Kalman Filter Momentum Strategy for SPY: Signal Design, Out-of-Sample Testing, and Statistical Validation

David Kumar
The Ohio State University, Fisher College of Business
B.S. Finance, Business Analytics Minor
kumar.1189@osu.edu

February 2026

Abstract

This paper presents a long-only momentum trading strategy for SPY using 15-minute intraday data. The strategy applies a Kalman filter to raw price data to estimate velocity and acceleration states, entering positions when both are positive and sizing based on relative momentum strength. The model is trained on data from 2014 to 2017 and tested across two separate out-of-sample periods: 2018 to 2022 and August 2025 to February 2026, with a three-year gap between them due to data availability constraints. Across 233 out-of-sample trades, the strategy produced a combined Sharpe ratio of 1.13, a total return of 199%, and a per-trade expected value of +0.518%. Monte Carlo simulation against 10,000 random entry benchmarks yielded a p-value below 0.001. Parameter sensitivity analysis is presented as an additional robustness check.

1. Introduction

Momentum strategies have a long history in quantitative finance. The basic observation is that assets which have historically tended to go up keep going up, at least for a while. Most implementations rely on lookback returns over some fixed window, typically 1 to 12 months, and use that to rank assets or generate entry signals. The challenge with these approaches on shorter timeframes, particularly intraday data, is noise. Raw price movements at the 15-minute level contain a lot of short-term fluctuation that can trigger false signals and lead to overtrading.

This paper explores a different approach: using a Kalman filter, originally developed for aerospace navigation, to extract momentum information from intraday price data. Rather than computing returns over a fixed window, the Kalman filter continuously estimates the underlying trend by modeling price as a system with position, velocity, and acceleration states. The velocity state provides a smoothed measure of trend direction, while the acceleration state captures whether the trend is strengthening or weakening. By conditioning entries on both states being positive, the strategy attempts to enter only when price is moving upward and that movement is gaining strength.

The goal of this project was to build a trading signal from scratch, test it properly on data the model never saw during development, and apply statistical methods to evaluate whether the observed results could be attributed to the signal design or to chance. Backtest overfitting is a well-documented problem in strategy development, and this paper addresses it by using strict temporal separation between training and testing periods and by benchmarking against random entry simulations. The strategy is intentionally kept simple (long-only, single asset, no leverage) to isolate the contribution of the Kalman-based signal itself.

2. Data

The primary dataset consists of 15-minute OHLC bars for SPY from January 2014 through December 2022, totaling 56,274 bars. The data was split into three non-overlapping periods:

Period	Role	Date Range	Bars
Training	Parameter estimation	Feb 2014 - Dec 2017	25,206
Validation	Out-of-sample test	Jan 2018 - Dec 2020	18,979
Test	Out-of-sample test	Jan 2021 - Jul 2022	9,682

The validation and test sets were never used during development. Although the original intent was to use validation for hyperparameter tuning, no parameters were adjusted based on validation results, so both periods function as pure out-of-sample tests. They are combined when reporting aggregate OOS statistics.

A second dataset was obtained from a Bloomberg Terminal, covering August 2025 through February 2026 (approximately 2,800 bars). Bloomberg's intraday history for SPY extends roughly six months back, which created an unavoidable three-year gap between the two test periods (mid-2022 to mid-2025). While not ideal, this gap provides an unintentional robustness test: the strategy had to perform in a market environment it had never been exposed to, after a significant passage of time. The parameters and logic were frozen before touching any Bloomberg data.

3. Methodology

3.1 Kalman Filter Signal Extraction

The strategy uses a Kalman filter to model SPY's 15-minute closing prices as a dynamic system with three states: filtered price (position), velocity (first derivative), and acceleration (second derivative). At each bar, the filter updates its estimates based on the new price observation, balancing between trusting the prediction from the previous state and trusting the new data point.

