and Jump Diffusion

Arnav Malhotra

December 2025

1 Abstract

Traditional machine learning models trained on financial market data are known to be highly susceptible to overfitting the past market. Historical data is limited, and models trained only on data from the past will only be able to perform well in financial conditions mimicking the past. To combat this data scarcity, many use mathematical generators of unlimited synthetic data. However, traditional synthetic data generators often rely on Geometric Brownian Motion (GBM), which fails to account for empirical phenomena such as volatility clustering, leverage effects, and "black swan" jump events[16]. This paper introduces fsynth, a multi-factor synthetic engine that integrates a Heston-based stochastic volatility model with Merton jump diffusion, governed by a hidden Markov regime-switching process. By using macro-regime states alongside company specific "genes," the engine generates plausible universes containing both price action and fundamental accounting data. We demonstrate that while the engine can be calibrated to replicate conditions of historical benchmarks, its primary utility lies in its parameterizable nature, allowing for the simulation of arbitrary, "out-of-sample" stress scenarios.

2 Introduction

The Efficient Market Hypothesis (EMH) and its mathematical foundation, Geometric Brownian Motion (GBM), assume that price returns are independent and identically distributed (i.i.d.), following a Gaussian distribution[5][10]. However, time and time again these assumptions are consistently inaccurate in reflecting true market data, which most notably exhibit leptokurtosis, volatility clustering, and the leverage effect[3].

In the context of modern financial machine learning, these contradictions impose catastrophic risk on the reliability of models trained on data relying on the EMH[2]. However, training on real historical data that do exhibit these characteristics foists problems of equal calamity. Historical data is scarce, with quality data being expensive and inaccessible for most practitioners. Additionally, historical data presents only one realized possibility of what the market could have been, meaning that machine learning models trained on historical data will only ever be able to perform well in conditions similar to the data trained on.

To address these issues, we created a high fidelity synthetic universe generator, `fsynth`. Unlike standard path generators, `fsynth` utilizes a hybrid stochastic architecture. It employs a mean-reverting stochastic volatility model based on the Heston model for options pricing[9], and a Merton Jump-Diffusion process to simulate high-impact price shocks[12]. Vitally, these parameters are not static; they are regulated by a hidden Markov regime-switching layer[8] that allows the financial universe to switch between distinct states of the market[1].

Furthermore, `fsynth` introduces an innovative method for fundamental analysis in the context of stock-price action. By assigning "Corporate Genes" to synthetic stocks-numerical traits such as leverage tolerance and margin stability-the engine is capable of generating self-consistent financial statements that can dynamically react to the macro-regime. This allows for the training of machine learning models that process both raw price data as well as fundamental quarterly earnings reports in a simulated environment.

3 Mathematical Framework

The `fsynth` engine uses a stochastic architecture in which the overall market state governs the evolution of all asset prices. We define this through a Regime-Switching Jump-Diffusion (RSJD) model that incorporates stochastic volatility.

3.1 Regime-Switching Process

The global market state is modeled as a discrete-time Markov chain $M_t \in \{0, 1\}$, representing "Normal" and "Crisis" regimes, respectively. The transition between these states is defined by a transition probability matrix P:

$$\begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} \tag{1}$$

where p_{01} represents the probability of entering a crisis regime and p_{10} the probability of recovery. In the `fsynth` implementation, parameters such as drift, mean-reversion speed, and jump intensity are functions of M_t.

3.2 Stochastic Price Dynamics

The logarithmic return of the market index S_t is governed by the following stochastic differential equation (SDE):

$$\frac{dS_t}{S_t} = (\mu_{M_t} - \lambda_{M_t} \bar{J})dt + \sqrt{v_t} dW_t^S + J_t dN_t \tag{2}$$

where μ_{M_t} is the regime-dependent drift, v_t is instantaneous variance, and W_t^S is the standard Wiener process. The term $J_t dN_t$ represents the Jump-Diffusion component, where N_t is a Poisson process with regime-dependent intensity λ_{M_t}, and J_t is the jump magnitude sampled from a log-normal distribution: $(1 + J) \sim \mathcal{LN}(\mu_j, \sigma_j^2)$.

3.3 Volatility Evolution

To account for volatility clustering and the leverage effect, the variance v_t follows a Cox-Ingersoll-Ross (CIR)[4] process as defined in the Heston model:

$$dv_t = \kappa_{M_t}(\theta_{M_t} - v_t)dt + \xi_{M_t}\sqrt{v_t}dW_t^v \tag{3}$$

where:

- κ_{M_t} is the speed of mean reversion.
- θ_{M_t} is the long-term variance level.
- ξ_{M_t} is the "volatility of volatility."
- dW_t^v is a Wiener process correlated with dW_t^S such that $d\langle W^S, W^v \rangle_t = \rho_{M_t} dt$

In our implementation, ρ_{M_t} is set to a more negative value during the "Crisis" regime ($\rho < \rho_0$) to replicate the intensified leverage effect observed during market crashes.

