

ORIGINAL RESEARCH

Predicting CSI 800 Sector Index Returns in China A Multifactor Statistical and Regularized Learning Approach

Caiyu Zhao

Shanghai Starriver Bilingual School

Faculty Advisor: Zhiyun Zheng, Economics Teacher, Shanghai Starriver Bilingual School

Corresponding author email: Sindy_2008@outlook.com

Accepted: October 11, 2026

Abstract

Industry allocation is an important component of asset allocation and a major source of excess returns in active investing. This study investigates the predictability of sector index returns in China using a multifactor framework. Using monthly data from January 2019 to June 2026 for eleven CSI 800 first-level sector indices, we construct 11 factors including indicator (CPI, PPI, PMI, M2), interest rates (Shibor), technical indicators (momentum and volatility), and a market factor. We model each industry separately and estimate by using two linear models: ordinary least squares (OLS) and Lasso regularized regression, with the period from January 2025 to June 2026 as an out-of-sample test window. A one-shot training scheme (fixed sample of January 2019 to December 2024) yields negative out-of-sample R^2 for both models (-0.72 for OLS and -0.12 for Lasso), which was normal in monthly financial return prediction. To improve the forecasts, we adopt rolling window re-estimation: at each month's end, the models are refitted on the most recent 72, 60, and 48 months of data. Rolling-window estimation is associated with improved numerical accuracy and strategy returns: the 72-month rolling Lasso raises the average out-of-sample R^2 to -0.07 and delivers a Top-3 industry rotation strategy with a cumulative return of 65.45%, an excess return of 48.03 percentage points over the equal-weight benchmark; the 72-month rolling OLS earns 37.46% (excess +20.04%). Since both models produce negative R^2 for out-of-sample prediction, we further investigate their ranking power using Spearman rank information coefficient (Rank IC) calculated on a monthly cross-section basis: using the 72-month rolling window approach, the mean Rank IC equals $+0.215$ ($t = 2.96$) for Lasso, compared to just $+0.057$ ($t = 0.60$) for OLS. Thus, the ranking power of Lasso strategy provides support to the presence of excess returns, while the ranking power of OLS strategy is not statistically different from zero. These results suggest that rolling-window estimation can help models adapt to changing market conditions. Compared with OLS, Lasso shows stronger out-of-sample ranking performance, which is consistent with the excess returns generated by the corresponding industry rotation strategy.

Keywords: multifactor model; Lasso regularization; rolling window; CSI 800 index; industry rotation; return predictability

JEL: G11; G17; C53

1. Introduction

1.1 Research Background

Industry allocation is a critical stage of asset allocation and an important source of excess returns in active management. In capital markets, industries respond differently to macroeconomic cycles, monetary policy, and technological change, resulting in significant divergence in sector performance. For example, from 2025 to the first half of 2026, the best-performing CSI 800 sector index, telecommunications, gained 162%, while the worst-performing sector, real estate, fell by 34%. The pronounced differences across industries shows the importance of sector selection. Therefore, identifying the key factors of industry returns and building predictive multifactor models is a core method for both financial research and practice [1].

1.2 Research Objectives

The specific objectives of this study are as follows:

- (1) to systematically identify the factors that drive industry index returns and construct a multifactor indicator system;
- (2) to estimate industry-specific linear regressions and identify significant factors for each industry;
- (3) to apply Lasso regularization to mitigate overfitting and achieve automatic variable selection;
- (4) to adopt rolling-window re-estimation to improve out-of-sample prediction when one-shot training performs poorly;
- (5) to construct a Top-3 industry rotation strategy based on out-of-sample predictions and compare it with an equal-weight benchmark and the CSI 800 index.

1.3 Significance of the Study

This study contributes in three ways. First, it provides sector-level evidence from China on the debate over return predictability, systematically comparing ordinary least squares with Lasso regularization. Second, it documents the heterogeneous responses of industries to macroeconomic, interest-rate, and technical factors, deepening our understanding of industry heterogeneity. Third, it shows that rolling-window re-estimation—a simple yet effective adaptation to structural change—is associated with improved out-of-sample prediction and constructs an implementable industry rotation strategy, offering practical implications for industry allocation.

2. Research Design and Data

2.1 Data Sources

The data used in this study are mainly obtained from the Tushare Pro financial data platform:

- (1) Daily index quotations for the 11 CSI 800 first-level sector indices;
- (2) Macroeconomic data: CPI (year-on-year), PPI (year-on-year), PMI, and M2 (year-on-year);
- (3) Interest-rate data: Shanghai Interbank Offered Rate (Shibor) at the 1-year and 3-month tenors;
- (4) Market benchmark: daily closing prices of the CSI 800 Index (000906.SH).

2.2 Sample Selection

The sample consists of 11 CSI 800 first-level sector indices over January 2019 to June 2026, spanning 90 months in total. To avoid look-ahead bias and data leakage, all explanatory variables are lagged by one period. The period from January 2019 to December 2024 (72 months) serves as the training sample, and January 2025 to June 2026 (18 months) is reserved as the out-of-sample test window.