The key tuning parameter is process noise, which controls how reactive the filter is. A low process noise value produces a smooth, slow-moving filter that resists reacting to short-term fluctuations. A high value makes the filter more responsive, tracking price changes more closely but also picking up more noise. The base configuration uses a process noise of 0.01, which was set during the training period. Section 5.2 tests the strategy's sensitivity to this parameter.

The velocity state provides a smoothed estimate of how fast price is moving up or down. The acceleration state captures whether that movement is speeding up or slowing down. Together, they describe not just the direction of the trend but its momentum, specifically whether the trend has conviction behind it.

3.2 Entry and Exit Rules

A long position is initiated when the Kalman velocity estimate is positive and the acceleration estimate is also positive. In plain terms: price is going up, and it's speeding up. Both conditions must be met simultaneously.

Once a position is entered, a minimum holding period of 130 bars (approximately 5 trading days) is enforced. This prevents the strategy from reacting to short-term noise that might briefly flip acceleration negative before the trend resumes. After the minimum hold, the position is closed when the acceleration state turns negative, indicating that the upward trend is losing momentum. A maximum hold of 520 bars is imposed as a safety cap.

3.3 Conviction-Based Position Sizing

Not all signals are equal. A velocity reading barely above zero suggests a weak trend, while a velocity at the 95th percentile of recent history suggests strong momentum. The strategy scales position size based on where the current velocity falls relative to the last 500 bars (approximately 19 trading days).

The position weight is calculated as: weight = 0.2 + 0.8 x percentile, where percentile is the rank of the current velocity over the trailing 500-bar window. This produces weights ranging from 0.2 (for a weak signal at the 0th percentile) to 1.0 (for a signal at the 100th percentile). The floor of 0.2 is a design choice, ensuring a minimum 20% allocation to any valid signal. The coefficient of 0.8 is the complement (1.0 minus 0.2), producing a linear scale from the floor to full allocation. This floor was not optimized; it reflects a preference for always participating in valid signals rather than sitting out entirely when conviction is low.

Only velocity drives sizing. Acceleration acts as a binary gate: it must be positive to enter, and when it flips negative the position is closed. It does not affect the size of the position.

Figure 1 illustrates the strategy in action during the Bloomberg period. The top panel shows SPY price with entry and exit markers. The middle and bottom panels display the velocity and acceleration states that drive trade decisions. Green shaded regions indicate periods where the strategy held a position.