3.4 Multi-Stock Factor Structure

Individual asset prices $S_{i,t}$ are derived using a multi-factor approach linked to the market return $r_{m,t}$[14][6]:

$$r_{i,t} = \alpha_i + \beta_i r_{m,t} + \gamma_i \epsilon_{sector,t} + \sigma_i \epsilon_{idio,t} \tag{4}$$

where ϵ_{sector} and ϵ_{idio} are sector specific and idiosyncratic shocks, respectively. This allows for the generation of a correlated universe of stocks that maintain structural consistency with the broader market.

3.5 Structural Interpretability vs Neural Synthesis

A newer trend in financial data generation is the use of Generative Adversarial Networks (GANs)[18][7]. While statistically accurate in the replication of moments in a historical distribution, GANs suffer from a lack of interpretability and being a "black box." A GAN might be able to generate a realistic flash crash, but it would be impossible to explain the causality behind it, nor could it guarantee that the generated price path adheres to fundamental no-arbitrage constraints.

By contrast, fsynth operates as a fully interpretable and explicable alternative. Every trend represented in a path generated by fsynth can be traced back to a specific, parameterizable, cause. If the synthetic market crashes, the practitioner can explicitly attribute it to a regime shift ($M_t \to 1$) or a discontinuous jump event ($dN_t > 0$). This interpretability allows for counterfactual reasoning—asking ”What if?” questions (e.g., ”What if the crisis regime lasted twice as long?”), an ability unattainable through analysis of GAN-generated paths, which are only ever able to interpolate within the manifold of their own training data.

4 Fundamental Synthesis and Corporate Genes

A core limitation of most synthetic financial data generators is the absence of fundamental accounting data consistent with price action[13]. To address this, fsynth introduces ”Corporate Genes:” a set of inherent parameters that govern how a synthetic stock reacts to macroeconomic shifts.

4.1 Genetic Parameterization

Each synthetic asset i is initialized with a vector of corporate genes $\mathbf{G}_i$:

$$\mathbf{G}_i = \{g_{growth}, m_{stab}, l_{tol}\} \tag{5}$$

where g_{growth} represents growth potential, m_{stab} represents margin stability, and l_{tol} represents leverage tolerance. These genes are sampled from distributions specific to the asset's sector. For instance, assets in the technology sector generally exhibit high potential for growth and variable margins, while utility assets exhibit lower growth potential but higher stability of margins.

4.2 Regime-Aware Fundamental Generation

Fundamental reports are generated on a quarterly basis (Q). The reported metrics, such as Revenue ($\mathcal{R}$), EBITDA ($\mathcal{E}$), and Earnings Per Share (EPS), are derived by putting the corporate genes through the integrated market regime M_Q:

$$\mathcal{R}_{i,Q} = \mathcal{R}_{i,Q-1} \times (1 + g_{growth,i} + \Delta M_Q + \epsilon_{idio}) \tag{6}$$

where ΔM_Q is a macro growth modifier determined by the mean regime state during the quarter.

4.3 Reflexivity and Debt Mechanics

To replicate ”death spiral” dynamics observed during credit crunches[11], fsynth implements a reflexive debt-service model. The interest expense $\mathcal{I}$ is modeled as a function of both the entity's leverage and the regime severity:

$$\mathcal{I}_{i,Q} = \text{Debt}_{i,Q} \times (r_f + \phi \cdot \bar{M}_Q) \tag{7}$$

where r_f is the base risk-free rate and ϕ is a sensitivity coefficient. In high volatility regimes ($M = 1$), the cost of debt increases, directly impacting Net Income and EPS. This linkage ensures that fundamental analysis models trained on `fsynth` learn the nonlinear relationship between macro-volatility and corporate solvency.

5 Empirical Validation

A primary metric for financial realism is excess kurtosis, which measures the frequency of extreme events. We calibrated the `fsynth` engine to the S&P 500 index over a five year period to determine if the engine can replicate the "Crisis" dynamics observed in empirical data-specifically the transition from low variance to high intensity, fat-tailed return events.