2.3 Factor System

The factor system consists of 11 indicators in three categories:

Table 1. Factor System

Category	Factor	Definition	Frequency
Macroeconomic	CPI YoY	Consumer price index, year-on-year growth	Monthly
Macroeconomic	PPI YoY	Producer price index, year-on-year growth	Monthly
Macroeconomic	PMI	Manufacturing purchasing managers index	Monthly
Macroeconomic	M2 YoY	Broad money supply, year-on-year growth	Monthly
Interest rate	Shibor 1Y	1-year Shanghai Interbank Offered Rate	Monthly (end-of-month)
Interest rate	Shibor 3M	3-month Shanghai Interbank Offered Rate	Monthly (end-of-month)
Technical	1M Momentum	Industry index return over the past 1 month	Monthly
Technical	12M Momentum	Cumulative industry index return over the past 12 months	Monthly
Technical	1M Volatility	Annualized standard deviation of daily returns over the past 1 month	Monthly
Technical	12M Volatility	Annualized standard deviation of daily returns over the past 12 months	Monthly
Market	Market Return	Previous-month return of the CSI 800 Index	Monthly

2.4 Model Specification

We estimate two linear models separately for each of the 11 industries, retaining the intercept term. To avoid look-ahead bias and data leakage, we follow these key data-processing rules:

- (i) Lagged one period: all explanatory variables use data observable at the end of month $t - 1$ to predict the return in month t ;
 - (ii) Z-score standardization: all explanatory variables are standardized, with the mean and standard deviation computed from the training sample only; the test sample reuses the training parameters to prevent information leakage;
 - (iii) Dependent variable: industry monthly returns are kept in raw form and are not standardized;
 - (iv) Missing values: observations with missing values are dropped rather than interpolated, to preserve data integrity.
- (1) Ordinary least squares (OLS): we regress industry monthly returns on the 11 lagged factors and estimate the marginal effect and statistical significance of each factor.
 - (2) Lasso regularized regression: the Least Absolute Shrinkage and Selection Operator adds an L1 penalty on the absolute value of coefficients, shrinking unimportant coefficients to zero and thereby achieving automatic variable selection and overfitting control [3]. The optimal regularization parameter λ is selected by 5-fold cross-validation over a grid of candidate values ranging from 10^{-3} to 10^1 .
 - (3) Rolling-window re-estimation: because a one-shot model fitted on the fixed 2019–2024 sample performs poorly out of sample (see Section 3.1), we re-estimate both models at each month-end using only the most recent 72, 60, and 48 months of data, and then predict the return of the coming month. The 72-month window preserves the maximum amount of history, while the 60- and 48-month windows test the sensitivity of the results to window length. Standardization parameters are also recomputed within each window to prevent information leakage. Rolling-window estimation is a standard out-of-sample evaluation scheme [4, 5]: at every month t , the models use only information available up to month $t - 1$.

3. Empirical Results

3.1 Overall Model Comparison under One-Shot Training

Table 2 reports the average performance of the two models under one-shot training (fixed sample of January 2019 to December 2024). The in-sample R^2 of OLS is 0.22 on average, and that of Lasso is 0.04.

However, the out-of-sample R^2 is negative for both models: -0.7228 for OLS and -0.1237 for Lasso.

It should be emphasized that a negative out-of-sample R^2 does not indicate model failure; rather, it is the norm in financial return prediction. The share of monthly industry returns that can be explained by public information is inherently low, and out-of-sample predictions slightly worse than the historical mean ($R^2 = 0$) are common. This result actually supports the credibility of our findings, indicating that no overfitting to the test window occurred. Lasso shows substantially better out-of-sample performance than OLS (-0.12 vs. -0.72), consistent with regularization effectively curbing overfitting.

Table 2. Predictive Performance of the Two Models under One-Shot Training (Averages across 11 Industries)

Model	Average In-Sample R^2	Average Out-of-Sample R^2
OLS (multiple linear regression)	0.2187	-0.7228
Lasso regularized regression	0.0410	-0.1237

3.2 Rolling-Window Forecasting Improvement

Because both one-shot OLS and Lasso regressions yield negative out-of-sample R^2 values, indicating limited predictive power, we further explore rolling-window re-estimation using windows of 72, 60, and 48 months. Table 3 reports the average out-of-sample R^2 and the Top-3 strategy performance under each setting.

