A Multi-Alpha Volatility-Driven Trading Strategy: Evidence from the Chinese A-Share Market

Jinjia Peng
Department of Mathematics
North Carolina State University

June 2025

1 Preface

A multi-factor (multi-alpha) trading strategy is a systematic approach that relies on the statistical properties of historical data to generate trading signals and execute trades. The central idea is to identify signals that exhibit a persistent relationship with future returns and to exploit these signals through repeated, rule-based trading. A useful analogy is that of an unfair coin: a valid trading signal can be thought of as a coin slightly biased toward the profitable side. Over a sufficiently large number of trades, the law of large numbers ensures that realized returns converge toward the expected value, provided the signal has been rigorously verified to yield a positive expectation. Importantly, this relationship between signal and return may be either positive or negative; in the latter case, shorting the asset converts the relationship into a profitable one. The central challenge in quantitative trading lies in reliably identifying such "unfair coins" that provide a true statistical edge.

To investigate this concept, the present study reproduces all alphas from the well-known paper *101 Formulaic Alphas* and evaluates their effectiveness within the Chinese A-share market, specifically the Shanghai Stock Exchange and Shenzhen Stock Exchange. The period from 2020 to 2023 serves as the training window to filter effective alphas, while the strategy is backtested from January 2023 to June 2025 under a set of simplifying assumptions. The methodology involves calculating the information coefficient (IC) and rank information coefficient (RIC), performing hypothesis testing and cross-sectional regressions, allocating daily cash conditional on market volatility, and simulating a full long/short trading process.

The strategy is implemented with daily rebalancing and is subject to several limitations and constraints. Its primary purpose is to provide an entry-level demonstration of core quantitative trading concepts and to serve as a foundation for further exploration in alpha mining and financial engineering.

2 Data Preparation

2.1 Basic Logic

The first step in implementing the strategy is the systematic collection and organization of all necessary data. Since the strategy covers the entire A-share market, the resulting data matrices must have N columns, where N is the total number of listed stocks on the Shanghai Stock Exchange and Shenzhen Stock Exchange. The number of rows corresponds to the time horizon T, spanning from the beginning of the training period (January 1, 2020) to the end of the backtesting period (June 13, 2025). This unified matrix structure allows the same

dataset to be used for both parameter estimation and backtesting, reducing computational overhead by enabling efficient subsetting of the matrices.

Several categories of data are required for this process:

- Daily trading data (open price, close price, high, low, trading volume, turnover, etc.)
- Dividend and corporate action data (cash dividends, bonus issues, stock splits, etc.)
- Stock sector classifications (Shenwan Classification for Chinese equities; GICS for U.S. equities if extended)
- Shareholding and capitalization data (total market capitalization, free-float market capitalization, shares outstanding, etc.)

For Chinese equities, the standard classification scheme follows the **Shenwan** methodology, which divides companies hierarchically into sectors, industries, and subindustries. For U.S. equities, the analogous scheme is the **GICS** classification. These classifications are necessary inputs for the computation of formulaic alphas.

At the conclusion of the data preparation process, the key datasets are stored in matrix form, including:

- `close`
- `open`
- `high`
- `low`
- `volume`
- `amount`
- `turnover`
- `market_cap`
- `market_cap_float`
- `div`
- `bonus`
- `split`

- `sector`
- `industry`
- `subindustry`
- `log return (explained in Section 2.2)`

Below are some mathematical definitions and examples of the matrix formations for reference:

Example: Close Matrix

$$\texttt{close} = \begin{bmatrix} P_{t_1,1} & P_{t_1,2} & \cdots & P_{t_1,N} \\ P_{t_2,1} & P_{t_2,2} & \cdots & P_{t_2,N} \\ \vdots & \vdots & \ddots & \vdots \\ P_{t_T,1} & P_{t_T,2} & \cdots & P_{t_T,N} \end{bmatrix},$$

where $P_{t,i}$ denotes the close price of stock i at time t, $t_1 =$ 2020-01-02, $t_T =$ 2025-06-13, and N is the number of stocks in the sample.

Table 1: Illustrative slice of the close price matrix

Date	sh600000	sh600004	$\cdots$	sz301662
2020-01-02	12.47	17.52	$\cdots$	NaN
2020-01-03	12.60	17.38	$\cdots$	NaN
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
2025-06-12	12.46	9.37	$\cdots$	93.05
2025-06-13	12.34	9.33	$\cdots$	98.39

From the above table, the `NaN` represent the stocks where their close prices on that date were missing, potentially due to halt, merge and acquisition, not in the market yet, etc.. Filling these `NaN`s with zeros would affect the true results because this action forces the missing prices to be zero, in which it is not mathematically accurate.

Example: Dividend/Bonus/Split Matrices

The dividend/bonus/split matrices follows the same logic, as shown below (dividend matrix example):

$$\texttt{div} = \begin{bmatrix} D_{t_1,1} & D_{t_1,2} & \cdots & D_{t_1,N} \\ D_{t_2,1} & D_{t_2,2} & \cdots & D_{t_2,N} \\ \vdots & \vdots & \ddots & \vdots \\ D_{t_T,1} & D_{t_T,2} & \cdots & D_{t_T,N} \end{bmatrix},$$

where $D_{t,i}$ denotes the dividend applied to stock i at time t. The dimension should be the same with bonus and split matrices. As before, $t_1 =$ 2020-01-02 and $t_T =$ 2025-06-13.

Table 2: Illustrative slice of the dividend matrix

Date	sh600000	sh600004	$\cdots$	sz301662
2020-01-02	NaN	NaN	$\cdots$	NaN
2020-01-03	NaN	NaN	$\cdots$	NaN
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
2025-06-09	NaN	NaN	$\cdots$	0.4
2025-06-10	NaN	NaN	$\cdots$	1.0
2025-06-13	NaN	NaN	$\cdots$	NaN

From the above table, the entries are mostly `NaN`, indicating no dividend corporate actions occurred on those dates for the given stocks. Non-`NaN` entries correspond to actual corporate action events (e.g., a split ratio or cash dividend). Here `NaN` carries structural meaning: it denotes the absence of any corporate action, not missing data. Thus, imputing zeros or replacing these values would also be misleading, since the distinction between "no action" and "action with a zero value" is essential in return adjustment.

Example: Sector/Industry/Subindustry Matrices

The sector/industry/subindustry classification matrices have similar formats. An example of the sector matrix is shown below:

$$\texttt{sector} = \begin{bmatrix} S_{t_1,1} & S_{t_1,2} & \cdots & S_{t_1,N} \\ S_{t_2,1} & S_{t_2,2} & \cdots & S_{t_2,N} \\ \vdots & \vdots & \ddots & \vdots \\ S_{t_T,1} & S_{t_T,2} & \cdots & S_{t_T,N} \end{bmatrix},$$

where $S_{t,i}$ denotes the *sector* of stock i on date t.

Notes:

- `Unknown (UNK)` indicates unmapped or newly listed stocks without sector assignment.