5.1 Calibration and Parameterization

Unlike traditional GBM, which is calibrated using only the first two moments (mean and variance), `fsynth` utilizes a high volatility of volatility (ξ) and a high jump intensity (λ_j) to replicate market turbulence. The calibration logic used in our validation script is summarized as follows:

- **Normal Regime (μ_0, θ_0):** A lower long-term variance level ($\theta_0 = 0.012$) and slower mean-reversion ($\kappa_0 = 0.5$) were utilized to simulate the "quiet" periods of market growth, resulting in a sharp, tall peak at the center of the return distribution.
- **Tail Generation (ξ_0, λ_j):** An increased volatility-of-volatility coefficient ($\xi_0 = 0.8$) and a Poisson jump intensity of $\lambda_j = 0.15$ were implemented to generate the "fat tails" that traditional models fail to capture.
- **Jump Dynamics (μ_j, σ_j):** Jumps were modeled with a mean magnitude of -3% to replicate the asymmetric downside risk observed in equity markets.

5.2 Leptokurtosis and Fat-Tail Analysis

The primary success metric for the validation was the Kurtosis score, which measures the extremity of outliers. Financial machine learning models' primary bottleneck is the lack of anticipation for the frequency of these outliers.

As shown in Table 1, while GARCH(1,1), the industry standard for volatility modeling, produces an excess kurtosis of 0.33, `fsynth` generates a score of 5.79. This captures most of the leptokurtic behavior found in the real market, which had a score of 8.05. This nearness is crucial in training robust agents and models capable of surviving market "flashes" and liquidity crises.

Data Source	**Excess Kurtosis**
S&P 500 (Target)	8.05
Gaussian Distribution	0
GARCH(1,1) Model	0.33
fsynth	**5.79**

Table 1: Validation statistics comparing the calibrated `fsynth` engine against historical S&P 500 returns. `fsynth` achieves a significantly higher fidelity to real market turbulence than the industry standard GARCH(1,1) and Gaussian models.

5.3 Log-Scale Distribution Analysis

To further investigate the "Fat Tails," we performed a log-scale histogram analysis of the returns. As shown in the log-density plot, a normal distribution appears as an inverted parabola. Real market returns, however, exhibit "straight" or flared edges, indicating a much higher probability of extreme movements.

Our validation confirms that the `fsynth` engine replicates this flaring. By coupling the Heston stochastic volatility with discrete Merton jumps, the engine generates returns that reside multiple standard deviations away from the mean—events that a standard model would categorize as "statistically impossible" but that occur frequently in actual trading environments.

5.4 Parameter Sensitivity and Robustness

To ensure that the leptokurtic behavior observed in fsynth is a structural property rather than a calibration artifact, we performed a parameter sensitivity analysis. We isolated two key stochastic drivers: the Poisson jump intensity (λ_j) and the Heston volatility-of-volatility (ξ).

As illustrated in Figure 2, we observe a clear positive relationship between the stochastic parameters and the excess kurtosis. The Right Plot shows that increasing the volatility-of-volatility (ξ) from 0.1 to 1.5 smoothly amplifies the kurtosis from near-zero to ¿6.0, extending the severity of volatility clusters.

The Left Plot highlights the discrete nature of the jump-diffusion process. As the intensity (λ_j) increases, the kurtosis rises in a step-like fashion, reflecting the discrete addition of "shock" events to the time series. This confirms that fsynth offers direct parametric controllability, allowing practitioners to target specific risk profiles by analytically adjusting the frequency of black swan events.

6 Case Study: Solvency Crisis

To demonstrate the structural utility of fsynth beyond simple price replication, we simulated a "Solvency Crisis" scenario. In this experiment, we forced the global market regime M_t into the "Crisis" state ($M_t = 1$) for a duration of four