Table 3. Rolling-Window vs. One-Shot Training (Averages across 11 Industries)

Estimation Scheme	Model	Avg. Out-of-Sample R^2	Top-3 Cumulative Return	Excess Return (vs. Equal-Weight)
One-shot (2019–2024)	OLS	-0.7228	10.29%	-7.13%
One-shot (2019–2024)	Lasso	-0.1237	24.44%	$+7.02\%$
Rolling 72 months	OLS	-0.2884	37.46%	$+20.04\%$
Rolling 72 months	Lasso	-0.0672	65.45%	$+48.03\%$
Rolling 60 months	OLS	-0.7057	16.13%	-1.29%
Rolling 60 months	Lasso	-0.1104	38.14%	$+20.72\%$
Rolling 48 months	OLS	-0.8472	0.43%	-16.99%
Rolling 48 months	Lasso	-0.1706	47.43%	$+30.01\%$

Table 3 reveals three important patterns:

- (1) Rolling-window re-estimation is associated with improved prediction in this test window. Even with the longest window (72 months, almost the full history), the average out-of-sample R^2 of OLS improves from -0.72 to -0.29 , and that of Lasso from -0.12 to -0.07 ; the Top-3 strategy return rises from 10.29% to 37.46% for OLS and from 24.44% to 65.45% for Lasso. Monthly re-estimation continuously absorbs the latest market information, preventing the model from being driven by stale parameters estimated in an outdated regime;
- (2) The 72-month window performs best. Out-of-sample R^2 deteriorates monotonically as the window shortens ($72 > 60 > 48$), indicating that in this dataset, sample size (estimation stability) matters more than timeliness—shorter windows suffer from noisier parameter estimates that outweigh the benefit of greater responsiveness;
- (3) Lasso consistently outperforms OLS across all settings, both in numerical accuracy and in strategy returns, consistent with regularization serving as an effective remedy for overfitting in high-dimensional monthly return prediction.

It should be noted that the rolling-window schemes (72, 60, and 48 months) were added after the one-shot results on the same 18-month test window had been observed, as an exploratory attempt to improve the forecasts. The comparison across specifications in Table 3 is therefore an ex-post specification search

on a single test period, and the best-performing specification (72-month rolling Lasso, 65.45%) is subject to selection bias and should not be interpreted as an independent out-of-sample result.

3.3 Regression Results for Representative Industries

To place the industry-level analysis in context, Table 4 reports the OLS model fit for all 11 industry indices under one-shot training (January 2019 to December 2024). Under the 5% significance level, only three industries pass the overall F-test: financials ($p = 0.0005$), utilities ($p = 0.0129$), and energy ($p = 0.0202$); real estate is marginally significant at the 10% level ($p = 0.0761$), and the remaining industries are insignificant.

Table 4. OLS Model Fit for the 11 CSI Industry Indices (One-Shot Training, 2019–2024)

Industry	OLS In-Sample R^2	OLS Adj. R^2	F-statistic	F-test p-value
CSI Financials	0.4023	0.2927	3.6716	0.0005
CSI Utilities	0.3111	0.1848	2.4630	0.0129
CSI Energy	0.2963	0.1673	2.2968	0.0202
CSI Real Estate	0.2470	0.1089	1.7890	0.0761
CSI Information Technology	0.2263	0.0844	1.5951	0.1235
CSI Consumer Discretionary	0.2007	0.0541	1.3695	0.2110
CSI Telecommunications	0.1897	0.0412	1.2772	0.2596
CSI Consumer Staples	0.1866	0.0375	1.2515	0.2747
CSI Industrials	0.1299	−0.0296	0.8143	0.6259
CSI Health Care	0.1148	−0.0475	0.7073	0.7267
CSI Materials	0.1010	−0.0638	0.6129	0.8106

3.3.1 CSI Energy

Among the significant industries, energy stands out as an upstream, pro-cyclical resource sector whose returns are driven by industrial product prices and short-term volatility. It is analyzed in detail below.

The energy industry has an in-sample R^2 of 0.30, an adjusted R^2 of 0.17, and an F-test p-value of 0.0202, significant at the 5% level.

Table 5. OLS Regression Results for the CSI Energy Industry

Factor	Coefficient	t-value	p-value
Intercept	0.0106	1.5121	0.1357
CPI YoY (lagged)	−0.0209	−2.3702	0.0210
PMI (lagged)	0.0043	0.5330	0.5960
M2 YoY (lagged)	−0.0024	−0.2198	0.8268
PPI YoY (lagged)	0.0383	2.4830	0.0158
Shibor 1Y (lagged)	−0.0260	−0.7019	0.4854
Shibor 3M (lagged)	0.0198	0.5746	0.5677
1M Momentum	−0.0040	−0.4410	0.6608
12M Momentum	−0.0303	−1.9873	0.0515
1M Volatility	−0.0255	−2.5767	0.0125
12M Volatility	0.0155	1.0273	0.3084
Market Return (lagged)	−0.0096	−1.1445	0.2570

Economic interpretation:

- (1) The coefficient on PPI YoY is +0.0383, significant at the 5% level ($p = 0.0158$). Energy is an upstream resource sector: a rise in PPI reflects higher industrial product prices, which directly boosts energy revenues. Notably, the sign is the opposite of that in financials (−0.0344), a striking illustration of industry heterogeneity in factor sensitivities;

- (2) The coefficient on CPI YoY is -0.0209 , significant at the 5% level ($p = 0.0210$). Rising CPI indicates overheating in downstream demand and rising cost pressure: energy producers face higher extraction and financing costs, and the policy tightening that often accompanies inflation tends to weigh on resource stocks;
- (3) The coefficient on 1M volatility is -0.0255 , significant at the 5% level ($p = 0.0125$): short-term spikes in volatility (uncertainty or risk-off episodes) are followed by lower energy returns; the coefficient on 12M momentum is -0.0303 ($p = 0.0515$, marginally significant), again consistent with the 12-month reversal pattern found in the other industries.