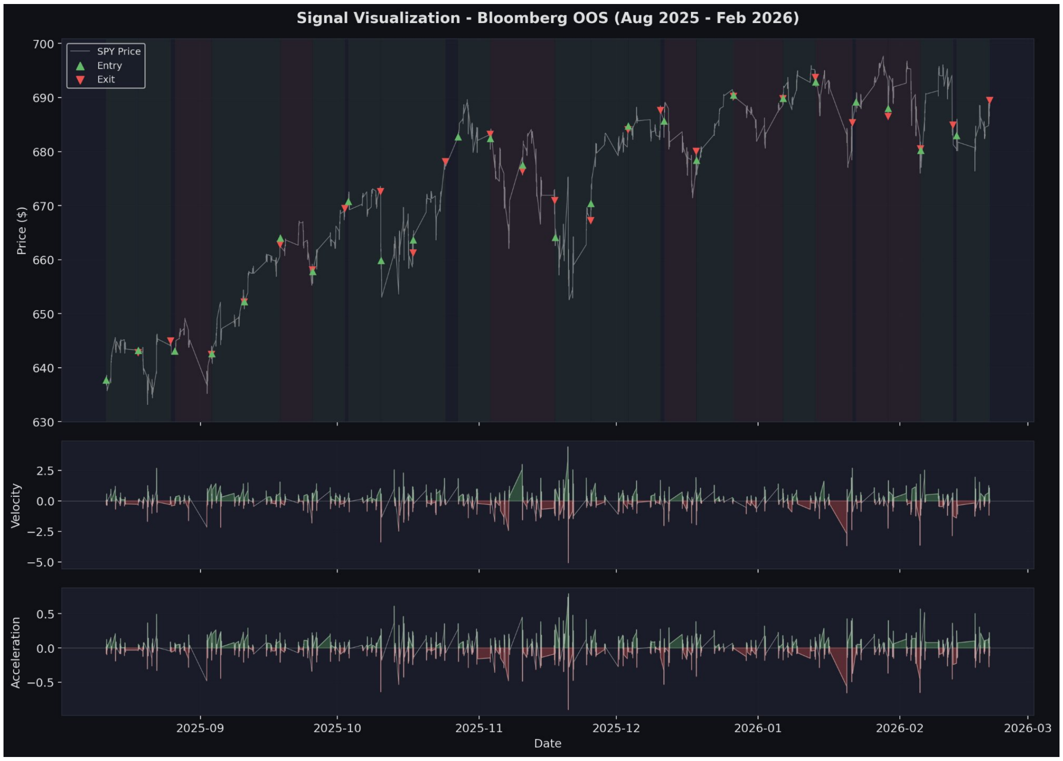

Figure 1. *Strategy behavior during the Bloomberg out-of-sample period. The same signal logic and parameters from the 2014 to 2017 training period were applied without modification to data collected three years after the original test set ended.*

4. Results

The strategy was run on all out-of-sample data with frozen parameters. No adjustments were made between periods. The following table summarizes performance across each test window and in aggregate.

Period	Trades	Sharpe	Return	Win Rate	Max DD

Validation (2018-2020)	119	0.97	+72.0%	52.1%	-23.4%
Test (2021-2022)	89	1.35	+63.6%	51.7%	-19.1%
Combined OOS (2018-2022)	208	1.10	+181.6%	51.9%	-23.4%
Bloomberg (Aug 2025-Feb 2026)	25	1.84	+6.3%	52.0%	-10.2%
All OOS Combined	**233**	**1.13**	**+199.0%**	**51.9%**	**-23.4%**

All returns are compounded, calculated bar-by-bar using the position weight from the prior bar. Sharpe ratios are annualized from 15-minute bar returns. Win rate is calculated on unweighted trade returns.

The Sharpe ratio is consistent across periods, ranging from 0.97 to 1.84. The win rate remains near 52% in every period. The strategy's positive expected value derives not from a high win rate but from an asymmetry in payoffs: average wins exceed average losses by approximately 10%, and conviction sizing allocates more capital to higher-percentile entries. Section 4.1 quantifies this expected value in detail.

The Bloomberg period produced the highest Sharpe ratio of any individual window (1.84) despite covering only six months and 25 trades. This period began three years after the original dataset ended, with no parameter adjustments or re-optimization applied to the Bloomberg data.

