BRINGING DISCIPLINE TO THE FACTOR ZOO: EVIDENCE FROM CHINA

BY

ZHANERKE DOSPAN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Finance
in the Graduate School of Business
Nazarbayev University, 2025

Astana, Kazakhstan

Advisor: Dr. Francesco Rocciolo

Abstract

Determining the true contribution of each factor recently found in the literature poses a challenge due to the multidimensionality of control factors and omitted variable bias. To address these issues, this thesis applies advanced model selection approach in machine learning, namely double-selection LASSO in the context of Chinese A-shares market. Results presented in this paper confirms that, although most of these new factors are redundant relative to existing factors, some factors have significant statistical explanatory power compared to the hundreds of factors proposed in the past. Further, we compared the results of alternative methods to identify whether the newly proposed factors that were significant in the main analysis are robust.

Keywords: DS LASSO, China, machine learning, factor zoo.

1 Introduction

As new factors were discovered, controlling the proliferation of factors has become a challenge in the field of empirical asset pricing, and many researchers attempted to bring discipline to these factors using different methodologies.

The main objective of this paper lays on filtering down the factor zoo in the Chinese A-shares market into a handful of significant factors using DS LASSO (double-selection least absolute shrinkage and selection operator) methodology. This method addresses the potential model selection errors caused by the omission of variables that are usually overlooked in standard methods.

We try to identify whether factors published between 2016 and 2020 perform better than already published factors. Furthermore, the country of interest was chosen due to the specificities of the Chinese A-shares market, such as changing market conditions, more pronounced investor sentiments, high volatility and unique market structure that is common to emerging markets (Yang et al, 2025).

Given any new factor, we can discriminate whether it is truly a new factor or a redundant relative to existing factors. Thus, allowing us to evaluate whether the new factors provide new exposures to risk or not.

Historically, all the factors that were introduced have been effectively benchmarked against the same model, usually Fama-French factor model or some variations. When evaluating the predictive power of a newly proposed factor relative

to such benchmark model, new factors might seem useful. But, when tested against the extended benchmark consisting of many existing factors, we end up rediscovering the same anomalies, the same factors, over and over again. This shows that whether the new factor is significant or not depends on the benchmark model. Consequently, the central question becomes whether the benchmark is sufficient or not. However, when each candidate factor is evaluated against an extensive benchmark that includes every possible control factor that has been introduced so far, employing conventional approaches are inefficient and lead to invalid inferences due to curse of dimensionality (Feng et al, 2019). Therefore, we exploit a double-selection (DS) LASSO methodology capable of handling the multidimensional scope of factor zoo.

2 Literature Review

In 1960s the Capital Asset Pricing Model was independently developed by Treynor (1961), Sharpe (1964), Lintner (1965), and Mossin (1966) and many works confirmed its predictions including Black, Jensen and Scholes (1972), Fama and MacBeth (1973). This model only accounted for a singular market risk factor.

Introduced by Ross in 1976, Arbitrage Pricing Theory showed that the model not only consisted of a single factor but multiple factors. Thus, it provided a

theoretical framework for multifactor models, making any test of an empirical linear multifactor asset pricing model eventually a test of the APT.

In their paper titled "The Cross-Section of Expected Stock Returns" Fama and French (1992) found that in the periods after 1960s market beta could not explain the cross-section of average stock returns, indicating that the relation between average return and market beta was flat. Hence, empirical evidence showed that CAPM did not hold anymore. In this paper they suggested that stock risks are multidimensional and found that there were significant size and value effects in average stock returns for the 1963-1990 period. As a result, they found that three systematic risk factors: market risk premium, size factor "SMB", value factor "HML" better explain the cross section of expected returns than a single factor, leading to the introduction of the three-factor model in 1993 (Fama and French, 1993). In 1997 Carhart proposed four-factor model by adding fourth factor "Up minus down" to account for the momentum factor. Fama-French 5 factor model was presented in 2015 by adding profitability "RMW" and investment "CMA" factors to the three-factor model.