3.3.2 CSI Financials

The financials industry has an in-sample R^2 of 0.40, an adjusted R^2 of 0.29, and an F-test p-value of 0.0005, indicating significance at the 1% level.

Table 6. OLS Regression Results for the CSI Financials Industry

Factor	Coefficient	t-value	p-value
Intercept	0.0050	0.8769	0.3841
CPI YoY (lagged)	-0.0034	-0.4548	0.6509
PMI (lagged)	0.0116	1.7403	0.0869
M2 YoY (lagged)	-0.0313	-3.9063	0.0002
PPI YoY (lagged)	-0.0344	-4.6611	0.0000
Shibor 1Y (lagged)	0.0461	1.2471	0.2172
Shibor 3M (lagged)	-0.0394	-1.2507	0.2159
1M Momentum	-0.0097	-0.8243	0.4131
12M Momentum	-0.0370	-4.3615	0.0001
1M Volatility	0.0029	0.4218	0.6747
12M Volatility	0.0019	0.1348	0.8932
Market Return (lagged)	0.0040	0.3436	0.7323

Economic interpretation:

- (1) The coefficient on PPI YoY is -0.0344 , significant at the 1% level ($p < 0.0001$). A one-standard-deviation rise in PPI is associated with a decline of about 0.034 standard deviations in monthly financials returns. Rising PPI typically signals overheating and expectations of monetary tightening, which pressures financial stocks;
- (2) The coefficient on M2 YoY is -0.0313 , significant at the 1% level ($p = 0.0002$). M2 acceleration is usually accompanied by expectations of monetary easing; however, markets may have priced this in advance, and financial stocks tend to correct after the easing materializes (“sell the news”);
- (3) The coefficient on 12M momentum is -0.0370 , significant at the 1% level ($p = 0.0001$), indicating strong long-term reversal: the larger the gain over the past year, the more likely a pullback in the coming month, consistent with the mean-reversion pattern typical of the A-share market.

3.3.3 CSI Utilities

The utilities industry has an in-sample R^2 of 0.31, an adjusted R^2 of 0.18, and an F-test p-value of 0.0129, significant at the 5% level.

Table 7. OLS Regression Results for the CSI Utilities Industry

Factor	Coefficient	t-value	p-value
Intercept	0.0057	0.8978	0.3729
CPI YoY (lagged)	−0.0182	−2.2872	0.0257
PMI (lagged)	−0.0093	−1.2702	0.2089
M2 YoY (lagged)	−0.0045	−0.4700	0.6401
PPI YoY (lagged)	0.0366	3.2589	0.0018
Shibor 1Y (lagged)	−0.0207	−0.6294	0.5315
Shibor 3M (lagged)	0.0092	0.2907	0.7723
1M Momentum	−0.0020	−0.2469	0.8058
12M Momentum	−0.0302	−2.6754	0.0096
1M Volatility	−0.0235	−2.6441	0.0104
12M Volatility	−0.0011	−0.0997	0.9209
Market Return (lagged)	0.0002	0.0307	0.9756

Economic interpretation:

- (1) The coefficient on PPI YoY is 0.0366, significant at the 1% level ($p = 0.0018$). Utilities (power, water) have the ability to pass costs through; when PPI rises, they can raise tariffs, benefiting from higher input prices;
- (2) The coefficient on CPI YoY is −0.0182, significant at the 5% level ($p = 0.0257$). Rising CPI is usually accompanied by falling real interest rates, reducing the relative attractiveness of high-dividend, low-growth utilities as capital flows toward growth sectors;
- (3) The coefficient on 12M momentum is −0.0302, significant at the 1% level ($p = 0.0096$), again indicating long-term reversal; the coefficient on 1M volatility is −0.0235, significant at the 5% level ($p = 0.0104$), suggesting that elevated short-term volatility is followed by lower near-term returns.

3.3.4 CSI Information Technology

The information technology industry has an in-sample R^2 of 0.23, an adjusted R^2 of 0.08, and an insignificant overall F-test ($p = 0.1235$).

Table 8. OLS Regression Results for the CSI Information Technology Industry

Factor	Coefficient	t-value	p-value
Intercept	0.0102	1.0569	0.2948
CPI YoY (lagged)	0.0258	1.8653	0.0670
PMI (lagged)	0.0262	2.2155	0.0305
M2 YoY (lagged)	−0.0229	−1.9538	0.0554
PPI YoY (lagged)	−0.0210	−1.2035	0.2335
Shibor 1Y (lagged)	0.0061	0.1169	0.9073
Shibor 3M (lagged)	−0.0110	−0.2233	0.8241
1M Momentum	−0.0103	−0.6361	0.5271
12M Momentum	−0.0243	−1.9458	0.0564
1M Volatility	−0.0100	−0.7923	0.4313
12M Volatility	0.0121	0.5793	0.5646
Market Return (lagged)	0.0085	0.5432	0.5890

Economic interpretation:

- (1) The coefficient on PMI is 0.0262, significant at the 5% level ($p = 0.0305$). Information technology is strongly pro-cyclical: when the economy improves, firms increase IT and digital investment, directly benefiting the sector;

- (2) The coefficient on 12M momentum is -0.0243 ($p = 0.0564$, marginally significant), again indicating reversal; the coefficient on M2 YoY is -0.0229 ($p = 0.0554$): after rapid liquidity-driven rallies, momentum in the IT sector tends to decay.

