Regime Based Multi-Asset Long Short Strategy

Applied Quantitative Macro Strategies

Mai Luo

Imperial Business School

June 2025

Table of Contents

Table of Figures

Tables

1. Introduction

Financial markets cycle through different "regimes"; For example, periods of robust growth, recessions, or sudden spikes in volatility. Recognizing which regime is unfolding can guide portfolio adjustments: during downturns, one might favour bonds and reduce equities, while in expansions, the opposite makes sense. Traditional signals such as an inverted yield curve help identify looming recessions but can be too coarse or lag actual stress. On the other hand, fitting complex clustering or hidden-state models to dozens of indicators risks overfitting and can become opaque when the economy shifts unexpectedly. Our approach is more transparent. Each Friday, we collect five macro-financial measures that together capture growth expectations, credit conditions, cross-asset sentiment, and volatility outlook:

- **Yield-Curve Slope (10-Year minus 2-Year Treasury).**

  The difference between long-term and short-term Treasury yields reflects market expectations of future growth and monetary policy. When the 2-year yield exceeds the 10-year (an inversion), recessions frequently follow within 6–18 months. We take the weekly spread and compute its z-score over the prior 52 weeks, so that large deviations (steep inversions or unusually steep curves) stand out relative to recent history.

- **Corporate Credit-Spread Proxy (CDX Index).**

  Direct historical corporate bond spreads are hard to find before 2000, so we use the Markit CDX North America Investment Grade index, which measures the cost of insuring a basket of investment-grade bonds. Each Friday, we subtract the comparable Treasury yield to approximate a credit spread and then z-score that level over the past 52 weeks. Widening credit spreads signal rising default risk and broader financial stress.

- **Equity-Bond Correlation.**

  In normal times, equities and bonds often move inversely; Stocks rise while bonds offer safety. During systemic stress, however, they can fall together. To capture this, we calculate weekly returns on an equity futures contract (e.g., S&P 500 E-mini) and a 10-year Treasury futures contract from Friday to Friday. We compute their 52-week rolling Pearson correlation and then z-score that correlation over the preceding ten years. A high positive z-score, meaning stocks and bonds are both underperforming, signifies "all-sell" market stress.

- **Implied Volatility (VIX).**

  The VIX index represents the market's expectation of 30-day S&P 500 volatility. Spikes in VIX typically coincide with or precede equity drawdowns. We take Friday's closing VIX level and compute a z-score over the past year, so that we flag unusually high (or low) volatility expectations relative to the recent past.

- **VIX Futures-Curve Slope.**

  Beyond just the front-month VIX, the slope of the VIX futures curve (difference between the second-month and first-month VIX futures prices) hints at whether markets expect volatility to rise. A steep upward slope often builds ahead of crises as participants price in heightened future uncertainty. Each Friday, we record front-month and second-month VIX futures prices, take their difference, and z-score it over a ten-year window.

After computing these five rolling z-scores, we form a vector representing that week's "macro-financial snapshot." To find today's closest historical peers, we compute the Euclidean distance between the current vector and every prior week's vector (excluding today). The handful of weeks with the smallest distances become our "historical analogues." We then gather each analogue's one-week forward returns for a modest asset universe including equity futures, Treasury bond futures, and commodity futures, and form a weighted average of those returns, where closer analogues carry higher weight. That weighted average becomes our forecast for the coming week. Finally, we convert forecasts into actual positions by volatility-scaling via an exponentially weighted covariance matrix, targeting a stable annualized risk.

Because we only ever use data available up to the decision date, there is no look-ahead bias. And by relying on only five well-chosen indicators (each expressed as a z-score), we keep the method simple yet powerful: if today’s environment most closely resembles late 2008, we know exactly which episode drives our signal.

2. Literature Review and Theoretical Background

2.1. Regime Signals in Macro-Finance

Economic regimes such as periods of expansion, recession, or heightened stress, have long been studied. The yield-curve slope (10-year minus 2-year Treasury yields) is a classic recession indicator: when short-term rates exceed long-term rates, recessions often follow within the next year or two (Estrella & Hardouvelis 1991; Wright 2006). Similarly, widening credit spreads (extra yield corporate borrowers pay over Treasuries) signal deteriorating credit conditions and often precede broader market downturns (Gilchrist & Zakrajšek 2012). Cross-asset behavior also matters: equity-bond correlations, normally slightly negative, can flip positive during systemic sell-offs, indicating “sell everything” panics (Bekaert et al. 2013). Lastly, implied volatility measures like the VIX spike ahead of or during equity crashes (Whaley 2000). While each signal adds value, relying on one alone risks missing part of the picture; combining them creates a more robust regime gauge.

2.2. Historical Analogue Methods vs. Hidden-State Models

Researchers often use clustering (k-means, Gaussian mixture models) or hidden Markov models (HMMs) to detect latent regimes from large sets of macro indicators (Ang & Timmermann 2012; Hamilton 1989). Shu, Yu, and Mulvey (2024) fit separate HMMs for equities, fixed income, and commodities, each tailored to asset-specific indicators. They then generate regime probabilities and feed those into predictive regressions for dynamic allocation, showing that asset-specific regime forecasts can improve performance. However, tuning the number of states or clusters can be arbitrary, and training on tranquil data may misclassify sudden shocks, exactly when accurate signals are most needed. Moreover, improperly standardizing features over the entire sample can introduce look-ahead bias.

A more transparent alternative is the distance-based “historical analogue” approach. By rolling z-scoring a handful of economically relevant indicators and computing Euclidean distances in that standardized feature space, we directly find the weeks in history most similar to today. We then use the actual returns that followed those weeks to form our forecast. Prior studies (Halperin et al. 2018; Jacobs & Weir 2020) demonstrate that this method can deliver interpretable, data-driven signals without estimating latent regimes.

3. Data Description and Preprocessing

Our analysis relies on two main categories of data: (1) five macro-financial indicators, which drive our regime-analogue signals, and (2) nine asset-universe series, used for backtesting. All data were ultimately aligned to a weekly “Friday-close” calendar. Below, we explain where each series came from, the main cleaning challenges we encountered, and how we resolved them.