Over the few decades, many other factors have been discovered in the asset pricing literature which led to proliferation of over hundreds of potential risk factors. These factors are known as "factor zoo" and it was first brought up by Cochrane in 2011. Several important questions were raised in John H. Cochrane's paper

"Presidential address: Discount rates" and many researchers attempted to address them using different methodologies.

Harvey et al., (2016) argued that using standard statistical significance cutoffs are not appropriate. Their analysis showed that most of the factors discovered are likely false discoveries and debated that t-statistic for the recently discovered factors should surpass 3.0. The paper provided series of recommended "cutoffs" from 1967 to 2016.

Green et al., (2017) explored the question regarding the characteristics that contain independent information about average returns. Their methodology followed that of Fama-French's (1992) by estimating series of Fama-MacBeth regressions while assessing all 94 firm characteristics simultaneously. They identified that 12 of these characteristics were independent in non-microcaps for 1980-2014, whereas the remaining 84 characteristics contained redundant information. However, only two characteristics passed the robustness check for post-2003 period.

Feng et al. (2020) utilized the DS procedure of Belloni, Chernozhukov, and Hansen (2014) that we will be employing in this paper. They constructed a data library consisting of 150 tradable factors and proposed a method for systematically selecting the most efficient control model from the set of hundreds of factors while accounting for model selection biases. Despite having a large set of control factors, they found that profitability, HXZ's investment factor, intermediary investment and

QMJ are strongly significant, while other newly proposed factors were statistically insignificant.

Harvey et al. (2021) addressed the following challenges in the asset pricing: first, they tackle the problem of test multiplicity using bootstrapping approach which enforced the null hypothesis to fit the data and by applying stepwise model selection method tailored for panel regressions; second, to avoid the arbitrary nature of portfolio formation they used individual stocks instead of portfolios. Lastly, they claimed that the market factor is superior than any other factor that are of second importance.

Our study focuses on one of the questions posed by Cochrane (2011), regarding "how many of the new factors in the "factor zoo" are really important?" in the context of China. As the current literature mainly concentrates on the developed markets such as the USA, we took on the challenge to fill this gap in the field of empirical asset pricing.

3 Data Description

The test assets were constructed using stock price data of 1093 Chinese A shares listed on Shanghai (SSE) and Shenzhen (SZSE) Stock Exchanges sourced from Bloomberg database. Then we matched monthly stock returns with annual accounting data collected from Compustat Global (WRDS) by ISIN. The sample

period that we covered in this study starts from January of 2001 and end in December of 2024 at a monthly frequency, totaling 288 months.

This paper uses 2x3 bivariate sorted 120 value-weighted portfolios formed on size and 20 various characteristics. We employed 2x3 sorting according to Fama-French 5 factor portfolio construction (Fama, E. F., & French, K. R., 2015). The choice of using 2x3 double sorting is in order to have enough stocks in each bin. Characteristic-sorted portfolios were constructed instead of using individual stocks mainly due to their stable risk exposures. Also, to limit the influence of micro caps we used value-weighted returns instead of equal-weighted. Financial firms were excluded because of their highly leveraged nature.

Breakpoints to allocate stocks were determined from June values each year based only on Shanghai Stock Exchange (SSE) stocks mainly because SSE stocks are larger and more liquid than Shenzhen (SZSE) stocks. Then applied it to sort all stocks using these thresholds into 2x3 bivariate sorted portfolios. First, we divided them by size using median split into small and big groups, then separately by 20 characteristics into high, medium, low groups, with the first cutoff at 30^{th} percentile, and second cutoff at 70^{th} percentile.

Accounting data from fiscal year-end (December year t-1) was used to form portfolios from July of year t to June of year t+1 to take into account look-ahead bias. We implemented annual rebalancing, thus the values of portfolios formed in

June each year were assigned for the next 11 months in order to align with the Fama-French methodology.

Portfolios were sorted based on the following 20 characteristics: book to market equity, operating profitability, gross profitability, return on assets, return on equity, operating cash flow to assets, EBITDA to assets, net profit margin, tangibility, net debt to assets, current ratio, quick ratio, z-score, capital expenditure to assets, 52-week high momentum, cash flow to price, sales to price, dividend to price, asset turnover, accruals.