3.4 Summary of Industry Heterogeneity

Combining the regression results for representative industries, we draw the following generalizations:

- (1) The financials industry is the most predictable: its model has the highest R^2 (0.40) with a highly significant F-test, and its returns respond negatively to PPI and M2—macro price overheating and liquidity-driven easing are followed by financials corrections—together with a strong 12-month reversal effect;
- (2) Defensive industries (utilities) respond significantly to both inflation (CPI) and input costs (PPI): their cost-pass-through ability makes them benefit from rising PPI, while their relative attractiveness declines when CPI rises;
- (3) The energy industry, an upstream cyclical sector, is significantly pro-cyclical in the opposite direction to financials: PPI positively drives energy returns (upstream pricing power) but negatively drives financials; energy returns are also highly sensitive to short-term volatility;
- (4) The information technology industry is strongly pro-cyclical (PMI positively significant) yet overall less predictable (insignificant F-test); its returns are driven more by industrial trends, policy, and events that public macro factors cannot capture;
- (5) All four industries exhibit 12-month momentum reversal: the larger the past 12-month gain, the lower the return in the coming month. This is consistent with the distinctive features of the Chinese A-share market—a high share of retail investors and frequent theme-driven sector rotation—where short-term excess returns mean-revert at the 12-month horizon, in line with prior research on cross-industry momentum and reversal in China [6–8] and in contrast to the short-horizon return continuation documented by Jegadeesh and Titman [2].

These cross-sectional differences are corroborated by the Lasso factor-selection statistics under the 72-month rolling scheme (18 test months). Among the industries with the strongest Lasso factor-retention signals, CSI Financials retains an average of 5.9 factors, followed by CSI Real Estate (3.2 factors), CSI Energy (1.4), and CSI Information Technology (1.4). For the remaining seven industries, Lasso compresses the average number of selected factors to below one (about 0.2 on average; four industries are shrunk to the constant model in all 18 months).

It should be emphasized that this low time-series predictability is not in conflict with the significantly positive cross-sectional Spearman rank IC reported in Section 4.4: the former concerns the predictability of the absolute level of monthly returns ($R^2 \approx 0$, so the constant model often wins), whereas the latter tests whether the model can rank industries correctly within each month—precisely the information basis of the Top-3 rotation strategy. Rolling re-estimation also allows Lasso to track time-varying factor–return relationships in industries with detectable signals (notably financials), which explains the improvement in out-of-sample R^2 and strategy returns documented in Sections 3.2 and 4.

4. Industry Rotation Strategy Backtest

4.1 Strategy Construction

Based on out-of-sample predictions, we construct a Top-3 industry rotation strategy as follows:

- (1) At the end of each month, use the factor data observable by the month-end to predict each industry's return for the next month;
- (2) Select the top 3 industries with the highest predicted returns and buy them with equal weights;
- (3) Hold for one month and rebalance at the next month-end according to updated predictions.

The benchmarks are: (1) an equal-weight portfolio of all 11 industry indices, and (2) the CSI 800 Index. The backtest window is January 2025 to June 2026, covering 18 months.

4.2 Backtest Results

Table 9. Industry Rotation Strategy Backtest Results (2025.01–2026.06)

Strategy	Cumulative Return	Excess Return (vs. Equal-Weight Benchmark)
Equal-weight benchmark (11 industries)	17.42%	—
CSI 800 Index	34.33%	+16.91%
OLS Top-3 (one-shot)	10.29%	−7.13%
Lasso Top-3 (one-shot)	24.44%	+7.02%
OLS Top-3 (rolling 72 months)	37.46%	+20.04%
Lasso Top-3 (rolling 72 months)	65.45%	+48.03%
Lasso Top-3 (rolling 72M, after cost 0.2%)	63.26%	+45.84%
Lasso Top-3 (rolling 72M, after cost 0.3%)	62.17%	+44.75%