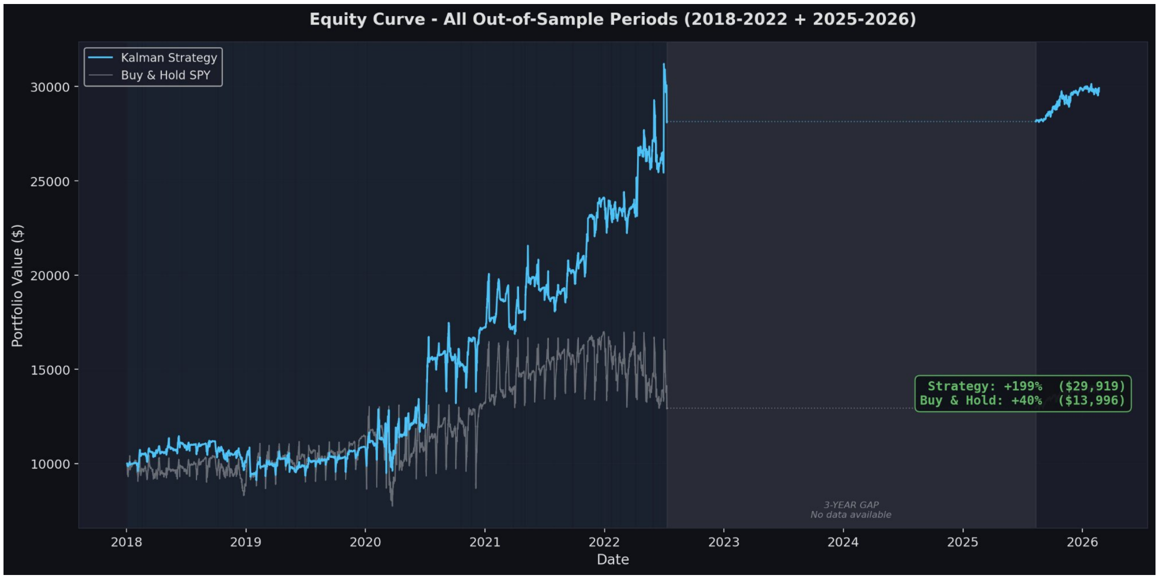

Figure 2. *Equity curve comparing the Kalman strategy to SPY buy-and-hold across all out-of-sample periods. Starting capital of $10,000. The shaded region indicates the three-year gap where no intraday data was available. Buy-and-hold comparison covers the 2018 to 2022 period only; the Bloomberg segment (2025 to 2026) shows the strategy's standalone continuation.*

4.1 Expected Value

Expected value (EV) per trade measures the average return a strategy generates each time it enters a position. It is calculated as: EV = (win rate x average win) + (loss rate x average loss). A positive EV indicates that the strategy produces a net gain per trade on average. The profit factor is the ratio of gross profits to gross losses: profit factor = (win rate x average win) / (loss rate x |average loss|). A profit factor above 1.0 means the strategy's winners outweigh its losers in aggregate.

Across all 233 out-of-sample trades, the strategy produced the following breakdown:

Metric	Value
Win Rate	51.9%
Average Win	+6.20%
Average Loss	-5.62%
EV per Trade	+0.518%
Profit Factor	1.19

Applying the formula: (0.519 x 6.20%) + (0.481 x -5.62%) = +0.518% per trade. The profit factor of 1.19 confirms that gross gains exceed gross losses by 19%. The win rate of 51.9% is only marginally above 50%; the positive EV is driven by the asymmetry in payoff size rather than prediction accuracy.

Figure 3 plots the cumulative sum of trade returns sequentially across all 233 out-of-sample trades. The dashed line represents the theoretical trajectory if every trade returned exactly +0.518% (the observed EV). If the actual cumulative path tracks this line consistently, it indicates the edge is distributed across the full sample. If the gains were concentrated in a small cluster of trades, the actual path would diverge significantly from the theoretical line for extended stretches.