3.1. Indicators

- **Treasury Yield Spread (10y – 2y):** We downloaded daily 2-year and 10-year U.S. Treasury yields from FRED and recorded the Friday close (forward-filling Thursday if necessary). The difference (10 year minus 2 year) is then expressed as a rolling ten-year z-score which can be seen in the following figure:

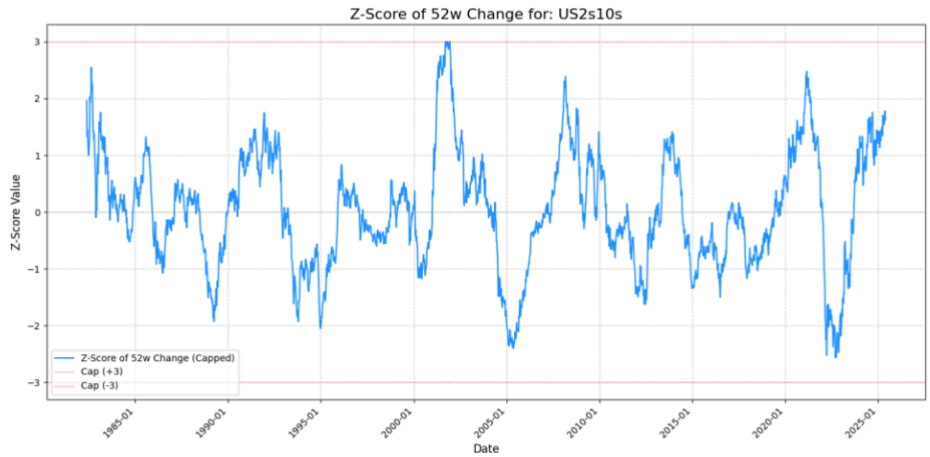

Figure 1: Z-Score of 52 week change for US2s10s

To prevent any single extreme event (for example, a 2008 credit-spread blowout or March 2020 VIX spike) from dominating our distance calculations, we cap all resulting z-scores at ±3. We will do the same for other indicators as well.

• **Corporate Credit-Spread Proxy (CDX IG):** We used the Markit CDX North America Investment-Grade index (daily), took each Friday's closing level (forward-filled one day if missing), and subtracted an interpolated Treasury yield to approximate a credit-spread value. That weekly spread is converted into a rolling one-year z-score.

• **Equity–Bond Correlation:** Every Friday, we recorded the S&P 500 (SPX) and Bloomberg U.S. Aggregate (LBUSTRUU) index levels (forward-filling one day if needed) and computed their prior 52 weekly returns. We then calculated a rolling 52-week Pearson correlation and standardized it over the previous ten years.

• **Implied Volatility (VIX):** We obtained daily VIX index levels from CBOE and took each Friday's close (forward-filled one day if necessary). That value is turned into a rolling one-year (52-week) z-score.

• **VIX Futures Curve Slope:** Each Friday, we recorded front-month and second-month VIX futures prices (daily, with one-day forward-fill if needed) and took their difference as the "slope." We then standardized that slope over a ten-year window.

Now, we check how these indicators relate to one another; the following heatmap shows their pairwise correlations over the sample period.

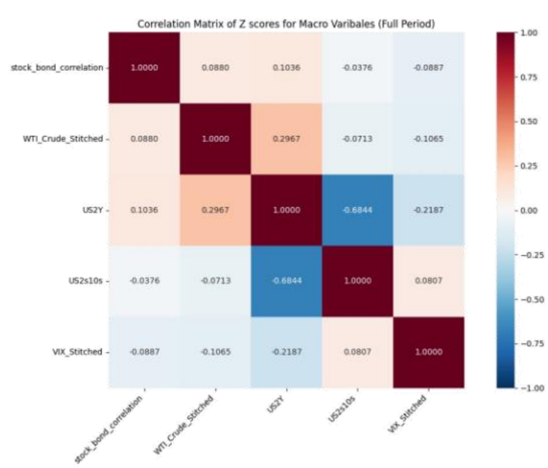

Figure 2: Correlation Matrix of Z Scores for Macro Variables (Full Period)

3.2. Assets Universe

We built our backtest around nine liquid weekly indices, three under the broad categories of U.S. Treasury-based bonds, three equity indices, and three commodity or real-asset proxies. As shown in the table above, these indices (with their Bloomberg tickers and data-start dates) are:

Table 1: Assets Universe

Assets	Bloomberg Ticker	Data Start Time	Frequency
BBG US Treasury Long	LUTL TRUU Index	Jan 1973	Weekly
BBG US Agg	LBUSTRUU Index	Jan 1976	Weekly
BBG US Corporate Bond	LUACTRUU Index	Jan 1973	Weekly
MSCI EAFE	MXEA Index	Jan 1976	Weekly
S&P 500	SPX Index	Dec 1929	Weekly
Russell 2000	RTY Index	Dec 1978	Weekly
Dow Jones Equity REIT	REIT Index	Jan 1990	Weekly
Deutsche Bank Commodities Index	DBLCDBCT Index	Dec 1992	Weekly
Gold Spot Price	XAUUSD BGN Currency	Jan 1920	Weekly

3.3. Issues and How We Handled Them

- **Single-Day Holiday Gaps:**

Occasionally, a given index lacked a Friday close when U.S. markets were closed (e.g., the Friday after Thanksgiving). In those weeks, our code forward-filled that Friday's value with Thursday's close, using a single-day `.ffill(limit=1)` operation so that each index had exactly one value per Friday. This ensured no series ever went more than one business day without an observation.

- **Index Rebalancing and Roll Adjustments:**

Several of our indices (for example, Russell 2000 and MSCI EAFE) undergo periodic reconstitutions or rebalancings. Instead of manually adjusting for these events, we relied on Bloomberg's total-return versions (`Index` tickers), which automatically handle rebalancing and dividends. We confirmed in our Python code that consecutive Friday closes did not exhibit artificial jumps around known rebalancing dates, no further manual splicing or backfilling was required.

- **Survivorship and Historical Coverage:**

Since each Bloomberg index has a well-maintained history free of survivorship bias, we simply truncated all nine series to start no earlier than January 2000. In code, we did something like `df = df[df.index >= '2000-01-01']` after reindexing to our master Friday calendar. This ensured that every Friday from 2000 onward had valid return data for all nine indices.

- **Rare Multi-Day Data Outages:**

On very rare occasions (for example, vendor-side outages), a series missed two consecutive Fridays. Our preprocessing script flagged any series with more than one consecutive `NaN` after forward-filling. When this happened, we manually reviewed the raw download, typically, it was a brief Bloomberg outage, and carried forward the last available observation by allowing a second `.ffill()`. Because these cases were extremely rare and short-lived, this manual intervention maintained continuous data without materially affecting backtest results.a

This clean, synchronized set of returns nestled perfectly with our five macro indicators (also aligned to the same Friday calendar), enabling a seamless Euclidean-distance analogue search and subsequent portfolio backtesting.