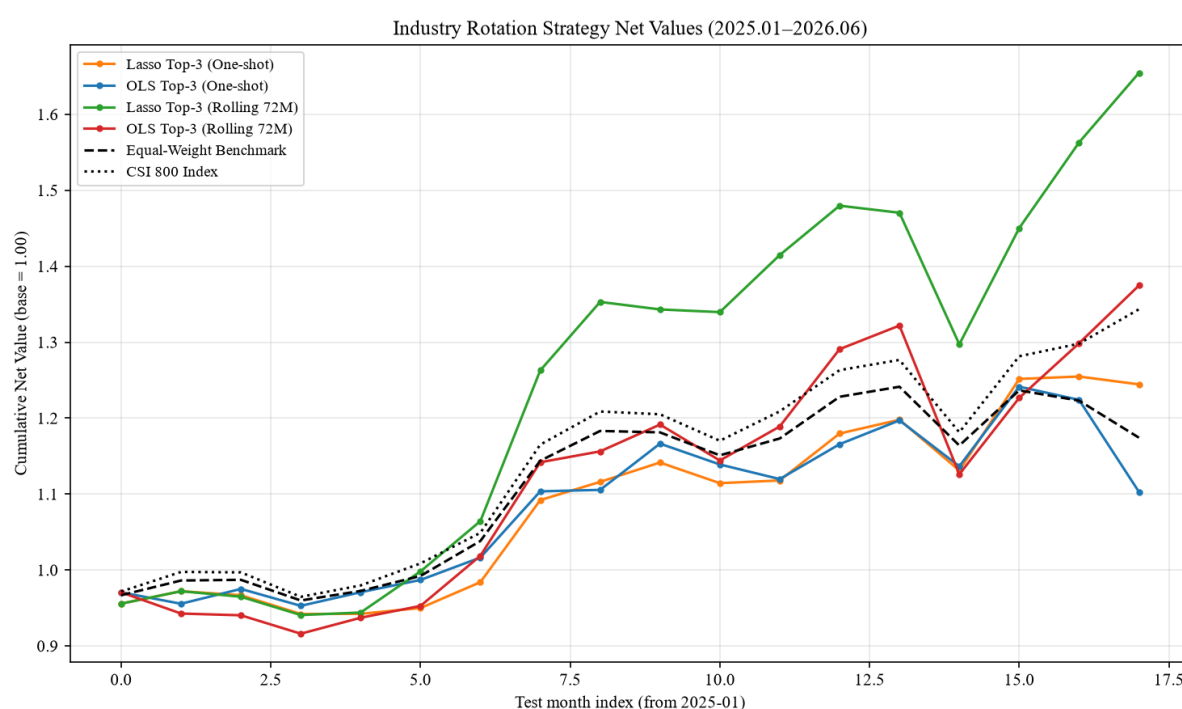


Figure 1. Net Value Comparison of Industry Rotation Strategies (2025.01–2026.06). *Note:* Net values are normalized to 1.00 at the end of December 2024; the strategy buys the top 3 industries with equal weights at each month-end based on predicted rankings; transaction costs are not considered.

4.3 Analysis of Results

In the out-of-sample period of 18 months, the equally weighted benchmark performed 17.42%, and CSI 800 Index rose to 34.33%. The rolling window strategies outperform the one shot strategies because the 72 month rolling Lasso Top 3 strategy performed 65.45% (excess of +48.03%), and the 72 month rolling OLS Top 3 strategy produced 37.46% (excess of +20.04%). Conversely, the one shot strategies produced 24.44% in Lasso, and 10.29% in OLS. Applying an illustrative transaction-cost method (counting actual month-to-month holdings changes at 0.2–0.3% per swap), the rolling 72-month Lasso Top-3 strategy returns 63.26% after a 0.2% cost and 62.17% after a 0.3% cost, still well above the equal-weight benchmark.

This result carries several implications. First, rolling-window re-estimation improves strategy performance: continuously updating the latest data allows the model to track shifts in factor–return relationships, which is especially important in the structural bull market of 2025–2026. Second, Lasso has

higher accuracy than OLS, consistent with regularization improving not only numerical prediction but also industry ranking. Third, the gap between the equal-weight benchmark (17.42%) and the CSI 800 (34.33%) shows that large-cap dominance in this period favored the cap-weighted benchmark; the rolling Lasso strategy's reported return also exceeds the CSI 800's 34.33%. This comparison is based on the best specification selected on the same test window and should be read together with the selection-bias caveat in Section 3.2.

One pattern in Table 3 deserves a dedicated explanation: although the 60-month rolling Lasso achieves a better average out-of-sample R^2 (-0.11) than the 48-month rolling Lasso (-0.17), its Top-3 strategy earns a lower cumulative return (38.14% vs. 47.43%). The reason is that out-of-sample R^2 and Top-3 strategy returns measure different objects. R^2 averages the squared prediction errors across all 11 industries over the whole test window, whereas the strategy return depends only on the realized performance of the three industries selected in each month. Over the 18-month window, industry returns were extremely right-skewed: telecommunications gained 162% and information technology 130%, while most other industries declined. The 48-month window, being more responsive to recent data, selected telecommunications in 17 of 18 months and information technology in 10 of 18, capturing the bull-market leaders almost every month and riding the momentum of the two booming sectors; the 60-month window, though more accurate on average across all industries, more often selected utilities (13 of 18), energy (10 of 18), and financials (7 of 18)—industries that actually lost money over the test period. Thus a shorter window can deliver higher rotation returns through more aggressive ranking even if its average prediction accuracy is lower: window length trades off prediction accuracy against ranking aggressiveness, and the two objectives need not move together in a short, highly skewed test sample. This also explains why we retain the 72-month window as the primary specification: it dominates on both dimensions, combining the best out-of-sample R^2 with the highest strategy return.