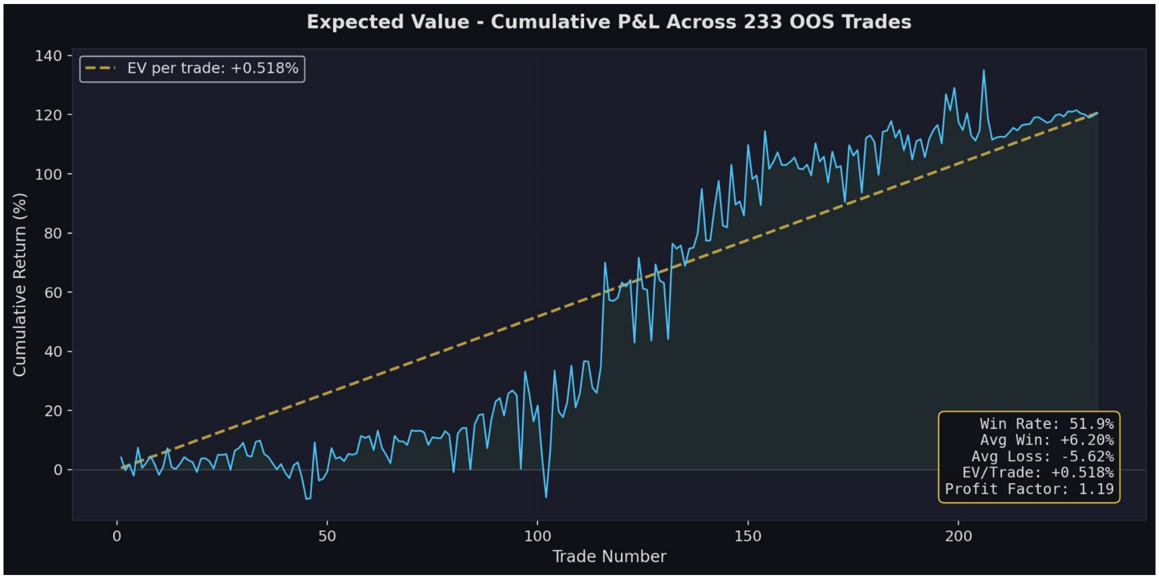

Figure 3. *Cumulative return plotted trade-by-trade for all 233 out-of-sample trades. The dashed line represents the theoretical EV of +0.518% per trade. The actual path oscillates around but trends with the EV line, indicating a distributed rather than concentrated edge.*

5. Statistical Validation

A positive backtest result alone does not establish that a strategy has a genuine edge. It could be the product of overfitting, luck, or data mining. This section describes two tests applied to evaluate the robustness of the observed results.

5.1 Monte Carlo Simulation

The primary statistical test asks: if we entered trades at random times instead of using the Kalman signal, how often would we observe a Sharpe ratio as high as the strategy's? To answer this, a Monte Carlo simulation generated 10,000 sets of random entries with the same constraints as the actual strategy (same number of trades, same minimum hold period, non-overlapping positions). A Sharpe ratio was computed for each random set.

On the combined 2018 to 2022 out-of-sample data, the strategy's Sharpe of 1.10 exceeded all 10,000 random benchmarks. When combined with the Bloomberg period (233 total trades), the p-value was below 0.001, meaning fewer than 10 out of 10,000 random simulations matched or exceeded the observed Sharpe of 1.13. This suggests the results are unlikely to be explained by chance entry timing alone.

It is worth noting what this test does and does not show. It shows that the specific timing of entries matters: entering when velocity and acceleration are both positive produces better results than entering at random. It does not prove the strategy will continue to work, nor does it rule out all forms of overfitting. It is one piece of evidence, not a proof.