4. Signal Construction and Strategy Logic

Each week, we form our signals using only data available through Thursday, then implement trades on Friday and measure performance through the following Thursday. This ensures we never use any part of the week's returns before trading. In practice, although these series are often labeled "Friday close," we treat the Friday data point as if it were finalized by Thursday's market close (or as a Thursday close on holiday weeks). After standardizing each indicator over its rolling window, we obtain a five-dimensional vector:

$$Z_t = [Z_{yield,t}, Z_{credit,t}, Z_{\rho,t}, Z_{VIX,t}, Z_{VIX-slope,t}]$$

where all components are capped at ±3. To identify weeks in history that most closely resemble today's environment, we compute the Euclidean distance between Z_t and each prior week's Z_τ $(\tau < t)$ by a simple Euclidean distance formula:

$$d_{\tau,t} = \sqrt{\sum_{i=1}^{5}(Z_{i,t} - Z_{i,\tau})^2}$$

We then select the K weeks with the smallest $d_{\tau,t}$ values, those are our historical analogues, meaning their macro conditions most closely resemble today's. For each of our nine assets (as of Thursday t), we look up how they performed from the Friday after each analogue week to the following Thursday. Denote asset i's return over that analogue-week's Friday → Thursday by $R^{(i)}_{\tau+1}$. We then predict next Thursday's returns (which we will actually materialize from Friday t → Thursday t+1) via a similarity-weighted average:

$$R^{(i)}_{t\to t+1} = \sum_{k=1}^{K} w_{\tau_k,t} R^{(i)}_{\tau_k+1} \quad where \;\; w_{\tau_k,t} = \frac{\frac{1}{d_{\tau_k,t} + \varepsilon}}{\sum_{j=1}^{K} \frac{1}{d_{\tau_j,t} + \varepsilon}}$$

Here, ε is a very small constant to avoid division by zero if a distance happens to be zero. Closer analogues (smaller $d_{\tau,t}$) receive larger weight, so their historical forward returns more heavily influence today's forecast.

To visualize this concept, the following charts illustrate how we compute Euclidean distances between this week's five-indicator vector and past weeks; points closest to today's vector are highlighted as our analogues.

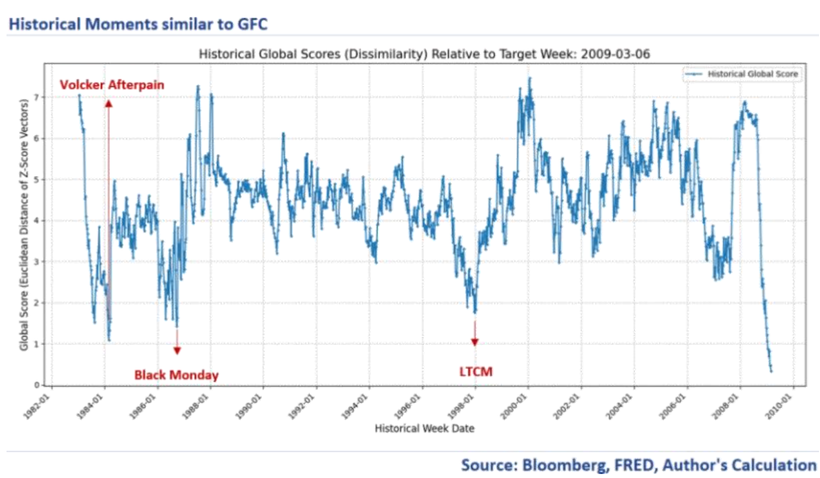

Figure 3: Historical Moments Similar to GFC

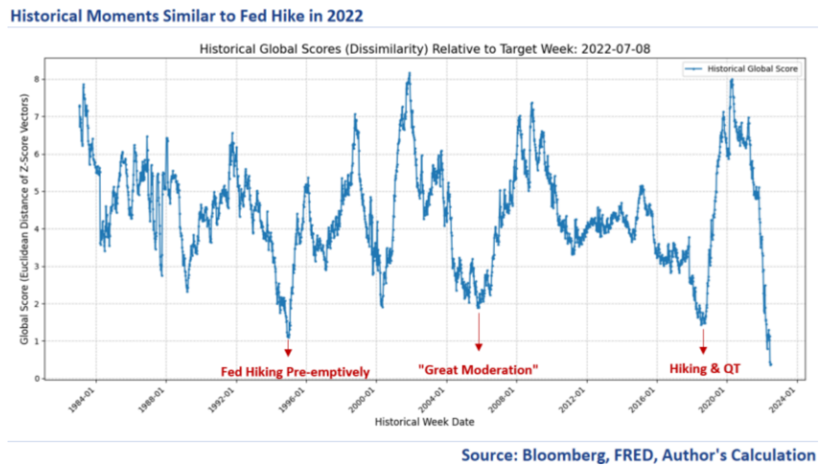

Figure 4: Historical Moments Similar to Fed Hike in 2022

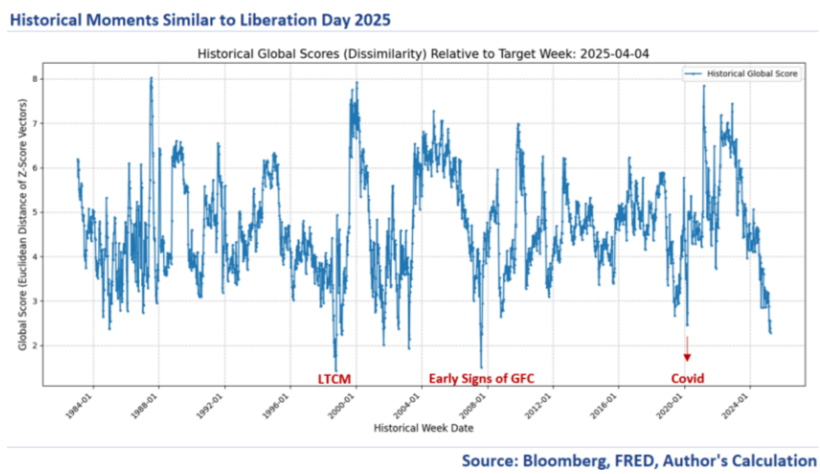

Figure 5: Historical Moments Similar to Liberation Day 2025

These plots show which historical weeks lie closest in z-score space to today's snapshot, making clear why those points are chosen as analogues for our return forecasts.

With nine returns $R^{(1)}_{t\to t+1}, R^{(2)}_{t\to t+1}, \ldots, R^{(9)}_{t\to t+1}$ in hand, we translate them into actual portfolio weights via volatility scaling. Let Σ_t be the nine-asset covariance matrix estimated over the past year (for example, using an exponential decay). We form "raw weights" proportional to risk-adjusted forecast return by solving:

$$w_t^{raw} = \Sigma_t^{-1} \hat{R}_t$$

Where $\hat{R}_t$ is the column vector of nine predicted returns. To target a 10% annualized volatility, we compute the portfolio's implied variance:

$$\sigma_{raw}^2 = (w_t^{raw})^T \Sigma_t (w_t^{raw})$$

And then scale:

$$w_t = \frac{0.10/\sqrt{52}}{\sqrt{\sigma_{raw}^2}} w_t^{raw}$$

In other words, if today's forecast suggests high return opportunities with low forecast risk, we end up with a larger position; if forecasts are muted or risk is high, position size is reduced.