Table 3: Illustrative slice of the `sector` matrix (English translation)

Date	sh600000	sh600004	$\cdots$	sz301665
2023-01-03	Banks	Transportation	$\cdots$	UNK
2023-01-04	Banks	Transportation	$\cdots$	UNK
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
2025-06-09	Banks	Transportation	$\cdots$	Basic Chemicals
2025-06-12	Banks	Transportation	$\cdots$	Basic Chemicals
2025-06-13	Banks	Transportation	$\cdots$	Basic Chemicals

- Sector labels (e.g., Banks, Transportation, Automobiles, Basic Chemicals) are translations of the Shenwan classification categories.
- The same logic applies to the `industry` and `subindustry` matrices: they are structured identically, differing only in classification granularity.

2.2 Return Adjustment

2.2.1 Properties of Log Return

Log return provides a very convenient way to calculate returns. Because of its mathematical properties, researchers can simply add up the rolling sum of log returns to calculate compound returns. Specifically, if R_t is the simple return at time t, then the cumulative return over T periods is:

$$\prod_{t=1}^{T}(1+R_t) = \exp\left(\sum_{t=1}^{T}\log(1+R_t)\right) \tag{1}$$

Thus, the strategy utilized log return to simplify the computation of compounded performance over multiple periods.

2.2.2 The Log Return Adjustment

For the computation of the log return matrix, all corporate actions must be taken into account. The calculations are derived from the previous dividend/bonus/split matrices. Let P_t denote the close price at time t, D_t the cash dividend per share, B_t the bonus share ratio, and S_t the split (conversion) ratio. Define the adjusted share capital factor as

$$F_t = 1 + B_t + S_t. \tag{2}$$

The gross return R_t between $t-1$ and t is then given by

$$1 + R_t = \frac{F_t P_t + D_t}{P_{t-1}}. \tag{3}$$

Accordingly, the log return is

$$\log(1+R_t) = \begin{cases} \log\left(\dfrac{P_t}{P_{t-1}}\right), & \text{if no dividend, bonus, or split event,} \\ \log\left(\dfrac{F_t P_t + D_t}{P_{t-1}}\right), & \text{on ex-dividend/bonus/split dates.} \end{cases} \tag{4}$$

This specification ensures that log returns consistently reflect both market price movements and the impact of corporate actions. The adjusted formulation avoids reliance on vendor-supplied "adjusted close" series and allows transparent reconstruction of return series directly from raw price and corporate action data. Pay attention that adjusted close prices are avoided here, only use unadjusted close prices and adjust accordingly when corporate event happens for the log-return matrix.

2.2.3 Log Return Matrix

Across the entire cross-section of N stocks and T trading dates, the adjusted log returns form the matrix

$$\texttt{logret} = \begin{bmatrix} r_{t_1,1} & r_{t_1,2} & \cdots & r_{t_1,N} \\ r_{t_2,1} & r_{t_2,2} & \cdots & r_{t_2,N} \\ \vdots & \vdots & \ddots & \vdots \\ r_{t_T,1} & r_{t_T,2} & \cdots & r_{t_T,N} \end{bmatrix},$$

where $r_{t,i} = \log(1 + R_{t,i})$ is the adjusted log return of stock i on date t.

3 Sample Space

After constructing the core data matrices, the next step is to define the *sample space*, which serves as the effective universe of tradable stocks. The purpose of the sample space is to filter out securities that are either inactive or insufficiently liquid, ensuring that only stocks meeting basic trading criteria are retained. Because liquidity conditions vary over time, the sample space is recomputed each day. Two conditions are imposed:

Table 4: Illustrative slice of the `logret` matrix

Date	sh600000	sh600004	$\cdots$	sz301662
2020-01-03	0.010371	−0.008023	$\cdots$	NaN
2020-01-06	−0.011173	0.017412	$\cdots$	NaN
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
2025-06-10	0.004454	−0.004264	$\cdots$	0.024690
2025-06-12	0.008867	−0.005322	$\cdots$	0.182343
2025-06-13	−0.009677	0.004278	$\cdots$	0.055802

Notes: Missing values (`NaN`) indicate that the stock was not listed or did not trade on that date, rather than representing zero returns. When constructing factor exposures or estimating covariance matrices, such entries must be handled carefully.

$$\frac{1}{10}\sum_{k=0}^{9}\mathrm{AMO}_{t-k,i} \;>\; 10^8 \text{ RMB}, \tag{5}$$

which requires that the 10-day average trading amount (AMO) of stock i exceeds 100 million RMB at date t; and

$$\mathrm{ListingPeriod}_{t,i} \;\geq\; 60, \tag{6}$$

which requires that stock i has been listed for at least 60 trading days.

These criteria ensure that the final universe consists of stocks with sustained trading activity and sufficient history for signal evaluation. By applying this daily filter, the resulting sample space reflects the dynamic composition of liquid and investable securities, minimizing distortions from newly listed or illiquid stocks. Observed that the number of stocks that satisfy the daily sample space is around 2000 for the Chinese A-share market, as illustrated in Figure 1.

4 The Alphas

4.1 Introduction

When attempting to reproduce the family of 101 alphas, one must first recognize that these signals are not mere theoretical constructs but real-life trading rules designed to capture subtle return patterns from daily price, volume, and corporate action data. Each alpha is essentially a functional transformation of basic market observables—open, high, low, close, volume, vwap, and sometimes fundamentals such as market capitalization or industry classification—expressed in formulaic code that is immediately translatable into implementable strategies. The critical task in replication is to ensure consistency in data alignment, treat-

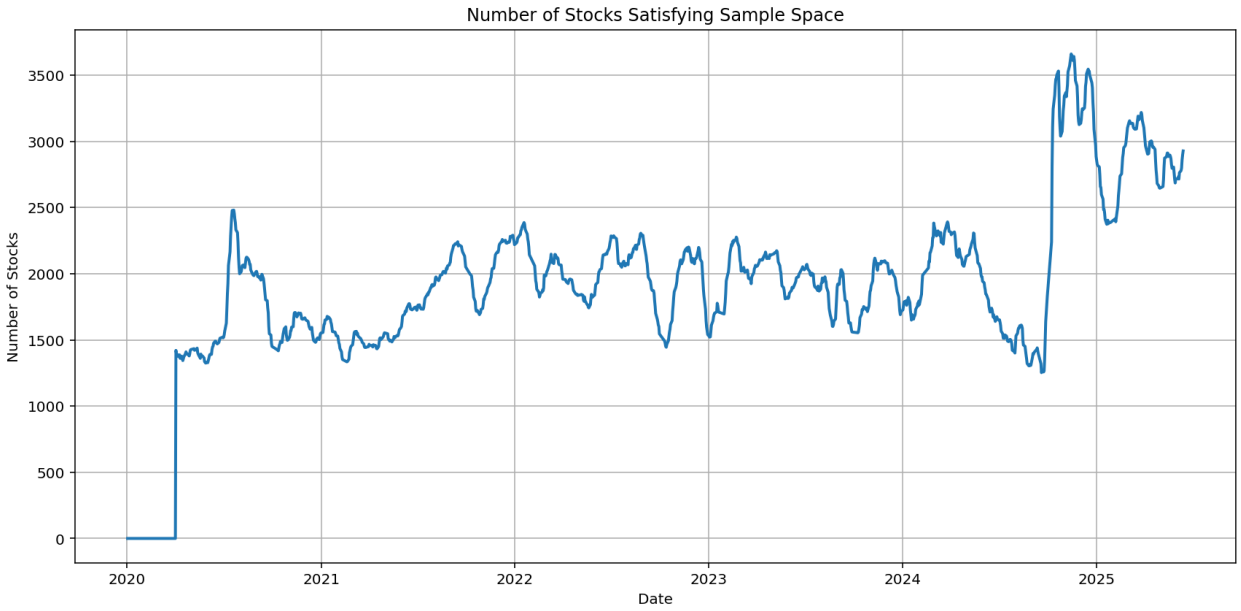

Figure 1: Illustration of the dynamic sample space.

ment of missing values, and application of corporate action adjustments, since even small deviations can materially distort the empirical performance of a signal. Therefore, before one evaluates whether an alpha is a viable trading predictor, it is necessary to construct the entire panel of alpha signals as structured matrices, where rows represent time and columns represent securities. This matrix-based format enables systematic backtesting, correlation analysis, and eventual portfolio construction.