Table 1 showcases the average monthly excess returns for 120 value-weighted 2x3 bivariate sorted portfolios formed on size and 20 various firm characteristics with table entries measured in percentages. Our sample dataset covers the period from January 2001 to December 2024. The monthly average excess returns for each size-characteristic pair are found on the intersection of respective columns indicating size, and the rows with 3 tiers for each characteristic.

Table 1: Descriptive statistics for the value-weighted test portfolios

Bivariate sorted 2x3 Portfolios			
	Low	Medium	High
		Panel: Size-B/M Portfolios	
Small	2.40	2.45	2.21
Big	1.48	1.13	1.30
		Panel: Size-OP Portfolios	
Small	2.65	2.33	2.36
Big	1.28	1.30	1.42
		Panel: Size-GP Portfolios	
Small	2.41	2.36	2.53
Big	0.97	1.43	1.50

		Panel: Size-ROA Portfolios	
Small	2.38	2.43	2.51
Big	1.13	1.49	1.41
		Panel: Size-ROE Portfolios	
Small	2.58	2.36	2.51
Big	1.28	1.34	1.44
		Panel: Size-OCF/AT Portfolios	
Small	2.59	2.36	2.49
Big	1.38	1.41	1.40
		Panel: Size-EBITDA/AT Portfolios	
Small	2.64	2.44	2.20
Big	1.34	1.46	1.36
		Panel: Size-NPM Portfolios	
Small	2.42	2.37	2.50
Big	1.43	1.52	1.33
		Panel: Size-Tangibility Portfolios	
Small	2.50	2.35	2.55
Big	1.33	1.52	1.35
		Panel: Size-Net Debt/AT Portfolios	
Small	2.62	2.32	2.18
Big	1.63	1.28	1.02
		Panel: Size-Current Ratio Portfolios	
Small	2.21	2.28	2.69
Big	1.08	1.44	1.51
		Panel: Size-Quick Ratio Portfolios	
Small	2.27	2.26	2.67
Big	1.02	1.49	1.49
		Panel: Size-Z-Score Portfolios	
Small	2.13	2.41	2.56
Big	0.97	1.34	1.66
		Panel: Size-CapEx/AT Portfolios	
Small	2.59	2.41	2.34
Big	1.49	1.32	1.31
		Panel: Size-Momentum Portfolios	
Small	3.09	2.21	1.28
Big	1.55	1.51	1.04
		Panel: Size-CF/P Portfolios	
Small	2.75	2.30	2.17
Big	1.43	1.38	1.23
		Panel: Size-S/P Portfolios	
Small	2.73	2.20	2.26
Big	1.38	1.39	1.18
		Panel: Size-D/P Portfolios	
Small	2.36	2.12	2.13
Big	1.40	1.26	1.33
		Panel: Size-Asset Turnover Portfolios	

Small	2.49	2.38	2.45
Big	1.20	1.44	1.48
	Panel: Size-Accruals Portfolios		
Small	2.45	2.51	2.39
Big	1.23	1.32	1.51

Note: All numbers are in percentages

From table 1 we can clearly see that small size corresponds to higher monthly average returns for every level of each characteristic, which is in line with size effect. The table also depicts a positive relationship between the excess returns and the firm characteristics for current ratio and z-score in both "small" and "big" sizes of firms. Inversely, Net Debt/Assets, CapEx/Assets, Momentum, CF/P exhibit negative effects to monthly average returns. While controlling for size, the average returns increase with characteristics OP, GP, ROA, ROE, Asset Turnover and Accruals. For small size, average returns decreased when EBITDA/Assets ratio increased, presenting an inverse relationship between the two. Other characteristics such as B/M, OCF/Assets, Net profit margin, Tangibility, Quick Ratio, S/P and D/P have demonstrated an ambiguous effect in relation to the average excess returns in our sample.