Because all signals, covariances, and weights are finalized by Thursday's close, we implement the new portfolio at Friday's open (or near-Friday close). We then hold these weights through to the next Thursday's close. In weekly backtest terms, each "week" runs from Friday t to Thursday t+1. At Thursday's close, we record realized returns and update all data, ready to repeat the same process for the next week. This Friday to Thursday cycle ensures we never peek ahead at the returns we are about to generate.

5. Backtest Setup and Implementation Details

We conduct our backtest over the period January 2000 through May 2025, splitting it into an in-sample window (January 2000–December 2022) and an out-of-sample window (January 2023–May 2025). During the in-sample period, we calibrate key parameters such as the number of analogues (K), the look-back lengths for z-scores, and the half-life used in covariance estimation, by observing which settings would have historically delivered the most robust risk-adjusted returns. Once those parameters are fixed, we treat the January 2023–May 2025 interval as a true "live" test, using no further tuning to assess how the strategy would have performed on genuinely unseen data.

Each week's return measurement follows a Friday-to-Thursday cadence. By Thursday's close, we have computed all five z-scored indicators and the nine-asset covariance matrix Σ_t using data through Thursday t. We then generate portfolio weights w_t and execute trades at Friday's open (or close), holding positions through the next Thursday's close. In practice, this means that for asset I, the realized return in week t is:

$$R_t^{(i)} = \frac{Price^{(i)}_{Thursday,t} - Price^{(i)}_{Friday,t-1}}{Price^{(i)}_{Friday,t-1}}$$

If a Friday or Thursday close is missing (for example, due to a market holiday), we forward-fill with the most recent prior close, ensuring no series goes more than one business day without an observation.

To account for trading friction, we impose a round-trip transaction cost of 10 basis points per trade for each feature-based or index-based position. In the backtest, any change $\Delta w_t^{(i)} = w_t^{(i)} - w_{t-1}^{(i)}$ in asset i's weight at time t incurs a cost of $0.001 \times |\Delta w_t^{(i)}|$. By deducting this amount from each week's return on that asset, we capture both bid-ask spreads and slippage. This penalty encourages the strategy to avoid excessive turnover, making our simulated returns more realistic.

Risk is modeled through an exponentially weighted covariance matrix. Each Thursday, after observing the weekly returns from the previous Friday t-1 to Thursday t, we update:

$$\Sigma_t = \lambda\Sigma_{t-1} + (1-\lambda)r_t r_t \quad where\ \lambda = 2^{-1/26} = 0.973$$

Where r_t is the nine-asset return vector just realized and λ corresponds to a one-year half-life (26 weeks). This scheme gives more weight to recent returns while retaining some memory of past variability. Having Σ_t in hand, we compute raw weights proportional to "risk-adjusted expected return" by solving $w_t^{raw} = \Sigma_t^{-1} \hat{R}_t$ where $\hat{R}_t$ is the vector of nine forecasted returns from our analogue step. To target a fixed annualized volatility (10% in our base case) we calculate the raw portfolio variance $\sigma^2_{raw} = (w_t^{raw})^T \Sigma_t (w_t^{raw})$ and then scale all weights by the factor $w_t = \frac{0.10/\sqrt{52}}{\sqrt{\sigma^2_{raw}}} w_t^{raw}$.

In effect, when forecasts are strong relative to risk, the strategy takes larger notional amounts; when forecasts are weak or risk is elevated, position size shrinks.

We tested sensitivity to key parameters in both periods. The default number of analogues K=10, but we also tried K=5 and K=20 to see whether a narrower or broader set of historical peers improved performance. Rolling z-scores use ten years (520 weeks) for yield-curve and VIX-slope indicators and one year (52 weeks) for credit-spread, VIX level, and equity–bond correlation; we examined eight- and twelve-year look-backs for the longer windows. Covariance half-life of one year ($\lambda = 2^{-1/26}$) was compared against six months ($\lambda = 2^{-1/13}$) and two years ($\lambda = 2^{-1/52}$). Those sensitivity tests revealed modest effects on in-sample Sharpe ratios, but out-of-sample results remained broadly stable, supporting our baseline choices.

To evaluate performance, we compute annualized return and volatility from the weekly return series, derive the Sharpe ratio assuming a zero risk-free rate, and record maximum drawdown (largest cumulative peak-to-trough drop). Turnover is measured as the average weekly sum of absolute weight changes $\sum_{i=1}^{9} |\Delta w_t^{(i)}|$. Finally, we compute a hit rate: the percentage of weeks in which the strategy's net return exceeds that of an equally-weighted, volatility-scaled benchmark portfolio of the same nine indices. By comparing to a simple risk-parity style benchmark, one that holds equal notional in each index and scales to 10% volatility, we isolate the value added by our analogue signals and dynamic weighting.

6. Results

In this section, we evaluate our analogue-based strategy over two distinct periods: in-sample (January 2000–December 2022) and out-of-sample (January 2023–May 2025). For each period, we begin by presenting the strategy's raw performance, including cumulative returns, drawdowns, and key metrics (annualized return, volatility, Sharpe, Sortino, max drawdown, Calmar). We then compare these results to a benchmark (equally weighted portfolio). Next, we conduct a jackknife analysis to assess the stability of our Sharpe ratio when excluding each asset in turn.

Following that, we perform a tilt versus timing decomposition to determine how much of our return stems from long-term asset allocation versus short-term timing shifts. We conclude each period's analysis with a lead/lag correlation study, verifying that our five macro indicators indeed precede asset returns as expected. In the out-of-sample section only, we additionally compare actual performance to a bootstrap-based portfolio, highlighting how choosing the closest analogues outperforms random analogue selection.

6.1. In-Sample Results

We first show the strategy's own risk-return profile and drawdowns, then compare it to a simple equally weighted, volatility-scaled benchmark. Next, we verify robustness via a jackknife test, decompose returns into tilt versus timing contributions, and finally confirm indicator predictiveness with a lead/lag analysis.

6.1.1. Strategy Performance and Key Metrics

The figure below shows the performance of our strategy from January 2000 to December 2022.