In the original specification of the 101 formulaic alphas, the signals are categorized into two types: delay-0 and delay-1. Delay-0 alphas incorporate information from the same trading day in which the signal is computed, making them unsuitable for realistic backtesting because they rely on contemporaneous data that would not have been observable at the decision point. Examples of such signals include Alpha #42, Alpha #48, Alpha #53, and Alpha #54, all of which explicitly use same-day information in their construction and therefore introduce unavoidable look-ahead bias. Delay-1 alphas, by contrast, are constructed such that only information available up to the previous trading day is used, ensuring implementability under a daily rebalancing framework. For the purposes of this study, the focus is restricted to delay-1 alphas, as they align with the practical constraints of daily frequency trading and eliminate the biases inherent in delay-0 formulations. This restriction ensures that the empirical evaluation of alphas reflects trading conditions that could feasibly be replicated in live markets.

At the same time, researchers should be mindful of the empirical caveats that prior stud-

ies highlight. The alphas' holding periods are short, typically on the order of one to several days, and their returns are strongly correlated with volatility while showing little dependence on turnover.This implies that naive reproduction may lead to overfitting or false positives if data-cleaning practices are not rigorous or if survivorship bias is not mitigated. Furthermore, pairwise correlations among alphas, though on average relatively low, can still yield highly singular covariance structures in large sets, complicating performance attribution. As such, storing the alpha outputs in matrix form is not just a matter of convenience—it is a prerequisite for empirical robustness checks, factor-model–based risk control, and for exploring whether these formulaic alphas retain predictive power under current market conditions. Proper reproduction of these alphas is thus as much an exercise in rigorous data engineering as it is in financial modeling.

4.2 Core Supportive Functions

A central component in the reproduction of alphas is the set of standardized operators that transform raw market observables into candidate signals. These functions allow researchers to extract cross-sectional structure, capture temporal dependencies, and isolate idiosyncratic effects. Alpha signals build on these functions, and among the most frequently applied are cross-sectional ranking, lagging, rolling correlations, and industry-neutralization. Their correct implementation is essential for empirical consistency, as even minor deviations in handling these operations can materially alter the statistical properties of the signals.

For instance, the cross-sectional rank operator maps raw values into relative scores across securities:

$$\text{rank}_{\text{cs}}(X)_{t,i} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}\{X_{t,j} \leq X_{t,i}\}, \tag{7}$$

where $X_{t,i}$ denotes the observable (e.g., close price or volume) of stock i at time t, N is the number of stocks, and the indicator function counts how many securities have values less than or equal to $X_{t,i}$. This operator standardizes each stock's value into a percentile rank within the cross-section at time t.

The delay operator introduces historical information explicitly:

$$\text{delay}_d(X)_{t,i} = X_{t-d,i}, \tag{8}$$

where d is the lag length in trading days. This function retrieves the value of variable X for stock i observed d days before t, allowing the construction of momentum or reversal-type signals.

Rolling correlations capture short-term co-movements:

$$\text{corr}_d(X,Y)_{t,i} = \frac{\sum_{k=0}^{d-1}(X_{t-k,i} - \overline{X}_{t,d,i})(Y_{t-k,i} - \overline{Y}_{t,d,i})}{\sqrt{\sum_{k=0}^{d-1}(X_{t-k,i} - \overline{X}_{t,d,i})^2}\,\sqrt{\sum_{k=0}^{d-1}(Y_{t-k,i} - \overline{Y}_{t,d,i})^2}}, \tag{9}$$

where $X_{t-k,i}$ and $Y_{t-k,i}$ are the lagged observations of two time series (e.g., returns and volume), and $\overline{X}_{t,d,i}$ and $\overline{Y}_{t,d,i}$ are their respective d-day moving averages. This statistic measures whether two series have moved together or diverged over the past d days.

Finally, the industry-neutralization operator is designed to remove common group-level effects, such as those arising from sectors, industries, or subindustries. Formally, it is expressed as

$$\text{indneutralize}_g(X)_{t,i} \;=\; X_{t,i} - \frac{1}{|G_{t,g_{t,i}}|}\sum_{j \in G_{t,g_{t,i}}} X_{t,j}. \tag{10}$$

Here, $X_{t,i}$ denotes the raw value of the variable of interest (for example, an alpha signal or return) for stock i on day t. The term $g_{t,i}$ represents the categorical group label assigned to stock i at time t, such as its sector, industry, or subindustry classification. The set $G_{t,g_{t,i}}$ consists of all stocks j that share the same group label as stock i on date t, and $|G_{t,g_{t,i}}|$ denotes the number of constituents in that group. The operator subtracts the cross-sectional average of the variable within each peer group from the raw value of the stock, thereby producing a demeaned series. In this way, the transformed value represents the idiosyncratic component of the signal, free from the influence of common sectoral or industry-level effects.

Together, these operators form the building blocks for constructing alphas. They ensure comparability across securities, embed relevant temporal structures, and mitigate the influence of common factors, thereby focusing attention on whether the derived signals contain genuine predictive power.

4.3 Alpha Matrices

Once the data foundation and sample space have been established, the next step is to construct the alphas in matrix forms. Each alpha corresponds to a quantitative trading signal, derived from transformations of market observables such as price, volume, or returns, often combined with statistical operators (e.g., rank, correlation, decay). For a universe of N stocks over T trading days, every alpha produces a $T \times N$ matrix, where entry $A_{t,i}^{(\ell)}$ represents the value of alpha ℓ for stock i on date t. These matrices form the backbone of the empirical analysis, allowing systematic evaluation, cross-sectional comparison, and eventual portfolio construction. Below is a visualization of a general alpha matrix:

Alpha Matrix General Form

$$\texttt{alpha}_\ell = \begin{bmatrix} A^{(\ell)}_{t_1,1} & A^{(\ell)}_{t_1,2} & \cdots & A^{(\ell)}_{t_1,N} \\ A^{(\ell)}_{t_2,1} & A^{(\ell)}_{t_2,2} & \cdots & A^{(\ell)}_{t_2,N} \\ \vdots & \vdots & \ddots & \vdots \\ A^{(\ell)}_{t_T,1} & A^{(\ell)}_{t_T,2} & \cdots & A^{(\ell)}_{t_T,N} \end{bmatrix},$$

where $A^{(\ell)}_{t,i}$ is the value of alpha ℓ for stock i at time t.