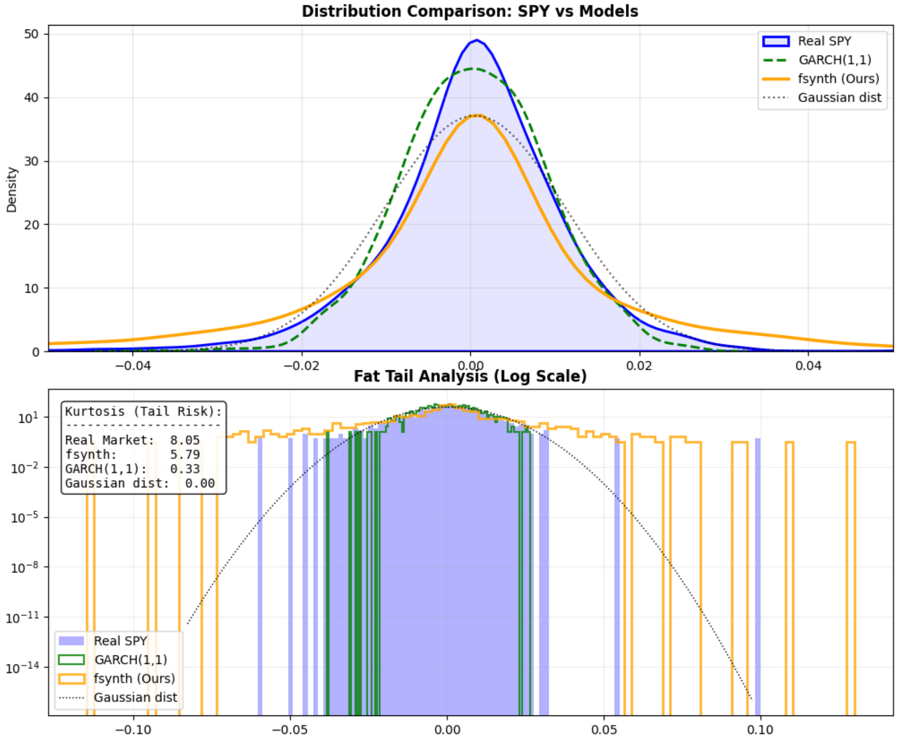

Figure 1: As shown by the second graph, the yellow bars (`fsynth`) track the blue bars (S&P 500) into the tails, unlike the standard GARCH(1,1) and Gaussian distributions.

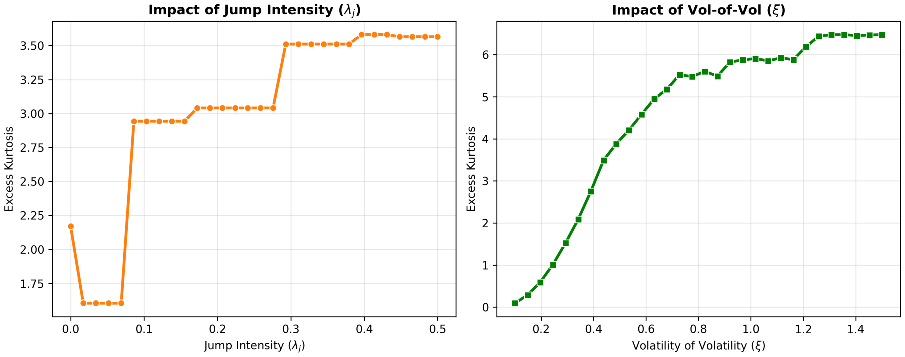

Figure 2: As shown on the left, the "step-like" shape of the graph reflects the discrete nature of the Poisson process, proving that `fsynth` is accurately simulating discrete "Black Swan" events. On the right, the smooth, almost monotonic increase proves that increasing volatility of volatility (ξ) reliably makes the tails heavier.

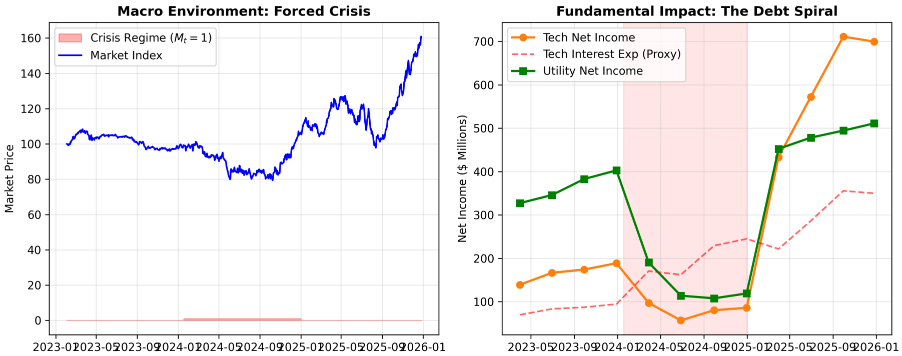

Figure 3: Simulation of a "Solvency Crisis" scenario. On the left, the simulation is forced into a Crisis Regime ($M_t = 1$, red shaded area), causing a drawdown in the market index. Shown on the right is the fundamental impact on two distinct assets. The high-leverage "Tech" asset (Orange) suffers a "death spiral" where the reflexive interest expense (Red Dashed) spikes due to the regime shift, driving Net Income down by $> 60\%$. In contrast, the conservative "Utility" asset (Green) remains solvent with stable earnings, demonstrating the model's ability to differentiate credit risk based on corporate genes.

quarters, representing a prolonged credit crunch similar to the 2008 financial crisis.

6.1 Experimental Setup

We initialized two distinct synthetic equities:

- Asset A (Tech-Growth): Characterized by high growth potential ($g_{growth} = 0.20$) but high leverage tolerance ($l_{tol} = 0.8$).
- Asset B (Utility-Stable): Characterized by low growth ($g_{growth} = 0.05$) and conservative leverage ($l_{tol} = 0.3$).