Proceeding to our factor dataset, it consists of Fama-French 5 factors and Momentum factor for emerging markets obtained from Kenneth R. French data library and added 153 factors that came from Global Factor data for the Chinese market developed by Jensen et al. (2023).

The factors that we used were excess returns in USD and were calculated using U.S. one month T-bill rate as a risk-free rate. Therefore, after converting stock returns expressed in CNY to USD, the one-month Treasury Bill was used as the risk-free rate for the purpose of calculating the excess returns of the stocks to stay consistent with the way the factors were constructed.

In this paper, g_t represents newly published candidate factors, while h_t represents all the factors that have been published before 2016 in our dataset. After removing h_t factors highly correlated with each other, our sample consisted of 98 h_t factors and 13 g_t factors.

In the appendix, the Table 5 displays the factors listed according to their publication year from 1972 to 2020. Factors from the Global Factor Dataset are classified into 13 fundamental themes. We split factors into two: factors released in the period from 2016 to 2020 were our new factors, and all the other factors were treated as a set of existing factors in the literature.

Table 2 displays the descriptive statistics with mean, standard deviation, skewness and kurtosis values of 13 candidate factors introduced between 2016-2020 period that were obtained from the Global Factor data library (Jensen et al., 2023).

Table 2: Summary statistics of 13 candidate factors

	Mean	SD	Skewness	Kurtosis	N obs
Cash-based operating profits-to-lagged book assets (cop_atl1)	0.292	3.195	0.141	5.687	288
Market correlation (corr_1260d)	-0.103	4.1	2.278	25.239	288

Cash-based operating profits-to-book assets (cop_at)	0.292	2.916	-0.390	4.507	288
Mispricing factor: Management (mispricing_mgmt)	0.237	2.469	0.214	4.662	288
Mispricing factor: Performance (mispricing_perf)	0.204	4.461	-0.177	3.847	288
Cash-based operating profits-to-book assets (ocf_at)	0.215	2.581	-0.361	4.369	288
Change in operating cash flow to assets (ocf_at_chg1)	0.048	1.289	-0.365	4.124	288
Operating profits-to- book assets (op_at)	0.166	3.503	-0.182	4.592	288
Operating profits-to-lagged book assets (op_atl1)	0.15	3.683	-0.020	4.488	288
Quality minus Junk: Composite (qmj)	0.319	3.553	-0.390	5.02	288
Quality minus Junk: Growth (qmj_growth)	0.036	2.443	-0.300	6.91	288
Quality minus Junk: Profitability (qmj_prof)	0.156	3.481	-0.279	4.077	288

Note: All numbers are in percentages

As presented in the table, quality minus junk: composite (qmj) yields the highest mean return of 0.319%, followed by cash based operating profits to lagged book assets (cop_atl1) and cash based operating profits to book assets (cop_at) both of which with the mean returns of 0.292%. The third largest average return is 0.237% corresponding to mispricing factor: management (mispricing_mgmt). Mispricing factor: performance (mispricing_perf) exhibits the highest standard deviation with the value of 4.461%. The factor with the second highest SD is market correlation (corr_1260d) at 4.1%, followed by operating profitability to lagged book assets (op_atl1) with 3.683%, representing their high divergence from the mean.

4 Methodology

Double Selection method was adapted from Belloni et al. (2014) to test the marginal contribution of factors recently found in the literature. The methodology we are implementing starts by using a linear specification of SDF for the demonstration of how the DS LASSO method was attained:

$$m_t := \gamma_0^{-1} - \gamma_0^{-1}\lambda_v^T v_t := \gamma_0^{-1}\left(1 - \lambda_g^T g_t - \lambda_h^T h_t\right),$$

where $\gamma_0$stands for the zero-beta rate, g_t represents a d × 1 vector of new factors to be tested, while h_t indicates a set of p × 1 vectors of existing factors. We took the demeaned values of both g_t and h_t, thus allowing factor innovations that satisfies the following conditions: $E(g_t) = 0$ and $E(h_t) = 0$. λ_g and λ_h are the respective d × 1 and p × 1 vectors of parameters. From now on λ_g and λ_h will be referred to as the SDF loadings of the factors g_t and h_t, respectively.