Example: Alpha #6

As an illustrative case, consider Alpha #6, defined by

$$A^{(6)}_{t,i} = -\operatorname{corr}_{10}\big(\text{open}_{\cdot,i},\, \text{volume}_{\cdot,i}\big)_t\,, \tag{11}$$

where $\text{open}_{\cdot,i}$ and $\text{volume}_{\cdot,i}$ denote the 10-day histories of the opening price and trading volume of stock i, respectively. The operator $\operatorname{corr}_{10}(\cdot,\cdot)$ computes the Pearson correlation coefficient over the most recent 10 trading days, and the negative sign implies that stronger positive correlation reduces the alpha's signal value. In economic terms, this alpha captures the tendency for stocks with a high contemporaneous co-movement between opening prices and trading volume to exhibit weaker expected returns.

Table 5: Illustrative slice of the `alpha_6` matrix

Date	sh600000	sh600004	$\cdots$	sz301665
2020-01-02	NaN	NaN	$\cdots$	NaN
2020-01-03	NaN	NaN	$\cdots$	NaN
2020-01-06	NaN	NaN	$\cdots$	NaN
2020-01-07	NaN	NaN	$\cdots$	NaN
2020-01-08	NaN	NaN	$\cdots$	NaN
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
2022-12-26	−0.044328	0.273676	$\cdots$	NaN
2022-12-27	0.041426	0.060764	$\cdots$	NaN
2022-12-28	0.165952	−0.304564	$\cdots$	NaN
2022-12-29	0.020330	−0.252056	$\cdots$	NaN
2022-12-30	0.001401	−0.377236	$\cdots$	NaN

It is important to note that the presence of `NaN` values in the alpha matrices has distinct interpretations. The missing entries at the beginning of the sample period occur because Alpha #6 requires a 10-day lookback window, and therefore cannot be computed until sufficient historical data becomes available. Similarly, `NaN` values on the right-hand side

of the matrix typically correspond to stocks that were not yet listed at the time or were temporarily suspended from trading, making the calculation of the correlation infeasible. In both cases, the `NaN` values represent unavailable observations rather than zero signals. For empirical applications, these entries should be handled with explicit masking or appropriate missing-data procedures, rather than imputed, to avoid distorting the underlying statistical properties of the signals.

At the end of the calculation, the complete set of nighty seven delay-1 alphas is assembled and stored as matrix variables, each of dimension $T \times N$ with trading days along the rows and securities along the columns. These alpha matrices serve as the foundational data objects that bridge raw market observables with systematic trading strategies.

5 Finding the Right Signal

5.1 IC and Rank IC

A standard approach to evaluating the predictive quality of alpha signals is through the Information Coefficient (IC) and its rank-based counterpart, the Rank Information Coefficient (Rank IC). These measures assess, in a purely cross-sectional sense, how well an alpha's forecasted scores align with subsequent realized returns. In quantitative finance research, they are widely adopted as model-free metrics of signal efficacy, with positive values indicating predictive power and values near zero suggesting little or no information content.

Formally, let $a_{t-1,i}$ denote the alpha score for stock i formed at the close of day $t-1$, and let $r_{t,i}$ be the realized log return of stock i on day t, taken from the `logret` matrix. Recall that on each evaluation date t, the sample space is defined as

$$\mathcal{U}_t = \{\, i : \texttt{sample}_{t-1,i} = 1 \,\},$$

which restricts attention to stocks that satisfy the liquidity and listing requirements as of $t-1$.

The cross-sectional Information Coefficient is then

$$\mathrm{IC}_t \;=\; \mathrm{corr}_{i \in \mathcal{U}_t}\big(a_{t-1,i},\, r_{t,i}\big), \tag{12}$$

where corr denotes the Pearson correlation computed across the cross-section on date t.

The Rank Information Coefficient replaces raw values with their cross-sectional ranks,

yielding the Spearman correlation:

$$\mathrm{RankIC}_t \;=\; \mathrm{corr}_{i\in\mathcal{U}_t}\big(\,\mathrm{rank}(a_{t-1,i}),\,\mathrm{rank}(r_{t,i})\big). \tag{13}$$

Across a sample of T trading days, the time series $\{\mathrm{IC}_t\}_{t=1}^{T}$ or $\{\mathrm{RankIC}_t\}_{t=1}^{T}$ is summarized by its mean, variability, and information ratio (IR). Specifically,

$$\overline{\mathrm{IC}} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{IC}_t, \tag{14}$$

Analogous definitions apply for Rank IC.

In practice, the absolute values of IC and Rank IC are typically small, as even strong predictive signals explain only a modest fraction of cross-sectional return variation. Daily-frequency alphas with an average IC in the range of 0.01 to 0.02 (or equivalently a Rank IC of similar magnitude) are generally considered meaningful in large universes, since such correlations can compound into economically significant performance once leveraged through portfolio construction. Ultimately, the interpretation depends on the trading frequency, transaction costs, and breadth of the investment universe, but even seemingly small positive IC or Rank IC values can translate into substantial long-run profitability.

It is essential that IC and Rank IC are computed only within the daily sample space $\mathcal{U}_t$, rather than across all listed stocks. This restriction ensures that the evaluation reflects a tradable and liquid investment universe. Including illiquid, suspended, or recently listed stocks would contaminate the correlation calculations with returns that cannot realistically be captured, thereby overstating or understating the true predictive power of the signals. By limiting the analysis to stocks that meet the liquidity and age requirements, the IC metrics provide an unbiased assessment of alpha performance under feasible trading conditions.

5.2 IC and RIC Matrices

Recall that given a cross-section of N stocks and T trading dates, each alpha ℓ and the realized log returns are stored as

$$\texttt{alpha}_\ell = \begin{bmatrix} A^{(\ell)}_{t_1,1} & A^{(\ell)}_{t_1,2} & \cdots & A^{(\ell)}_{t_1,N} \\ A^{(\ell)}_{t_2,1} & A^{(\ell)}_{t_2,2} & \cdots & A^{(\ell)}_{t_2,N} \\ \vdots & \vdots & \ddots & \vdots \\ A^{(\ell)}_{t_T,1} & A^{(\ell)}_{t_T,2} & \cdots & A^{(\ell)}_{t_T,N} \end{bmatrix}, \qquad \texttt{logret} = \begin{bmatrix} r_{t_1,1} & r_{t_1,2} & \cdots & r_{t_1,N} \\ r_{t_2,1} & r_{t_2,2} & \cdots & r_{t_2,N} \\ \vdots & \vdots & \ddots & \vdots \\ r_{t_T,1} & r_{t_T,2} & \cdots & r_{t_T,N} \end{bmatrix}.$$

For each date t, the IC is computed cross-sectionally by pairing the *row at* $t-1$ from the alpha matrix with the *row at* t from the log return matrix. Writing the row vectors

$$a_{t-1}^{(\ell)} = \big(A_{t-1,1}^{(\ell)}, \ldots, A_{t-1,N}^{(\ell)}\big), \qquad r_t = \big(r_{t,1}, \ldots, r_{t,N}\big),$$

and letting $\mathcal{U}_t$ denote the daily sample space (eligible, liquid names), the IC and Rank IC for alpha ℓ on date t are

$$\mathrm{IC}_t^{(\ell)} = \mathrm{corr}_{i\in\mathcal{U}_t}\big(a_{t-1,i}^{(\ell)},\, r_{t,i}\big), \qquad \mathrm{RIC}_t^{(\ell)} = \mathrm{corr}_{i\in\mathcal{U}_t}\big(\,\mathrm{rank}\, a_{t-1,i}^{(\ell)},\, \mathrm{rank}\, r_{t,i}\big).$$

Stacking these over time yields the IC vector and the IC sample mean for alpha ℓ,

$$\mathrm{IC}^{(\ell)} = \big(\mathrm{IC}_{t_1}^{(\ell)},\, \mathrm{IC}_{t_2}^{(\ell)},\, \ldots,\, \mathrm{IC}_{t_T}^{(\ell)}\big)^\top, \quad \overline{\mathrm{IC}}^{(\ell)} = \frac{1}{T}\sum_t \mathrm{IC}_t^{(\ell)},$$

with analogous definitions for $\mathrm{RIC}^{(\ell)}$ and $\overline{\mathrm{RIC}}^{(\ell)}$.