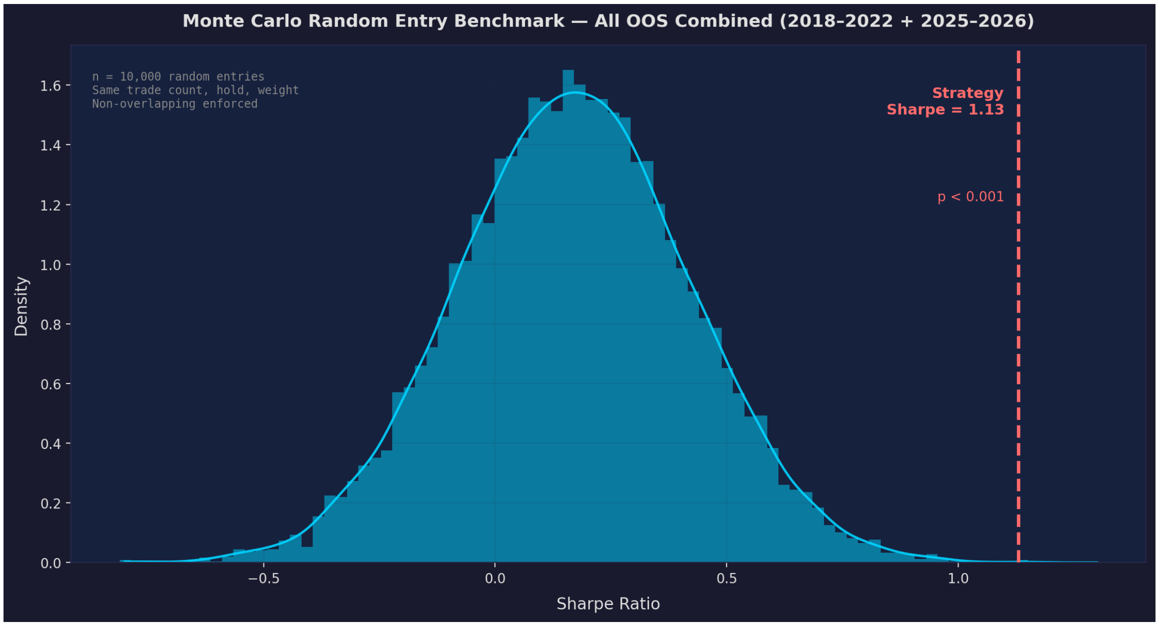

Figure 4. *Distribution of Sharpe ratios from 10,000 Monte Carlo random entry simulations on the combined out-of-sample data. The strategy's observed Sharpe of 1.13 (dashed line) falls in the extreme right tail, with p < 0.001.*

5.2 Parameter Sensitivity

A strategy that only works at one exact setting is likely overfit to the training data. To test whether the Kalman signal is genuinely capturing momentum or whether the results depend on a specific calibration, the two primary tunable parameters were varied across a grid of values and the strategy was re-run on out-of-sample data for each combination.

The first parameter is the minimum hold period, which determines how long the strategy must hold a position before it is allowed to exit. This was tested from 70 bars (approximately 2.7 trading days) to 200 bars (approximately 7.7 trading days). The second parameter is the Kalman filter's process noise, as described in Section 3.1. A low value (0.003) produces a very smooth filter that reacts slowly to price changes; a high value (0.1) produces a responsive filter that tracks price closely. The base configuration uses a process noise of 0.01, roughly in the middle of this range on a logarithmic scale.

Across all 42 parameter combinations, every configuration produced a positive out-of-sample Sharpe ratio, ranging from 0.22 to 1.11. The base configuration (min_hold = 130, process_noise = 0.01) produced a Sharpe of 0.77, which falls near the middle of the distribution rather than at the peak. This indicates two things: the chosen parameters were not cherry-picked from an optimal region, and the underlying momentum signal the filter captures is present regardless of how aggressively or conservatively the filter is tuned.