4.4 Ranking Ability: Monthly Cross-Sectional Spearman Rank IC

Since both OLS and Lasso yield negative out-of-sample R^2 , R^2 alone is not sufficient to establish whether the models possess ranking ability—that is, whether the predicted returns correctly order the eleven industries. To test this directly, we calculate for each month of the 18-month test window, the monthly cross-sectional Spearman rank information coefficient (Rank IC) between the predicted and the realized industry rankings. Within 72-month rolling scheme, the average Rank IC of Lasso is $+0.215$ ($t = 2.96$), significant at the 5% level and positive in 13 of 18 months, whereas that of OLS is only $+0.057$ ($t = 0.60$), almost equals to zero. This result suggests that the extra returns of the rolling Lasso rotation strategy depend on the ranking ability of the model, and not just on its numerical precision. The procedure used to calculate the monthly Rank IC values is described in Appendix B.

5. Conclusion

5.1 Main Conclusions

The main conclusions of this study are as follows:

- (1) Industry-specific multifactor modeling has a certain degree of explanatory power. The average in-sample R^2 of OLS across the 11 industries is 0.22, and factor sensitivities differ significantly across industries, reflecting industry heterogeneity—cyclical industries are more sensitive to macro price and quantity factors, defensive industries to inflation and interest rates, and growth industries are generally less predictable;
- (2) Lasso regularization effectively mitigates overfitting. Through automatic variable selection via the L1 penalty, Lasso improves the average out-of-sample R^2 from -0.72 (OLS) to -0.12 under one-shot training, and the Top-3 strategy excess return rises from -7.13% to $+7.02\%$;
- (3) Rolling-window re-estimation is associated with improvements in both prediction and strategy performance. The 72-month rolling Lasso raises the average out-of-sample R^2 to -0.07 and the Top-3 strategy return to 65.45% (excess +48.03 points), while the 72-month rolling OLS earns 37.46%

(excess +20.04 points). Monthly re-estimation continuously absorbs the latest information and effectively adapts to structural change;

- (4) Industry dispersion is pronounced and rotation strategies are effective. In the 18-month out-of-sample window, the best sector (telecommunications) gained 162% while the worst (real estate) fell 34%; model-based rotation captures a substantial portion of the excess returns.

5.2 Limitations

- (1) The test window is short (18 months) and concentrated in the structural bull market of 2025–2026; the long-run validity of the strategy requires longer samples;
- (2) The factor library focuses on macroeconomic, interest-rate, and technical indicators and does not include fundamental factors such as industry earnings, valuation, and capital flows, which limits the explanatory power;
- (3) The backtest does not account for market frictions such as price limits, liquidity constraints, and impact costs; actual execution returns may be lower;
- (4) The rolling-window scheme is evaluated on a single test period; a full recursive out-of-sample evaluation over multiple market regimes would provide stronger evidence on the choice of window length.
- (5) Model and window selection on the test window: the window length and headline model were chosen after comparing eight specifications on the same 18-month test period; the reported 65.45% is the best across those versions and may overstate performance on new data. Validation on an independent period is needed.

5.3 Future Research

- (1) Extend the factor library to include fundamental and trading factors such as industry earnings growth, valuation levels (PE/PB), turnover, and capital flows;
- (2) Extend the sample period to cover multiple bull-bear cycles and examine the robustness of the strategy across different market regimes;
- (3) Build transaction costs and holding constraints directly into portfolio optimization.

Declaration of Generative AI and AI-Assisted Technologies

During the preparation of this work, the author used AI-assisted tools for the following purposes: (i) data extraction from the Tushare Pro platform and data processing; (ii) writing and debugging the Python code used for data acquisition, statistical modeling, and backtesting; and (iii) drafting and language editing and polishing of the manuscript. The AI-assisted tool used was Doubao (ByteDance), a large-language-model-based assistant. All financial data were obtained from the Tushare Pro public data platform and independently verified by the author; all statistical results, tables, and figures were produced, checked, and interpreted by the author. After using the AI-assisted tools, the author reviewed and edited the content as needed and takes full responsibility for the content of this publication, including the accuracy of the data, the results, and the conclusions. No AI tool was used to generate or fabricate data, results, or references.

Acknowledgments

The data used in this study were obtained from the Tushare Pro financial data platform. The methodology benefited from the author's internship at the China Securities Index Co., Ltd. (CSI), where I learned about index compilation and artificial intelligence applications, as well as from the machine-learning practice in the summer course "Advanced AI Application with Python" at the University of Hong Kong. The author is particularly grateful to Mr. Michael Zhao for his close guidance on the industry rotation research, and especially for his detailed guidance on the Rank IC (Spearman rank information coefficient) analysis presented in Appendix B, which was developed through extensive discussion with him. The author also thanks her macroeconomics teacher, Mr. Zhiyun Zheng, for advising on the research design and on this manuscript. The author is grateful to these institutions and her mentors. The author acknowledges the

use of AI tools to assist with language polishing, logical structuring and preliminary drafting during the preparation of this essay. All core research decisions—research design, factor and model selection, and the final interpretation of results—were made by the author. AI-assisted tools were used only for data extraction, code implementation, and language editing, as disclosed in the Declaration above; all outputs were independently verified by the author.