In words: treat each alpha matrix *horizontally*. On every date t, correlate the row of alpha scores formed at $t-1$ with the row of realized log returns at t, using only stocks in the day-t sample space $\mathcal{U}_t$. Repeating this over all dates produces an IC (or RIC) time series for that alpha, whose sample mean summarizes its average predictive power.

5.3 Hypothesis Testing

Core idea

To assess whether an alpha exhibits genuine predictive power, the below method tests whether its cross–sectional association with future returns is statistically different from an economically irrelevant benchmark. For a given alpha ℓ, the object of inference is the population mean

$$\mu = \mathbb{E}\Big[\mathrm{IC}_t^{(\ell)}\Big],$$

with $\overline{\mathrm{IC}}^{(\ell)}$ serving as its sample estimate. An entirely analogous interpretation applies to the relative RIC calculations.

HAC rationale

Daily IC (and RIC) series are not well described by an i.i.d. assumption: they typically display *heteroskedasticity* (time–varying volatility) and *serial correlation* (dependence across days, e.g., due to regime persistence or overlapping influences). If one applies standard

variance formulas under i.i.d. assumptions, the estimated standard errors will be downward–biased and hypothesis tests will appear spuriously significant.

To address this, a heteroskedasticity– and autocorrelation–consistent (HAC) estimator is employed to capture the sampling variance of the mean IC. The construction relies on the long–run variance (LRV) of the process. Let the sample length be T and denote the sample mean as

$$\overline{\mathrm{IC}}^{(\ell)} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{IC}_t^{(\ell)}.$$

Autocovariances. Consider the population mean $\mu^{(\ell)} = \mathbb{E}[\mathrm{IC}_t^{(\ell)}]$ and define the demeaned process

$$\widetilde{\mathrm{IC}}_t^{(\ell)} = \mathrm{IC}_t^{(\ell)} - \mu^{(\ell)}.$$

For lag $l \geq 0$, the corresponding population autocovariance is

$$\gamma_l^{(\ell)} = \mathrm{Cov}\big(\mathrm{IC}_t^{(\ell)},\, \mathrm{IC}_{t-l}^{(\ell)}\big) = \mathbb{E}\Big[\widetilde{\mathrm{IC}}_t^{(\ell)}\, \widetilde{\mathrm{IC}}_{t-l}^{(\ell)}\Big].$$

The lag–0 autocovariance $\gamma_0^{(\ell)}$ is the variance of the process. In finite samples, autocovariances are estimated as

$$\widehat{\gamma}_l^{(\ell)} = \frac{1}{T}\sum_{t=l+1}^{T} \big(\mathrm{IC}_t^{(\ell)} - \overline{\mathrm{IC}}^{(\ell)}\big)\big(\mathrm{IC}_{t-l}^{(\ell)} - \overline{\mathrm{IC}}^{(\ell)}\big), \qquad l = 0, 1, \ldots, L.$$

In implementation, missing values are handled by computing means and covariances only over valid observations, with effective sample sizes $n^{(\ell)}$ and $n_l^{(\ell)}$ tracked explicitly.

Long–run variance. The asymptotic variance of the sample mean can then be expressed as the long–run variance divided by T:

$$\mathrm{Var}\Big(\overline{\mathrm{IC}}^{(\ell)}\Big) \;\approx\; \frac{\Omega^{(\ell)}}{T}, \qquad \Omega^{(\ell)} = \gamma_0^{(\ell)} + 2\sum_{l=1}^{\infty} \gamma_l^{(\ell)}.$$

This $\Omega^{(\ell)}$ aggregates contemporaneous variance and all lagged covariance terms, thereby incorporating both heteroskedasticity and persistence.

Newey–West HAC estimator. Because only a finite sample of length T is available, the infinite sum of autocovariances must be truncated at lag L, and Bartlett kernel weights are

used to stabilize the estimate. The Newey–West estimator is defined as

$$\widehat{\Omega}_{\mathrm{NW}}^{(\ell)} = \widehat{\gamma}_0^{(\ell)} + 2\sum_{l=1}^{L} w_l\, \widehat{\gamma}_l^{(\ell)}, \qquad w_l = 1 - \frac{l}{L+1}, \quad l = 1, \ldots, L.$$

The weights w_l decay linearly from 1 at lag 0 to 0 at lag $L+1$, down–weighting more distant autocovariances and ensuring positive semidefiniteness. In practice, L is bounded above by a user–specified parameter `maxlags` and by the effective sample size of each series, $L \le \min(\texttt{maxlags}, n^{(\ell)} - 1)$. For finite samples, an adjustment factor $n^{(\ell)}/\max(n^{(\ell)} - l, 1)$ is also applied to correct for the diminishing number of valid pairs at higher lags.

HAC variance and standard error. The HAC estimator of the variance of the sample mean is therefore

$$\widehat{\mathrm{Var}}\big(\overline{\mathrm{IC}}^{(\ell)}\big) = \frac{1}{T}\,\widehat{\Omega}_{\mathrm{NW}}^{(\ell)} = \frac{1}{T}\left(\widehat{\gamma}_0^{(\ell)} + 2\sum_{l=1}^{L} w_l\, \widehat{\gamma}_l^{(\ell)}\right),$$

with corresponding HAC standard error

$$\mathrm{SE}_{\mathrm{HAC}} = \sqrt{\widehat{\mathrm{Var}}\big(\overline{\mathrm{IC}}^{(\ell)}\big)}.$$

In computation, the estimator is applied simultaneously across all signals in a vectorized fashion, treating each column of IC (or RIC) values as a separate series.

Choice of truncation lag L. The truncation parameter L governs the bias–variance trade-off. A larger L incorporates longer–horizon serial dependence (reducing bias) but adds sampling noise (inflating variance). Common practice is to set L using plug–in rules such as

$$L = \left\lfloor 4\left(\frac{T}{100}\right)^{2/9} \right\rfloor \quad \text{or} \quad L = \lfloor c\, T^{1/4} \rfloor,$$

with c a small constant. In the strategy, L is further restricted by `maxlags`, ensuring truncation never exceeds a predetermined bound, and is adjusted for sample size to maintain validity. Sensitivity checks across a range of L values are performed to confirm robustness.