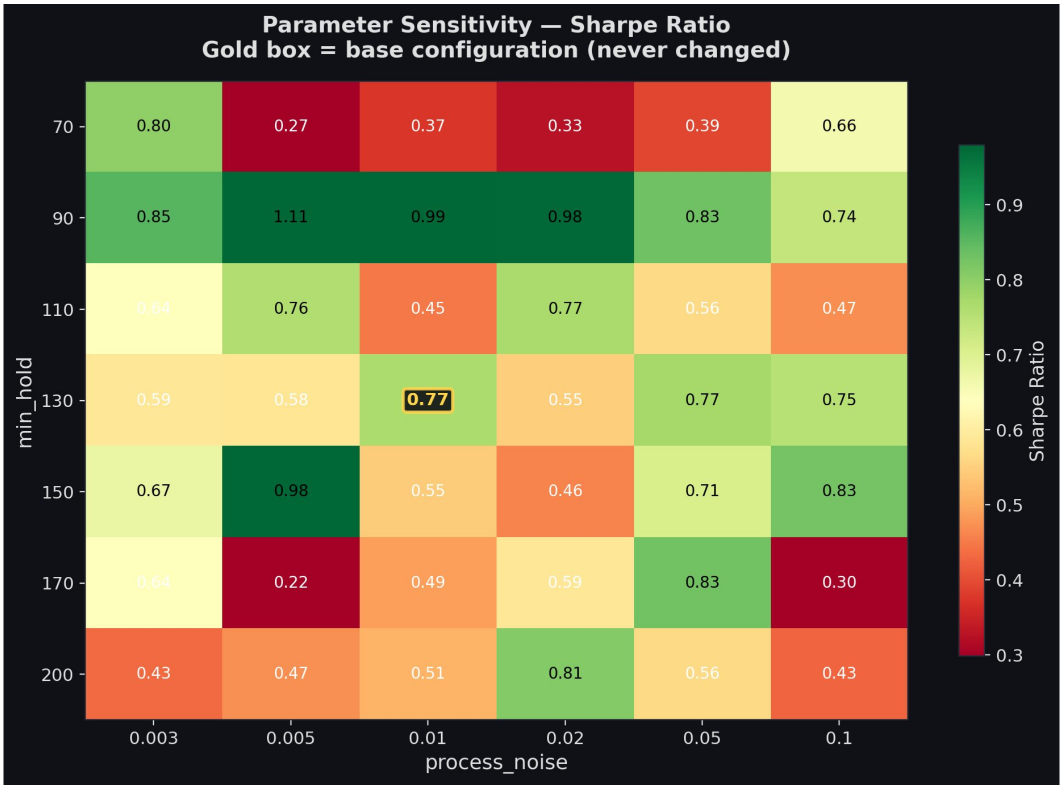

Figure 5. *Heatmap of out-of-sample Sharpe ratios across different combinations of minimum hold period and process noise. The gold box marks the base configuration. All 42 parameter combinations produced positive Sharpe ratios, indicating the strategy is not dependent on precise calibration.*

6. Limitations and Future Work

Minimum hold period and tail risk. The 130-bar minimum hold means the strategy cannot exit during sudden market drops within the first five trading days of a position. If a flash crash or sharp reversal occurs during this window, the strategy is locked in. The conviction sizing provides some natural protection, since the strategy is unlikely to be at full size during sudden reversals (those typically don't occur when velocity is at a high percentile), but it does not eliminate the risk. Future iterations could explore stop-loss mechanisms or volatility-adjusted hold periods.

Long-only constraint. The strategy only takes long positions. It has no way to profit from downtrends. A natural extension would be to mirror the logic for short entries (negative velocity, negative acceleration), though short-side momentum has historically been less reliable and more expensive to implement.

Single asset. All testing was performed on SPY. Whether the signal generalizes to other ETFs, individual stocks, or other asset classes is an open question.

Transaction costs. The backtest does not model transaction costs, slippage, or market impact. SPY is highly liquid with tight spreads, so these costs would be small on a per-trade basis, but they are not zero. With an average of roughly 29 trades per year and an EV of 0.518% per trade, the edge is large enough to absorb reasonable transaction costs, but this has not been formally quantified.

Data gap. The three-year gap between the two test periods (mid-2022 to mid-2025) means the strategy has not been tested on that specific window. While the gap serves as an indirect robustness test, a continuous backtest across the full period would be more rigorous.

Sample size. 233 trades is a reasonable sample for statistical testing, but not a large one. The Bloomberg period in particular contains only 25 trades. Longer out-of-sample histories would strengthen the evidence.

Forward testing. Paper trading the strategy in real time would remove any remaining possibility of lookahead bias or data snooping. This is the intended next step for this work.

7. Conclusion

This paper documented the process of building and testing a Kalman filter momentum strategy for SPY using 15-minute intraday data. The strategy uses the filter's velocity and acceleration states to time entries, conviction-based sizing to weight positions, and an acceleration reversal to time exits.

Across 233 out-of-sample trades spanning 2018 to 2022 and 2025 to 2026, the strategy produced a Sharpe ratio of 1.13, a cumulative return of 199%, and an expected value of +0.518% per trade. Monte Carlo testing against random entries yielded a p-value below 0.001. Parameter sensitivity analysis showed positive Sharpe ratios across all 42 tested configurations, indicating the results are not dependent on a narrow set of parameter choices.

What this paper demonstrates is a process: formulating a hypothesis about what drives short-term momentum, translating it into a concrete signal, testing it on unseen data, and evaluating the statistical significance of the outcome. The natural next step is forward testing in real time, which would provide additional validation that a backtest alone cannot.