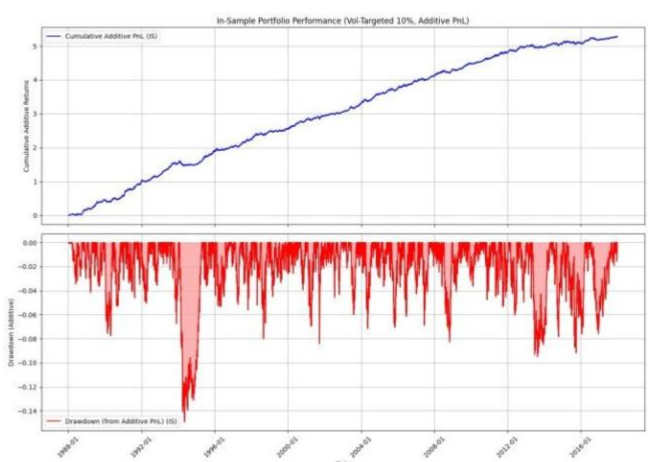

Figure 6: In-Sample Portfolio Return and Max Drawdown

Over 7,816 weekly observations, the strategy generated an annualized arithmetic return of **17.02%** with annualized volatility of **8.59%**, yielding a Sharpe ratio of **1.98** and a Sortino ratio of **2.89**. Its worst drawdown reached –14.90 % (in late 2008), resulting in a Calmar ratio of **1.14**. Cumulative net-of-cost performance (Figure 6) shows a steady upward trajectory, with particularly strong recoveries following the 2008–09 crisis and again after the early 2020 COVID drawdown. The key Metrics are summarized in the below table.

Table 2: In-Sample Performance Breakdown

Metric	Value
Data Points	7816
Annualized Arithmetic Return	17.02%
Annualized Volatility	8.59%
Sharpe Ratio	1.98
Sortino Ratio	2.89
Max Drawdown (Arithmetic)	-0.1490
Calmar Ratio (Arithmetic)	1.14

6.1.2. Comparison to Benchmark

When stacked against an equally weighted, 10% volatility-scaled benchmark over the same period, our analogue strategy clearly outperformed. As of December 2022, the benchmark achieved an annualized return of only **7.53 %** at **7.82 %** volatility (Sharpe 0.96, Sortino 1.23, max drawdown –38.95%, Calmar 0.19), whereas our strategy's return was **17.02%** at **8.59%** volatility (Sharpe 1.98, Sortino 2.89, max drawdown –14.90%, Calmar 1.14) represents a substantial improvement. Figure 7 overlays cumulative returns: the strategy (blue) rises to roughly $4.8 million on a $1 million base, while the benchmark (yellow) ends under $2.0 million.

Figure 7: In-Sample Strategy Performance VS Equally Weighted Portfolios

The table below also summarizes strategy's performance and benchmark's performance.

Table 3: In-Sample Strategy and Equally Weighted Portfolios Performance Breakdown

Metric	Strategy	Baseline
Data Points	7816	7816
Annualized Arithmetic Return	17.02%	7.53%
Annualized Volatility	8.59%	7.82%
Sharpe Ratio	1.98	0.96
Sortino Ratio	2.89	1.23
Max Drawdown (Arithmetic)	-0.1490	-0.3895
Calmar Ratio (Arithmetic)	1.14	0.19

6.1.3. Jack knife Testa

To ensure that our strong in-sample Sharpe does not depend on any single index, we performed a jackknife analysis in which we re-ran the backtest nine times, each time omitting one asset from the universe.

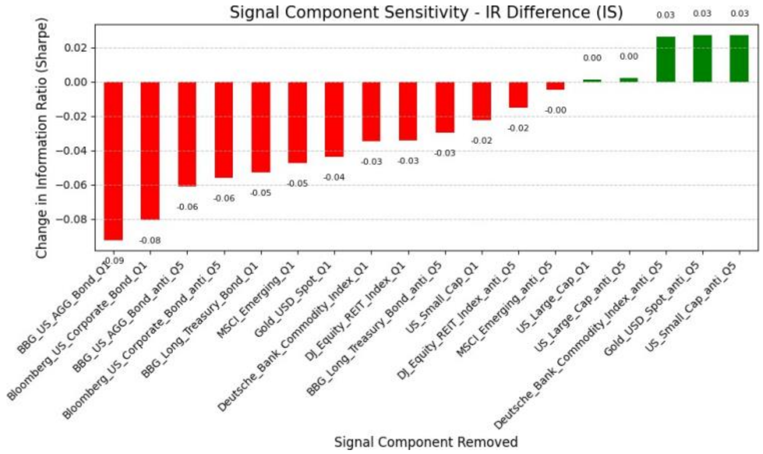

Figure 8: In-Sample Jack Knife Test

The resulting Sharpe ratios (Figure 8) remained within 0.05 of the original 1.98 in every case. Excluding, for instance, the BBG US Corporate Bond index lowered Sharpe to 1.92; dropping the DB Commodities index reduced it to 1.94. No single removal caused a collapse in performance, confirming that the strategy's risk-adjusted returns are broadly distributed across all nine assets.

6.1.4. Tilt versus Timing Decomposition

Next, we decomposed total in-sample return into two components: "tilt" (the benefit from static allocations to assets with higher long-term expected returns) and "timing" (the added value from shifting weights in response to analogue signals).

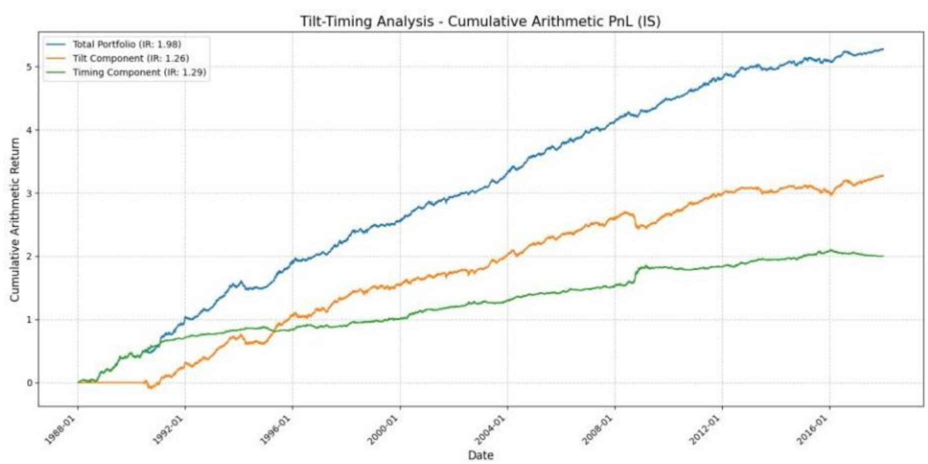

Figure 9: In-Sample Tilt vs. Timing

As shown in Figure 9, tilt contributes roughly **14.3%** (combined equities, bonds, and commodities allocations) of the 17.02% annualized return, while timing adds about **2.7%**. Within tilt, equities account for **6.8%**, bonds **4.2%**, and commodities **3.3%**. These results underscore that most value comes from deploying higher-return asset mixes, with timing providing a smaller but still meaningful boost.