Aside from the two groups of factors g_t and h_t, another vector n×1 of test portfolio returns was included, denoted by r_t and in consistency with (1) it should satisfy the following equation:

$$E(r_t) = \iota_n\gamma_0 + C_v\lambda_v = \iota_n\gamma_0 + C_g\lambda_g + C_h\lambda_h,$$

where ι_n is an n-dimensional vector with all elements of 1 in order to align the dimension, $C_a = Cov(r_t, a_t)$, for $a = g, h$ or some random term v. In addition, the r_t dynamics exhibit the standard linear factor model:

$$r_t = E(r_t) + \beta_g g_t + \beta_h h_t + u_t,$$

where n × d and n× p factor loading matrices are denoted by β_g and β_h and the company's characteristic income in n× 1 vector is indicated by ut with $E(u_t)$= 0 and Cov(u_t, v_t) = 0.

The double-selection (DS) approach procedure consists of the following steps:

Step 1. We run a cross-sectional LASSO regression on the sample covariances between set of h_t factors and average returns,

$$min_{\gamma,\lambda}\{n^{-1} \parallel \bar{r} - \iota_n\gamma - \overline{C_h}\lambda \parallel^2 + \tau_0 n^{-1} \parallel \lambda \parallel_1\}$$

where $\widehat{C_h} = \widehat{Cov}(r_t, h_t) = T^{-1}\bar{R}\bar{H}^T$. The first step retrieves significant covariances and selects factors in h_t that best explain the cross section of average returns.

Step 2. Next, we run a cross-sectional second LASSO regression for each *j*th factor of g_t of the covariance between returns and *j*th factor of g_t on the covariance between returns and all h_t factors. This second step is required to select factors that may be omitted in the first step.

$$\min_{\xi_j,\chi_{j,\cdot}}\{n^{-1} \parallel \left(\hat{C}_{g,\cdot,j} - \iota_n\xi_j - \hat{C}_{h,\chi_{j,\cdot}^T}\right) \parallel^2 + \tau_j n^{-1} \parallel \chi_{j,\cdot}^T \parallel_1\},$$

Step 3. Then, we run cross-sectional OLS regression using the covariances between the selected factors and asset returns from two-step regression,

$$\left(\widehat{\gamma_0}, \widehat{\lambda_g}, \widehat{\lambda_h}\right) = arg \min_{\gamma_0,\lambda_g,\lambda_h}\{\left\|\bar{r} - \iota_n\gamma_0 - \hat{C}_g\lambda_g - \hat{C}_h\lambda_h\right\|^2 : \lambda_{h,j} = 0, \forall_j \notin \hat{I} = \hat{I}_1 \cup \hat{I}_2\}.$$

5 Results

We implemented Double Selection method alongside two other methods, which are Single Selection method, that runs one cross-sectional LASSO without the second step and controls only for factors selected in the first stage; and Fama-French Five Factor Model as a benchmark model. Our selected tuning parameters were chosen by minimizing the average cross-validation error from 3-fold cross-validation using 100 random seeds.

We started by running the first cross-sectional LASSO regression which selected 3 controls from 98 h_t factors. Then, we run the second cross-sectional LASSO regression and it selected between 3 and 98 $h_t$factors. This aligns with the results of Feng et al. (2020) that single selection method is more conservative and parsimonious compared to the Double Selection method which includes more control factors to ensure omitted variable bias is avoided.

Table 3 reports the results on the marginal explanatory power of each candidate factor proposed between 2016-2020 period, which can be clustered into Quality, Seasonality, Investment, Profit Growth themes according to Jensen et al. (2023). The first columns of each method report the point estimate of SDF loading and the second columns report associated t-statistic for each factor.