This construction mirrors the Newey–West procedure and provides consistent inference on $\overline{\mathrm{IC}}^{(\ell)}$ (and analogously on $\overline{\mathrm{RIC}}^{(\ell)}$) under general time–dependence, while incorporating missing data handling, effective sample adjustments, and small–sample corrections.

Test statistic. The Newey–West construction above yields the HAC standard error $\mathrm{SE}_{\mathrm{HAC}}$ for $\overline{\mathrm{IC}}^{(\ell)}$, which enters directly into the test statistic. Given a null mean μ_0 (typically 0; an

economic threshold can also be used), the statistic is

$$t^{(\ell)} = \frac{\overline{\mathrm{IC}}^{(\ell)} - \mu_0}{\mathrm{SE}_{\mathrm{HAC}}},$$

which is evaluated against a Student–t reference with $T-1$ degrees of freedom (large–sample normality provides a similar decision rule).

Directional one–sided tests and trading interpretation. Because trading is directional, one–sided significance aligned with the observed effect is reported:

$$\begin{aligned}
&\text{(Positive–side test)} \quad H_0 : \mu \le \mu_0 \quad \text{vs.} \quad H_A : \mu > \mu_0, \qquad p_+ = 1 - F_t\big(t^{(\ell)}\big);\\
&\text{(Negative–side test)} \quad H_0 : \mu \ge \mu_0 \quad \text{vs.} \quad H_A : \mu < \mu_0, \qquad p_- = F_t\big(t^{(\ell)}\big),
\end{aligned}$$

where $F_t(\cdot)$ is the t–distribution CDF. For a long–only implementation, the sign of the estimated information determines how the score is used: if the alpha's mean IC is positive, larger scores at $t-1$ are associated with larger returns at t, so high–score names are long candidates; if the mean IC is negative, the signal is contrarian, so low–score names are the long candidates (equivalently, the score's sign may be flipped before portfolio formation). The same testing and interpretation apply to the rank–based series (RIC). In all cases IC/RIC are computed only over the daily sample space $\mathcal{U}_t$ to ensure results reflect a tradable, liquid universe.

5.4 Alpha Filtering Logic

Having all 97 delay-1 alphas as the trading signals are too much. Therefore, filtering the effective ones is an important step. Each delay-1 alpha is evaluated through both statistical and economic filters. The statistical component is a one–sided hypothesis test at the 5% level with a one–sided 95% confidence interval. The null means for both directions are set to zero. To avoid retaining signals that are statistically significant but economically negligible, the strategy additionally impose an absolute threshold of 0.001 for both IC and RIC.

An alpha is retained only if its IC and RIC satisfy the following conditions simultaneously:

- **Directional consistency:** IC and RIC must have the same sign, ensuring that both linear (Pearson) and rank–based (Spearman) measures point in the same direction.
- **Statistical significance:** Both IC and RIC reject the null at the 5% level, with one–sided confidence intervals lying strictly above zero (for positive cases) or strictly below zero (for negative cases).

- **Economic materiality:** The average IC and RIC must exceed the threshold of 0.001 in absolute value.

This combined filter guarantees that retained alphas are (i) directionally coherent across correlation measures, (ii) statistically credible under one–sided HAC inference, and (iii) economically meaningful in magnitude.

5.5 Testing Results

Applying the filtering procedure to the delay–1 alpha set yields a final collection of 49 alpha signals out of the 97 originally tested. These retained alphas consistently satisfy the dual requirements of statistical significance and economic materiality, while also showing directional agreement between IC and RIC.

Within this selected group, certain alphas stand out. Alpha 40 exhibits the strongest positive average IC, with $\overline{\text{IC}} = 0.019601$ and a corresponding $\overline{\text{RIC}} = 0.035144$. Alpha 42 produces the highest positive average RIC, with $\overline{\text{RIC}} = 0.036430$ and a moderate $\overline{\text{IC}} = 0.003644$. On the negative side, Alpha 101 is the most contrarian predictor, showing the lowest mean values across both metrics with $\overline{\text{IC}} = -0.016897$ and $\overline{\text{RIC}} = -0.021540$.

Overall, these results confirm that roughly half of the formulaic delay–1 alphas in the sample exhibit statistically credible and economically meaningful predictive power under daily rebalancing. The distribution of outcomes—ranging from strongly positive to strongly contrarian—highlights the diversity of underlying trading rules and provides a robust basis for subsequent portfolio construction and performance attribution.

5.6 Fama–MacBeth Cross-Sectional Pricing Test

The IC and Rank IC analysis demonstrated that certain delay–1 alphas possess statistically credible and economically meaningful predictive power. While such evidence establishes cross-sectional association between signals and next-day returns, it does not guarantee that exposures to these signals are systematically compensated in expectation. To address this, the Fama–MacBeth (FM) two-step procedure is employed. This method estimates the average price of risk for each retained alpha and tests whether it is significantly different from zero, thereby linking predictive association to priced premia.

Data aggregation by date. Fix a trading day t. Let $N_t = |\mathcal{U}_t|$ denote the number of eligible stocks (i.e., those with valid $r_{t,i}$ and $a^{(j)}_{t-1,i}$). The lagged alpha scores are stacked into

the $N_t \times K$ regressor matrix

$$X_t \;=\; \begin{bmatrix} a^{(1)}_{t-1,1} & a^{(2)}_{t-1,1} & \cdots & a^{(K)}_{t-1,1} \\ a^{(1)}_{t-1,2} & a^{(2)}_{t-1,2} & \cdots & a^{(K)}_{t-1,2} \\ \vdots & \vdots & \ddots & \vdots \\ a^{(1)}_{t-1,N_t} & a^{(2)}_{t-1,N_t} & \cdots & a^{(K)}_{t-1,N_t} \end{bmatrix}, \qquad y_t \;=\; \begin{bmatrix} r_{t,1} \\ r_{t,2} \\ \vdots \\ r_{t,N_t} \end{bmatrix}.$$

If an intercept is included, a leading column of ones is appended to X_t. The daily cross-sectional regression is then expressed as

$$y_t \;=\; X_t\beta_t \;+\; \varepsilon_t, \tag{15}$$

where $\beta_t \in \mathbb{R}^{K\times 1}$ is the vector of date-t risk premia, and $\varepsilon_t \in \mathbb{R}^{N_t\times 1}$ are residuals. Crucially, β_t is common to all stocks on date t; the cross-section provides the N_t observations required to estimate this coefficient vector.

Daily estimation. The least-squares solution to (15) is

$$\hat\beta_t \;=\; (X_t^\top X_t)^{-1}X_t^\top y_t, \qquad \text{or} \quad \hat\beta_t \;=\; (X_t^\top X_t{+}\lambda I)^{-1}X_t^\top y_t \;\text{ (ridge, if regularization is applied)}, \tag{16}$$

with $\lambda \geq 0$ a penalty parameter introduced only to mitigate numerical instability from multicollinearity. Each $\hat\beta_t$ is a $K \times 1$ vector summarizing the daily pricing of all retained alphas.