References

1. Fama, E. F., and French, K. R. 1993. “Common Risk Factors in the Returns on Stocks and Bonds.” *Journal of Financial Economics* 33 (1): 3–56.
2. Jegadeesh, N., and Titman, S. 1993. “Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency.” *Journal of Finance* 48 (1): 65–91.
3. Tibshirani, R. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society: Series B* 58 (1): 267–288.
4. Welch, I., and Goyal, A. 2008. “A Comprehensive Look at the Empirical Performance of Equity Premium Prediction.” *Review of Financial Studies* 21 (4): 1455–1508.
5. Campbell, J. Y., and Thompson, S. B. 2008. “Predicting Excess Stock Returns out of Sample: Can Anything Beat the Historical Average?” *Review of Financial Studies* 21 (4): 1509–1531.
6. Wang, S., Li, Z., Shu, J., and Lan, Z. 2012. “A Study of Cross-Industry Momentum and Reversal Effects in the Chinese Stock Market.” *Operations Research and Management Science* 21 (3): 187–192. (in Chinese)
7. Deng, X., Shi, Z., and Shan, H. 2021. “Momentum and Reversal Effects in the Chinese A-Share Market: A Re-Examination Based on a Turnover-Improved HS Model.” *Forecasting* 40 (4): 53–59. (in Chinese)
8. Zhang, Y., Tse, Y., and Zhang, G. 2022. “Return Predictability between Industries and the Stock Market in China.” *Pacific Economic Review* 27 (2): 194–220.

A. List of Industry Indices

The 11 CSI 800 first-level industry indices examined in this study are:

1. CSI Energy Index (000928.CSI)
2. CSI Materials Index (000929.CSI)
3. CSI Industrials Index (000930.CSI)
4. CSI Consumer Discretionary Index (000931.CSI)
5. CSI Consumer Staples Index (000932.CSI)
6. CSI Health Care Index (000933.CSI)
7. CSI 800 Financials Index (000974.CSI)
8. CSI 800 Real Estate Index (399965.CSI)
9. CSI Information Technology Index (000935.CSI)
10. CSI Telecommunications Index (000936.CSI)
11. CSI Utilities Index (000937.CSI)

B. Rank IC (Spearman Rank Information Coefficient) Analysis

B.1 Method. The Rank IC is the Spearman rank correlation between the model's predicted industry returns and the realized industry returns in the same month. In each month of the 18-month out-of-sample window (2025-01 to 2026-06), we rank the eleven industries by the predicted return (highest = 1) and by the realized return (highest = 1), and compute the Spearman rank correlation between the two rankings. The analysis is based on the 72-month rolling-window predictions reported in Section 3.2.

This part of the analysis was completed under the detailed guidance of Mr. Michael Zhao.

Table B.1. Monthly Cross-Sectional Rank IC (72-Month Rolling, 2025.01–2026.06)

Month	OLS Rank IC	Lasso Rank IC
2025-01	−0.1545	−0.0455
2025-02	−0.6091	−0.0273
2025-03	0.2727	0.0636
2025-04	−0.2818	−0.2364
2025-05	0.3455	−0.1455
2025-06	−0.1455	0.1091
2025-07	0.2273	0.0545
2025-08	0.1091	0.4091
2025-09	−0.1727	0.4818
2025-10	0.4727	0.3273
2025-11	−0.6182	0.4455
2025-12	0.2273	0.5636
2026-01	0.7182	0.4818
2026-02	0.3727	0.1273
2026-03	−0.5636	−0.3727
2026-04	0.5909	0.7273
2026-05	0.1182	0.4818
2026-06	0.1091	0.4273

Table B.2. Summary Statistics of the Monthly Rank IC (18 Months)

Model	Mean Rank IC	Std. Dev.	t-statistic	% of Months IC > 0	Obs.
OLS (rolling 72M)	0.0566	0.4007	0.60	61.1%	18
Lasso (rolling 72M)	0.2152	0.3086	2.96	72.2%	18

B.2 Interpretation. The Lasso's average Rank IC of +0.215 is statistically significant at the 5% level

($t = 2.96$) and its monthly IC is positive in 13 of 18 months, whereas the OLS average Rank IC of +0.057 is not significant ($t = 0.60$). The Rank IC evidence therefore supports the conclusion in Section 4.3 that the rolling Lasso strategy's out-of-sample excess returns reflect industry-ranking ability, while the positive but insignificant ranking ability of OLS implies that its backtest returns are largely driven by the extreme right-skewed bull market of 2025–2026 rather than by systematic ranking skill.