Table 3: Testing Factors Introduced between 2016 and 2020

Factors introduced in 2016-2020	Double Selection LASSO		Single Selection LASSO		Fama-French 5 Factor Model	
	λ_s *(bp)*	*tstat*	λ_s *(bp)*	*tstat*	λ_s *(bp)*	*tstat*
Cash-based operating profits-to-lagged book assets (cop_atl1)	-41	-3.06***	-2	-0.22	-4	-3.35***
Market correlation (corr_1260d)	-11	-1.65*	9	1.59	1	0.62
Cash-based operating profits-to-book assets (cop_at)	-60	-2.93***	2	0.18	-4	-2.89***
Mispricing factor: Management (mispricing_mgmt)	15	0.90	10	0.81	10	4.27***
Mispricing factor: Performance (mispricing_perf)	70	3.16***	2	0.23	-3	-3.01***
Cash-based operating profits-to-book assets (ocf_at)	7	0.33	2	0.19	-6	-3. 32***
Change in operating cash flow to assets (ocf_at_chg1)	-19	-3.05***	-20	-3.19***	-31	-5.13***
Operating profits-to- book assets (op_at)	-36	-1.46	0	0.03	-4	-3.36***
Operating profits-to-lagged book assets (op_atl1)	42	1.47	-1	-0.12	-4	-3.56***
Quality minus Junk: Composite (qmj)	-15	-0.94	1	0.08	-4	-3.38***
Quality minus Junk: Growth (qmj_growth)	20	1.85*	-10	-1.66*	-10	-3.96***
Quality minus Junk: Profitability (qmj_prof)	8	0.71	0	0.06	-4	-3.39***
Quality minus Junk: Safety (qmj_safety)	88	3.46***	7	0.71	-2	-1.98**

*Note: Significance levels: *(p < 0.10), **(p < 0.05), ***(p < 0.01)*

As demonstrated in Table 3, Double Selection LASSO identified 7 out of 13 candidate factors as significant at 10%, and 5 of them, comprising of cash based operating profits to lagged book assets, cash based operating profits to book assets, mispricing factor: performance, change in operating cash flow to assets, quality minus junk: safety were statistically significant at 1%. Single Selection LASSO

found only 2 out of 13 to be significant at 10% and change in operating cash flow to assets factor showed significance at 1% level. Fama-French 5 Factor Model found 12 out of 13 factors significant. Results show that most factors are not explained by the 5 Factor Model, and DS LASSO method did much better job than the conventional benchmark model.

Table 4: Robustness to regularization methods of candidate factors

	Double Selection LASSO		**Elastic Net**		**Forward Stepwise**	
Factors introduced in 2016-2020	λ_s *(bp)*	*tstat*	λ_s *(bp)*	*tstat*	λ_s *(bp)*	*tstat*
Cash-based operating profits-to-lagged book assets (cop_atl1)	-41	-3.06***	-39	-2.85***	-46	-3.18***
Market correlation (corr_1260d)	-11	-1.65*	-12	-1.84*	5	0.75
Cash-based operating profits-to-book assets (cop_at)	-60	-2.93***	-1	-0.04	-22	-1.35
Mispricing factor: Management (mispricing_mgmt)	15	0.90	33	2.01**	27	1.38
Mispricing factor: Performance (mispricing_perf)	70	3.16***	65	3.10***	7	0.51
Cash-based operating profits-to-book assets (ocf_at)	7	0.33	-32	-1.64	-57	-3.43***
Change in operating cash flow to assets (ocf_at_chg1)	-19	-3.05***	-26	-4.19***	-33	-3.56***
Operating profits-to- book assets (op_at)	-36	-1.46	-36	-1.46	-45	-3.19***
Operating profits-to-lagged book assets (op_atl1)	42	1.47	18	0.73	-57	-3.41***
Quality minus Junk: Composite (qmj)	-15	-0.94	-12	-0.72	12	0.93
Quality minus Junk: Growth (qmj_growth)	20	1.85*	3	0.27	-13	-1.47
Quality minus Junk: Profitability (qmj_prof)	8	0.71	50	2.22**	5	0.38
Quality minus Junk: Safety (qmj_safety)	88	3.46***	85	3.89***	17	1.15