Stacking across time. Repeating (16) for $t = 1, \ldots, T$ produces a panel of estimated premia:

$$\hat B \;=\; \begin{bmatrix} \hat\beta_{1,1} & \hat\beta_{2,1} & \cdots & \hat\beta_{K,1} \\ \hat\beta_{1,2} & \hat\beta_{2,2} & \cdots & \hat\beta_{K,2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat\beta_{1,T} & \hat\beta_{2,T} & \cdots & \hat\beta_{K,T} \end{bmatrix} \;\in\; \mathbb{R}^{T\times K}.$$

Each row corresponds to one trading day's coefficient vector, while each column is the time series of estimated premia for a particular alpha. The FM estimate of the unconditional price of risk for factor j is the column average:

$$\bar\beta_j \;=\; \frac{1}{T_j}\sum_{t=1}^{T_j}\hat\beta_{j,t}, \tag{17}$$

with T_j denoting the number of valid daily estimates for factor j.

Sampling variability and robust standard errors. The statistical question is whether the average premium $\bar{\beta}_j$ differs significantly from zero. Hypothesis testing is essential here because even if an alpha appears predictive on average, its estimated premia $\{\hat{\beta}_{j,t}\}$ may fluctuate randomly across time. Only if the mean effect is statistically distinguishable from noise can one conclude that the factor is genuinely priced. The simplest variance estimator assumes independence across t:

$$\mathrm{se}_j^{\mathrm{iid}} \;=\; \frac{s_j}{\sqrt{T_j}}, \qquad s_j^2 \;=\; \frac{1}{T_j - 1} \sum_{t=1}^{T_j} \big(\hat{\beta}_{j,t} - \bar{\beta}_j\big)^2.$$

However, in financial data the coefficient series often exhibits autocorrelation, reflecting persistence in regimes or clustering of shocks. To account for this, a heteroskedasticity- and autocorrelation-consistent (HAC) estimator is used:

$$\mathrm{se}_j^{\mathrm{hac}} \;=\; \sqrt{\frac{1}{T_j}\left(\gamma_0 + 2\sum_{\ell=1}^{L} w_\ell\, \gamma_\ell\right)}, \qquad w_\ell \;=\; 1 - \frac{\ell}{L+1},$$

where γ_ℓ is the lag-ℓ autocovariance of $\{\hat{\beta}_{j,t}\}$. The truncation lag L follows the Newey–West convention, with declining weights w_ℓ, as mentioned in Section 5.3. In practice, autocorrelation is diagnosed using the autocorrelation function (ACF) and partial autocorrelation function (PACF): the ACF measures correlation between $\hat{\beta}_{j,t}$ and $\hat{\beta}_{j,t-\ell}$ at lag ℓ, while the PACF isolates the direct effect at lag ℓ after controlling for shorter lags. If either shows significant dependence, HAC errors are applied; otherwise the i.i.d. form suffices.

Hypotheses and decision rule. Formally, the null and alternative hypotheses are

$$H_0 :\; \mathbb{E}[\hat{\beta}_{j,t}] = 0 \qquad \text{versus} \qquad H_A :\; \mathbb{E}[\hat{\beta}_{j,t}] \neq 0.$$

The test statistic is

$$t_j \;=\; \frac{\bar{\beta}_j}{\mathrm{se}_j}, \qquad p\text{-value} \;=\; 2\big(1 - F_t(|t_j|;\, T_j - 1)\big),$$

where se_j is chosen according to the ACF/PACF rule and $F_t(\cdot;\, T_j - 1)$ is the Student–t distribution function. A two-sided test is adopted, since both positively priced and contrarian (negatively priced) factors are economically meaningful. The sign of $\bar{\beta}_j$ remains critical for

portfolio construction, but statistical inference is conducted without imposing a directional prior.

Interpretation and linkage to IC/RIC. The IC and Rank IC metrics demonstrated predictive ordering power at the daily horizon, with one-sided tests used to align with trading direction and economic thresholds. The FM regression complements this by establishing whether the corresponding exposures are systematically priced in the cross-section. A significant positive $\bar{\beta}_j$ implies that higher alpha scores at $t-1$ translate into systematically higher returns at t, while a significant negative $\bar{\beta}_j$ implies that the factor is contrarian, penalizing exposure. Taken together, IC/RIC and FM provide a coherent framework: the former identifies signals with day-to-day predictive validity, while the latter tests whether such signals correspond to persistent cross-sectional premia.

5.7 Final Selected Alphas and Diagnostic Report

The IC and Rank IC analysis previously provided an initial filter of delay–1 alphas, ensuring that retained signals were directionally coherent, statistically credible under HAC inference, and economically meaningful above a minimum threshold of 0.001. Having established this foundation, a second-stage screening is implemented based on the outcomes of the Fama–MacBeth regressions and subsequent correlation analysis. The purpose is to refine the signal set further, ensuring that retained alphas are not only cross-sectionally predictive but also systematically priced and sufficiently distinct from one another.

Integration with Fama–MacBeth results. For each alpha j, the Fama–MacBeth procedure yields an average price of risk $\bar{\beta}_j$, its robust standard error se_j, and a two-sided significance test of $H_0 : \mathbb{E}[\hat{\beta}_{j,t}] = 0$. To be retained, an alpha must pass the following condition:

$$p_j < \alpha_{\mathrm{FM}}, \qquad \alpha_{\mathrm{FM}} = 0.10,$$

which corresponds to rejecting the null at the 10% level (i.e., 90% confidence). This requirement ensures that the factor is systematically priced in the cross-section rather than displaying predictive ability that is indistinguishable from noise. A two-sided test is deliberately chosen, because both positively priced and contrarian factors are economically relevant: the former align naturally with long exposure, while the latter can be incorporated via sign reversal at the portfolio stage.

Revised economic materiality. In addition to statistical significance, retained factors must satisfy

$$|\overline{\mathrm{IC}}^{(j)}| \;\geq\; 0.005,$$

a more stringent economic filter compared to the earlier threshold of 0.001. Raising the bar from 0.001 to 0.005 ensures that the final factor set reflects not only statistically credible but also materially impactful predictive relationships. This adjustment is necessary because small but statistically significant IC values, while non-negligible in large universes, may translate into economically weak signals once costs and capacity constraints are accounted for. The stricter cutoff thus prioritizes alphas that combine statistical robustness with practical relevance.

Correlation diagnostics. After statistical and economic filters, the retained set is subjected to pairwise correlation analysis. Specifically, the cross-sectional time series of each alpha is compared using the Pearson correlation coefficient. The resulting correlation matrix is examined to detect redundancies and clusters of highly collinear signals. Alphas exhibiting extremely high pairwise correlations (e.g., $|\rho| > 0.8$) risk introducing instability in subsequent portfolio construction, as their contributions to forecasted returns are not independent. Addressing this correlation structure—either by clustering, pruning, or orthogonalizing—is essential to obtain a parsimonious, diversified, and robust factor set.

Rationale. This reporting step consolidates three necessary conditions: (i) systematic pricing under Fama–MacBeth significance, (ii) elevated economic materiality via a stricter IC threshold, and (iii) non-redundancy via correlation diagnostics. Together these filters ensure that the final alphas are statistically sound, economically meaningful, and informationally distinct. By applying this hierarchy of screens, the strategy transitions from an initial broad universe of delay–1 signals to a focused set of priced factors, suitable for subsequent risk-controlled portfolio construction and attribution.