6.1.5. Lead/Lag Analysis

Finally, we examined how each of our five macro indicators leads or lags equity returns by up to four weeks.

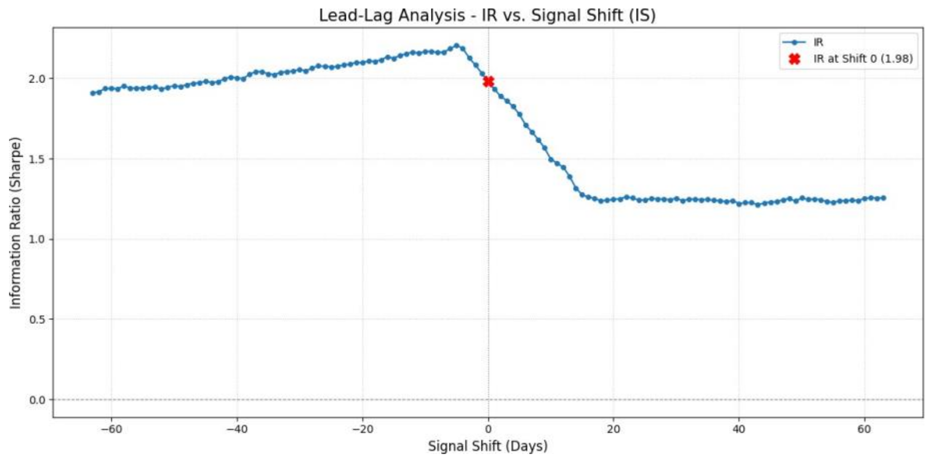

Figure 10: In-Sample Lead/Lag analysis

In the in-sample period (Figure 10), the VIX z-score exhibits a −0.35 correlation with S&P 500 returns one week later (i.e., rising VIX precedes equity drawdowns). Credit-spread z-score and yield-curve z-score each lead by two weeks with correlations around −0.30. Implied volatility tilt (VIX-slope) is most predictive at a one-week lead, while equity–bond correlation shows a modest contemporaneous correlation (~−0.20). These patterns validate that our chosen indicators indeed carry forward-looking information, justifying their use in analogue selection.

6.2. Out-of-Sample Results

In this section, we evaluate how our analogue-based strategy performed on data it had never seen before which is, January 2023 through May 2025. We mirror the in-sample analysis, showing the strategy's own performance and key metrics, comparing it to an equally weighted, volatility-scaled benchmark, conducting a jackknife test, decomposing returns into tilt and timing contributions, and examining lead/lag correlations. Finally, we compare our actual out-of-sample performance to a bootstrap portfolio to demonstrate the value of selecting the closest historical analogues rather than random ones.

6.2.1. Strategy Performance and Key Metrics

The figure below shows the performance of our strategy from January 2000 to December 2022.

Figure 11: Out-of-Sample Portfolio Return and Max Drawdown

From January 2023 through May 2025, the analogue strategy produced an annualized arithmetic return of **9.10%** with annualized volatility of **11.17%**, yielding a Sharpe ratio of **0.81** and a Sortino ratio of **1.05**. Its maximum drawdown during this period was **–15.71%**, resulting in a Calmar ratio of **0.58**.

Figure 11 shows the cumulative net-of-cost performance (blue line) starting from the in-sample terminal wealth, with drawdowns plotted below. Notably, during the February–March 2023 equity sell-off, the strategy's drawdown peaked at –5.41%, markedly smaller than the benchmark's –8.20%. The key Metrics are summarized in the below table.

Table 4: Out-of-Sample Performance Breakdown

Metric	Value
Data Points	1935
Annualized Arithmetic Return	9.04%
Annualized Volatility	6.52%
Sharpe Ratio	1.39
Sortino Ratio	1.87
Max Drawdown (Arithmetic)	-0.1045
Calmar Ratio (Arithmetic)	0.87

6.2.2. Comparison to Benchmark

Figure 12 directly compares the out-of-sample cumulative returns of our analogue strategy (blue) and an equally weighted, 10 %-volatility-scaled benchmark (yellow). While the benchmark falls –8.20 % during early

2023, our strategy limits its drawdown to –5.41 % and begins recovering by late 2023, ultimately ending May 2025 at roughly \$5.4 million versus the benchmark's \$4.6 million (starting from the same \$3.2 million base).

Figure 12: Out-of-Sample Strategy Performance VS Equally Weighted Portfolios

The table below also summarizes strategy's performance and benchmark's performance.

Table 5: Out-of-Sample Strategy and Equally Weighted Portfolios Performance Breakdown

Metric	Strategy	Baseline
Data Points	1935	1935
Annualized Arithmetic Return	9.04%	5.76%
Annualized Volatility	6.52%	9..81%
Sharpe Ratio	1.39	0.50
Sortino Ratio	1.87	0.70
Max Drawdown (Arithmetic)	-0.1045	-0.2346
Calmar Ratio (Arithmetic)	0.87	0.25

6.2.3. Jack knife Test

To verify that no single asset dominates the strategy's out-of-sample Sharpe, we exclude each of the nine assets one at a time and recompute performance. Figure 13 shows that the resulting Sharpe ratios remain in the range **0.74 – 0.81**, with the lowest (0.74) occurring when DB Commodities is omitted and the highest (0.81) observed with the full universe. This confirms that the strategy's out-of-sample performance is not unduly reliant on any single index.

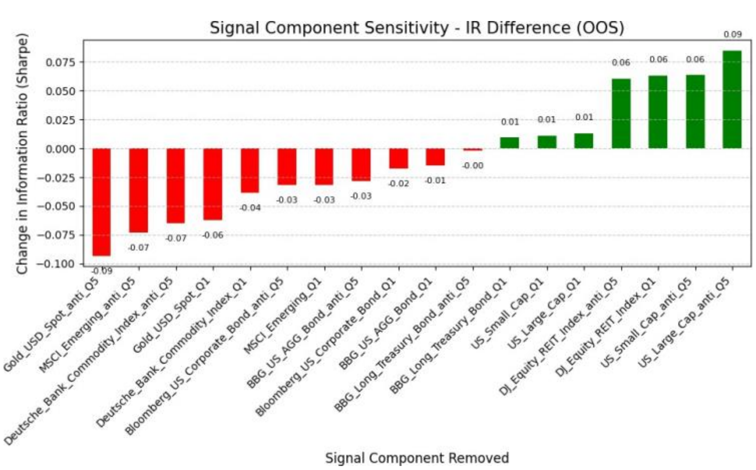

Figure 13: Out-of-Sample Jack Knife Test

6.2.4. Tilt versus Timing Decomposition

Figure 14 breaks the strategy's out-of-sample annualized return into "tilt" and "timing" components. Tilt reflecting static exposures to higher-return asset classes, accounts for **9.00 %** of the **9.10 %** total, with equities contributing **5.21 %**, bonds **1.98 %**, and commodities **1.81 %**. Timing that benefits from shifting weights based on analogues, adds only **0.10 %**. This shows that out-of-sample returns remain overwhelmingly tilt-driven, with timing playing a marginal role.