*Note: Significance levels: *(p < 0.10), **(p < 0.05), ***(p < 0.01)*

Table 4 displays the robustness of g_t factors introduced in 2016-2020 period to several alternative dimension-reduction methods. The first column of the table replicates the results of the Double Selection method that we produced in Table 3 for comparison with other methods. Elastic Net used the combined L1 and L2 penalties of LASSO and Ridge, respectively. Whereas, Forward Stepwise regression (Harvey et al., 2016) penalizes using the Bayesian Information Criterion (BIC). The following 2 factors were highly robust across all 3 different regularization methods: change in operating cash flow to assets (ocf_at_chg1), cash-based operating profits to lagged book assets (cop_atl1), whereas mispricing factor: performance (mispricing_perf) and quality minus junk: safety (qmj_safety) passed the robustness check for both DS LASSO and Elastic Net, cash based operating profits to book assets (cop_at) which showed significance in DS LASSO selection did not appear robust across other two methods.

Overall, our results are in line with Chinese A-shares market which exhibits more information asymmetries and higher investor sentiments compared to developed markets, meaning that cash-based metrics were underestimated, creating exploitable mispricings for informed investors.

6 Conclusions

As a result of the two-step procedure, we were able to pin down and select factors that truly contribute to explain the cross-section of returns. We attempted to reduce the control factors to a parsimonious benchmark against which we evaluated each recently introduced factor. Our results reveal that different dimension-reduction methods produce similar, but not identical results. Overall, our findings show that only a few dominant factors from a set of newly proposed factors contain unique information. Regardless of methods chosen, 2 factors out of 13 g_t factors, namely change in operating cash flow to assets (ocf_at_chg1) and cash-based operating profits to lagged book assets (cop_atl1) showed great significance. The remaining supposedly "new" factor were subsumed by other factors. Therefore, by employing the double-selection LASSO, we were able to compress the zoo of factors for Chinese A-shares market.

References

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies, 81*(2), 608-650. https://doi.org/10.1093/restud/rdt044

Cochrane, J. H., (2011). Presidential address: Discount rates. *The Journal of Finance, 66*(4), 1047-1108. https://doi.org/10.1111/j.1540-6261.2011.01671.x

Fama, E. F., & French, K. R. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance, 47*(2), 427-465. https://doi.org/10.2307/2329112

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics, 33*(1), 3-56. https://doi.org/10.1016/0304-405X(93)90023-5

Fama, E. F., & French, K. R. (1995). Size and Book-to-Market Factors in Earnings and Returns. *The Journal of Finance, 50*(1), 131-155. https://doi.org/10.1111/j.1540-6261.1995.tb05169.x

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics, 116*(1), 1-22. https://doi.org/10.1016/j.jfineco.2014.10.010

Feng, G., Giglio, S., Xiu D. (2020). Taming the Factor Zoo: A Test of New Factors. *The Journal of Finance, 75*(3), 1327-1370. https://doi.org/10.1111/jofi.12883

Green, J., Hand, J. R. M., & Zhang, X. F. (2017). The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns. *The Review of Financial Studies, 30*(12), 4389–4436. https://doi.org/10.1093/rfs/hhx019

Harvey, C.R., Liu, Y., & Zhu, H. (2016). and the Cross-Section of Expected Returns. *The Review of Financial Studies, 29*(1), 5-68. https://doi.org/10.1093/rfs/hhv059

Harvey C.R., Liu, Y. (2021). Lucky factors. *Journal of Financial Economics. 141*, 413-435. https://doi.org/10.1016/j.jfineco.2021.04.014

Jensen, T. I., Kelly, B., & Pedersen, L. H. (2023). Is There a Replication Crisis in Finance? *The Journal of Finance, 78*(5), 2465-2518. https://doi.org/10.1111/jofi.13249

Yang, B., Gao, Q., & Li, J. (2025). Market ambiguity, investor sentiment and market anomalies – Evidence from the Chinese A-share market. *Research in International Business and Finance, 75*, Article 102721. https://doi.org/10.1016/j.ribaf.2024.102721