6.2 The Death Spiral Mechanics

As illustrated in Figure 3, the onset of the crisis regime triggers the reflexive debt mechanism defined in Equation 7. The cost of debt rises sharply as the risk-free rate sensitivity ϕ interacts with the crisis state $M_t = 1$.

For Asset A (Orange Line), this creates a "death spiral." The increased interest expense (Red Dashed Line) outpaces its EBITDA, driving Net Income down by over 60% during the crisis window. This effectively simulates a fundamental default risk triggered by macro conditions.

In contrast, Asset B (Green Line), with its lower debt load and stable margins, experiences a minor dip in earnings but remains solvent.This experiment confirms that `fsynth` successfully encodes the non-linear relationship between macro-volatility and corporate fundamentals, enabling the training of risk-management algorithms that can detect solvency traps before they appear in price data.

7 Discussion and Practical Utility

The primary objective of `fsynth` is to bridge the "Sim-to-Real" gap in financial machine learning. While the engine is shown to demonstrate high fidelity to historical benchmarks, its true value lies in its ability to generate unseen, parallel histories for robust model training and stress testing.

7.1 Domain Randomization and Generalization

Traditional models trained on historical data often overfit to the specific volatility regimes of the training period. Through `fsynth`, practitioners can employ domain randomization-varying the parameters of the stochastic engine such as jump intensity λ_j or the regime transition probabilities p_{01} to expose machine learning models to a multitude of different historically unrealized, yet mathematically plausible, market conditions[17]. This ensures that the agents are not merely learning to perform well in the past, but are learning how to profit in a variety of environments[15].

7.2 Stress Testing and Black Swan Simulation

The parameterizable nature of `fsynth` allows for the targeted simulation of "Black Swan" events. By artificially inflating parameters related to the "Crisis" regime, `fsynth` can be used as a rigorous stress-testing framework to evaluate the solvency of a portfolio during extreme liquidity crunches. The linkage between price action and the reflexive debt-service model further allows for the testing of strategies based on fundamentals against systemic credit risk.

8 Conclusion

This paper introduced `fsynth`, a high fidelity synthetic universe generator designed to combat data scarcity and risks of overfitting in financial machine learning. By integrating regime-switching stochastic volatility with Merton jump diffusion and a novel "Corporate Gene" fundamental layer, we have created a framework capable of generating self-consistent financial environments.

Our empirical validation confirms that `fsynth` effectively replicates the leptokurtosis and fat-tailed distributions characteristic of real-world equity markets, far exceeding the capabilities of traditional Gaussian-based models. As financial AI continues to evolve, the necessity for high-fidelity synthetic environments like `fsynth` will become essential in ensuring the stability and resilience of automated trading systems in a non-linear global market.

References

[1] Andrew Ang and Geert Bekaert. International asset allocation with regime shifts. *Review of Financial Studies*, 15:1137–1187, 07 2002.

[2] David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, 61:458, 05 2014.

[3] Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1:223–236, 02 2001.

[4] John C. Cox, Jonathan E. Ingersoll, and Stephen A. Ross. A theory of the term structure of interest rates. *Econometrica*, 53:385, 03 1985.

[5] Eugene Fama. Efficient capital markets: a review of theory and empirical work. *The Journal of Finance*, 25:383–417, 05 1970.

[6] Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33:3–56, 02 1993.

[7] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 06 2014.

[8] James D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–384, 03 1989.

[9] Steven L. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6:327–343, 04 1993.

[10] Benoit Mandelbrot. The variation of certain speculative prices. *The Journal of Business*, 36:394, 01 1963.

[11] Robert C. Merton. On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29:449–470, 05 1974.

[12] Robert C. Merton. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3:125–144, 01 1976.

[13] James A Ohlson. Earnings, book values, and dividends in equity valuation. *Contemporary Accounting Research*, 11:661–687, 03 1995.

[14] Stephen A Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13:341–360, 12 1976.

[15] Richard S Sutton and Andrew Barto. *Reinforcement learning: An introduction.* The Mit Press, 2nd edition, 2018.

[16] Nassim Nicholas Taleb. *The Black Swan: The Impact of the Highly Improbable.* Taylor And Francis, 2007.

[17] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. *arXiv:1703.06907 [cs]*, 03 2017.

[18] Magnus Wiese, Robert Knobloch, Ralf Korn, and Peter Kretschmer. Quant gans: deep generation of financial time series. *Quantitative Finance*, 20:1419–1440, 04 2020.