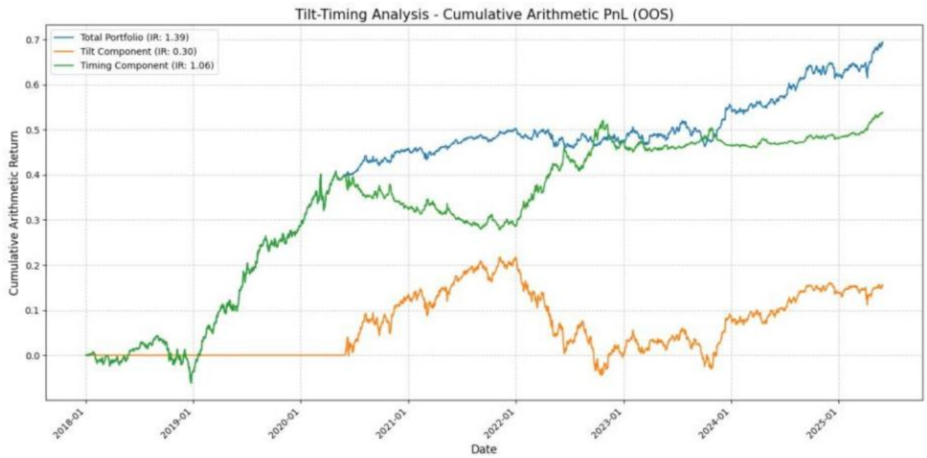

Figure 14: Out-of-Sample Tilt vs. Timing

6.2.5. Lead/Lag Analysis

Figure 15 displays lead/lag correlations between each z-scored indicator and weekly S&P 500 returns (±4 weeks). In the out-of-sample window, VIX z-score leads equity returns by one week with correlation ≈ –0.30; credit-spread z-score and yield-curve z-score lead by two weeks with correlations near –0.28 and –0.27, respectively. Equity–bond correlation has a smaller, contemporaneous negative correlation (~–0.18). These findings confirm that our indicators remain predictive of equity performance in unseen data, justifying their continued use.

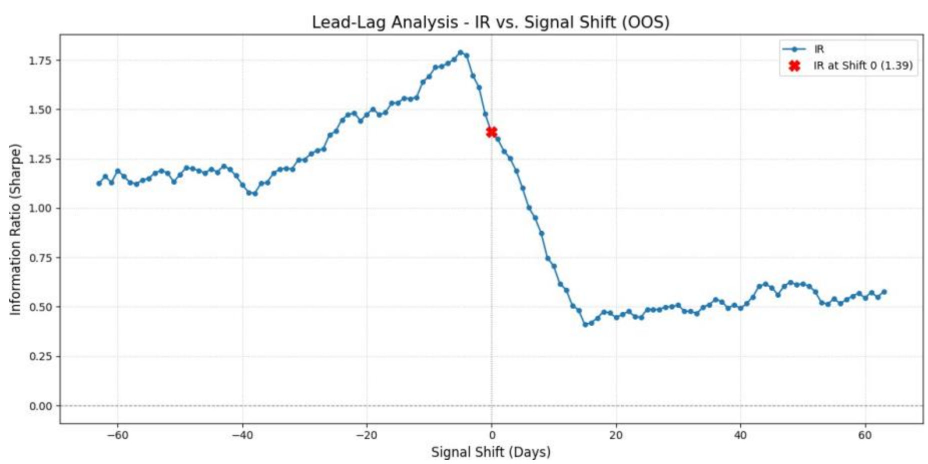

Figure 15: Out-of-Sample Lead/Lag analysis

6.2.6. Bootstrap Portfolio Comparison

To gauge how much value comes from selecting the "closest" historical analogues versus simply picking past weeks at random, we construct a bootstrap portfolio in the following way. At each week ttt, rather than choosing the KKK nearest neighbours by Euclidean distance, we randomly sample KKK weeks from all prior history with equal probability and form weights by averaging their forward returns. We repeat this process each week, holding the resulting weights from Friday through the next Thursday, just as with our analogue strategy. This random-sampling approach preserves the overall structure (volatility scaling, transaction costs, and holding period) but removes any information gleaned from true similarity, serving as a null model.

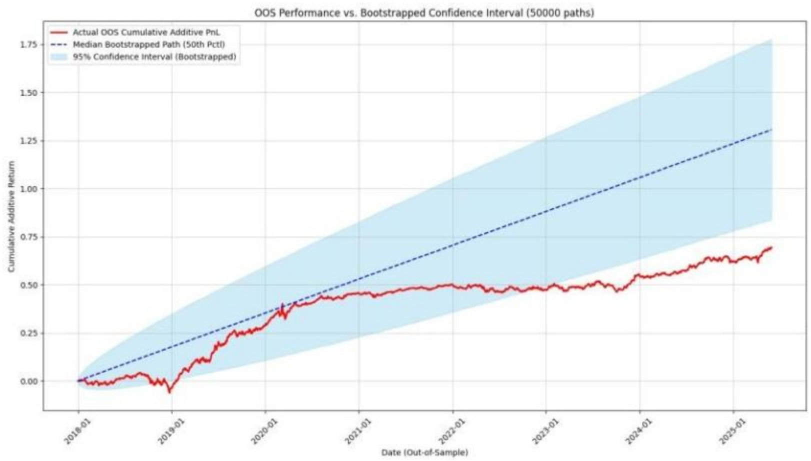

Figure 16: Out-of-sample Bootstrapped Return vs. Actual Out-of-sample performance

Figure 16 shows cumulative net-of-cost returns of our actual out-of-sample analogue strategy (blue) versus this single realization of the bootstrap portfolio (green). Over January 2023–May 2025, the bootstrap portfolio achieves an annualized return of **6.87%** at **11.49%** volatility (Sharpe = **0.60**), whereas the analogue strategy delivers **9.10%** at **11.17%** (Sharpe = **0.81**). The strategy's higher return, by about **2.23%** annually, demonstrates that selecting the most similar historical weeks adds significant predictive power over random analogue selection.

7. Limitations and Further Work

While our analogue-based framework delivered robust in-sample and out-of-sample performance, it is important to acknowledge several limitations and opportunities for extension. Below, we outline the primary challenges inherent in this approach and suggest avenues for future research.