The final screening yields a parsimonious set of seven alphas: `alpha_40`, `alpha_38`, `alpha_19`, `alpha_16`, `alpha_69`, `alpha_94`, and `alpha_90`. Each of these factors passes the joint criteria of Fama–MacBeth significance at the 10% level (two-sided), an elevated economic materiality threshold of $|\overline{\mathrm{IC}}| \geq 0.005$, and non-redundancy checks through correlation analysis.

Autocorrelation diagnostics. For each factor, the ACF/PACF test determined whether HAC-robust or i.i.d. standard errors were appropriate. The results show that all selected

alphas exhibited behavior consistent with the i.i.d. assumption, indicating that serial correlation in their estimated premia was not statistically significant. This reduces the need for HAC adjustments and confirms the stability of their time-series coefficient processes.

Fama–MacBeth estimates with IC. The joint regression and information coefficient results for the seven retained alphas are presented in Table 5.7.

Alpha	Method	$\hat{\beta}$	SE	t-stat	T_k	df	p-value	$\overline{\text{IC}}$
alpha_40	IID	0.001273	0.000394	3.230	477	476	0.0013	0.0196
alpha_38	IID	0.001376	0.000756	1.819	477	476	0.0695	0.0184
alpha_19	HAC	0.000211	0.000095	2.209	477	476	0.0277	0.0179
alpha_16	IID	-0.000797	0.000463	-1.721	477	476	0.0858	0.0119
alpha_69	IID	0.000496	0.000286	1.733	477	476	0.0838	0.0094
alpha_94	IID	0.000725	0.000392	1.849	477	476	0.0651	0.0112
alpha_90	IID	-0.000799	0.000371	-2.155	477	476	0.0317	0.0073

These results show that all seven alphas are statistically significant at the 10% two-tailed level, with magnitudes of $\hat{\beta}$ reflecting their average risk premia and $\overline{\text{IC}}$ confirming predictive strength in cross-sectional sorting. Importantly, `alpha_16` and `alpha_90` exhibit negative estimated premia despite positive IC values. This apparent inconsistency highlights the distinction between correlation-based predictability and priced exposures: while these signals correctly rank returns in the cross-section, the market compensates them with negative risk premia, classifying them as contrarian factors. In practice, such factors are not discarded; their signs can be inverted when constructing portfolios, enabling profitable use of their predictive structure. The combination of positive IC and significant FM coefficients—whether positive or negative—provides strong evidence that these alphas contain economically meaningful information.

Correlation analysis. The correlation heatmap highlights the degree of overlap across the final seven alphas. While certain pairs, such as `alpha_38` and `alpha_90`, exhibit moderate correlation ($\rho \approx 0.59$), most pairwise relations remain below 0.50, indicating that the set is not dominated by redundant signals. This ensures that each retained factor contributes distinct information to the composite signal space. The presence of moderate correlations is expected, as signals often capture overlapping market structures, but the absence of extremely high values ($|\rho| > 0.8$) confirms that the selection retains informational diversity. This diversity is crucial for robust portfolio construction, as it mitigates the risk of overexposure to narrow market phenomena.

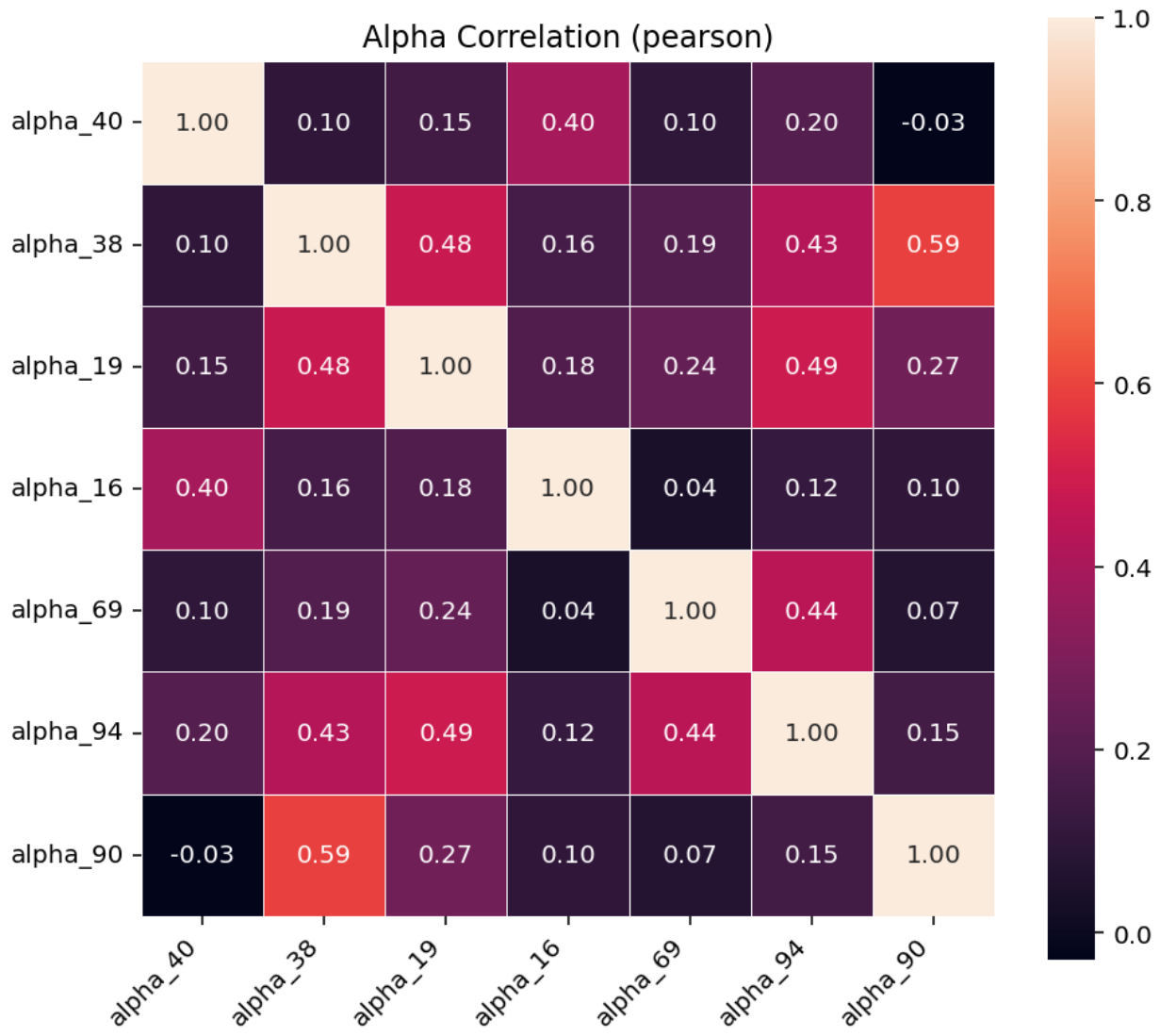

Figure 2: The Pearson Correlation Heatmap of the Final Alphas

Summary. The joint report confirms that the final seven alphas are statistically significant under Fama–MacBeth testing, economically material with respect to IC, and sufficiently independent according to correlation diagnostics. They therefore form a strong and diversified basis for subsequent portfolio construction and risk attribution analysis.