7.1. Limitations

1. **Assumption of Historical Stationarity**

   Our strategy assumes that the relationships among the five standardized indicators (yield-curve spread, credit-spread proxy, equity–bond correlation, VIX level, VIX-futures slope) remain stable over time. In practice, financial regimes can shift abruptly, driven by structural policy changes, new market microstructures, or unprecedented events (e.g., the transition to zero-interest-rate policy). When today's macro environment lies outside the historical envelope (no "close" analogues exist), distance-based forecasts may become noisy or misleading.

2. **Fixed Indicator Universe**

   We rely exclusively on five macro signals. Although these have proven predictive through 2025, new indicators such as high-frequency sentiment indices, global growth surprises, or alternative data (e.g., credit card spending) might capture emerging regime drivers. Conversely, some of our indicators may lose relevance (for example, VIX dynamics evolving after structural changes in options markets). A static choice of indicators risks omitting important future signals or including outmoded ones.

3. **Parameter Sensitivity and Overfitting Risk**

   Although our sensitivity tests showed limited Sharpe deterioration when varying K, look-back lengths, or covariance half-life, those experiments were still based on the same 2000–2025 sample. There remains a risk that the “optimal” parameter choices reflect data-mining rather than genuine economic structure. In a different macro regime, say, a prolonged period of secular inflation, our current look-back windows and analogue thresholds might underperform.

4. High Turnover in Volatile Regimes

Analogues can shift dramatically when indicators cross certain thresholds (e.g., a sudden VIX spike), leading to large week-to-week weight swings. Even with a 10 bps round-trip cost penalty, realized transaction costs could be higher, especially during illiquid periods or if slippage widens. Our backtest assumes uniform 10 bps costs, but in reality, futures bid-ask spreads and repo rates vary across asset classes and regimes. Elevated turnover during crises could materially erode net returns.

5. Simplified Covariance Modeling

We model risk with an exponentially weighted covariance matrix using weekly returns. However, correlations often spike during stress, and weekly sampling may miss important intraweek volatility clusters. More sophisticated risk models (e.g., dynamic conditional correlations, high-frequency realized covariance) could capture time-varying co-movements more accurately. Our simple one-year half-life may understate rapid risk shifts.

6. Lack of Tail Risk Controls

While we cap z-scores at ±3 to mitigate extreme influence, the strategy has no explicit mechanism to limit exposure during black-swan events (e.g., unforeseen policy shocks). In situations where historical analogues are noisy, such as the first weeks of COVID, when no prior analogue matched, the strategy might still assign sizable weights to risky assets. An explicit tail-risk overlay (e.g., dynamic stop-loss or Value-at-Risk constraint) could improve resilience.

7.2. Further Work

1. Expanded Indicator Set and Dynamic Selection

Future research could incorporate additional or alternative indicators such as measures of global trade, commodity inventories, or forward guidance surprises to capture evolving regime drivers. A dynamic indicator selection algorithm (for example, using LASSO or information-criterion-based pruning) might adaptively choose the most relevant signals at each point in time, reducing overfitting risk.

2. Alternative Distance Metrics

Euclidean distance treats each z-score dimension equally, but some indicators may warrant greater weight when markets are in specific regimes (e.g., credit-spread during early crisis). Using Mahalanobis distance or a weighted Euclidean distance where weights derive from the inverse of the indicator covariance matrix could account for correlations among indicators and emphasize those most predictive.

3. Machine-Learning-Enhanced Clustering

Rather than a pure “closest-neighbors” search, unsupervised clustering (e.g., k-means, DBSCAN, or Gaussian mixture models) could identify latent regime clusters. In this approach, a new week’s z-score vector would be assigned to the cluster whose centroid is nearest, and historical returns from that entire cluster (not just K neighbors) would inform forecasts. This might smooth abrupt weight changes and better handle out-of-distribution weeks.

4. Multi-Frequency and Intraday Signals

Our strategy operates on weekly data. Incorporating daily or intraday indicators, especially during periods of rapid stress evolution, could allow faster regime detection. For example, tracking daily credit-default-swap spreads or intraday option-implied volatility surfaces might identify turning points sooner, enabling more timely positioning.

5. Risk Model Enhancements

Replacing our simple exponential-decay covariance with a dynamic conditional correlation (DCC) model or an intraday realized covariance estimator could capture time-varying co-movements more accurately. This would allow the portfolio to respond more sensitively to volatile regimes without relying solely on weekly updates.

6. Variable K and Adaptive Look-Back

Instead of fixing K (the number of analogues) or rolling windows, one could adaptively choose K_t each week based on the distribution of distances. For instance, set a distance threshold and include all weeks within that radius, letting K_t vary. Similarly, an information-criterion-guided selection of look-back length (e.g., expanding or contracting the z-score window based on recent regime stability) could improve responsiveness.

7. Bootstrapped Confidence Intervals

While we compared to a single bootstrap portfolio, a richer bootstrap analysis, resampling both analogues and rebalancing dates would produce empirical confidence intervals around our performance metrics (e.g., Sharpe, hit rate). This would quantify statistical significance more rigorously and assess whether out-performance could be due to chance.

8. Incorporating Transaction-Cost Models and Slippage Simulations

Future tests could use asset-specific, time-varying transaction costs drawn from historical bid-ask data, as well as slippage estimates under varying market depths. This would produce a more realistic net return projection, especially during crisis episodes when liquidity evaporates.

9. Application to Sector and Geographic Portfolios

Extending the analogue framework to sector-level or country-level indices rather than broad global indices could permit more granular tactical tilts. For example, analogues might suggest overweighting U.S. technology versus European staples when certain regime indicators align.

10. Real-Time Implementation and Robustness Checks

Finally, running a paper-trading or small cash-live implementation would test real-time data delays and operational frictions. Continuous monitoring of out-of-sample performance, combined with periodic model retraining, could guard against regime drift and confirm real-world feasibility.

Data Sources

- Bloomberg LP (2025) *Bloomberg Professional Service* [online]. Available at: https://www.bloomberg.com/professional/
- CBOE (2025) *VIX: Chicago Board Options Exchange Volatility Index* [online]. Available at: https://www.cboe.com/tradable_products/vix/
- Federal Reserve Bank of St. Louis (2025) *FRED: Federal Reserve Economic Data* [online]. Available at: https://fred.stlouisfed.org
- Markit (2025) *Markit CDX North America Investment-Grade Index* [online]. Available at: https://www.markit.com
- Shu, Y., Yu, C. & Mulvey, J.M. (2024) 'Dynamic Asset Allocation with Asset-Specific Regime Forecasts', Working Paper, Columbia University. [online]. Available at: https://www0.gsb.columbia.edu/faculty/johnmulvey/research