Appendix

Factor	Acronym	Citation	Publication year
Excess Market Return	MktRf	Jensen, Black and Scholes (1972)	1972
Market Beta	beta_60m	Fama and MacBeth (1973)	1973
Market Equity	market_equity	Banz (1981)	1981
Price per share	prc	Miller and Scholes (1982)	1982
Earnings-to-price	ni_me	Basu (1983)	1983
Book-to-market equity	be_me	Rosenberg, Reid, and Lanstein (1985)	1985
Long-term reversal	ret_60_12	De Bondt and Thaler (1985)	1985
Debt-to-market	debt_me	Bhandari (1988)	1988
Short-term reversal	ret_1_0	Jegadeesh (1990)	1990
Book leverage	at_be	Fama and French (1992)	1992
Small Minus Big	SMB	Fama and French (1993)	1993
High Minus Low	HML	Fama and French (1993)	1993
Price momentum t-12 to t-1	ret_12_1	Jegadeesh and Titman (1993)	1993
Price momentum t-6 to t-1	ret_6_1	Jegadeesh and Titman (1993)	1993
Price momentum t-3 to t-1	ret_3_1	Jegadeesh and Titman (1993)	1993
Price momentum t-9 to t-1	ret_9_1	Jegadeesh and Titman (1993)	1993
Sales Growth (1 year)	sale_gr1	Lakonishok, Shleifer, and Vishny (1994)	1994
Free cash flow-to-price	fcf_me	Lakonishok et al. (1994)	1994
Sales Growth (3 years)	sale_gr3	Lakonishok et al. (1994)	1994
Return on equity	ni_be	Haugen and Baker (1996)	1996
Sales-to-market	sale_me	Barbee Jr, Mukherji, and Raines (1996)	1996
Capital turnover	at_turnover	Haugen and Baker (1996)	1996
Operating accruals	oaccruals_at	Sloan (1996)	1996
Momentum	WML	Carhart (1997)	1997
Change sales minus change Inventory	dsale_dinv	Abarbanell and Bushee (1998)	1998
Ohlson O-score	o_score	Dichev (1998)	1998
Change sales minus change SG&A	dsale_dsga	Abarbanell and Bushee (1998)	1998

Change gross margin minus change sales	dgp_dsale	Abarbanell and Bushee (1998)	1998
Change sales minus change receivables	dsale_drec	Abarbanell and Bushee (1998)	1998
Altman Z-score	z_score	Dichev (1998)	1998
Dollar trading volume	dolvol_126d	Brennan, Chordia, and Subrahmanyam (1998)	1998
Pitroski F-score	f_score	Piotroski (2000)	2000
Coefficient of variation for share turnover	turnover_var_126d	Chordia et al. (2001)	2001
Kaplan-Zingales index	kz_index	Lamont, Polk, and Saaa´-Requejo (2001)	2001
CAPEX growth (1 year)	capx_gr1	Xie (2001)	2001
Coefficient of variation for dollar trading volume	dolvol_var_126d	Chordia, Subrahmanyam, and Anshuman (2001)	2001
Inventory change	inv_gr1a	J. K. Thomas and Zhang (2002)	2002
Amihud Measure	ami_126d	Amihud (2002)	2002
Change in long-term net operating assets	lnoa_gr1a	Fairfield, Whisenant, and Yohn (2003)	2003
Idiosyncratic volatility from the CAPM (252 days)	ivol_capm_252d	Ali, Hwang, and Trombley (2003)	2003
Equity duration	eq_dur	Dechow, Sloan, and Soliman (2004)	2004
Earnings variability	earnings_variability	Francis et al. (2004)	2004
Change in net operating assets	noa_gr1a	Hirshleifer et al. (2004)	2004
Earnings volatility	ni_ivol	Francis et al. (2004)	2004
Earnings persistence	ni_ar1	Francis, LaFond, Olsson, and Schipper (2004)	2004
Current price to high price over last year	prc_highprc_252d	George and Hwang (2004)	2004
Net operating assets	noa_at	Hirshleifer, Hou, Teoh, and Zhang (2004)	2004
Taxable income-to-book income	pi_nix	Lev and Nissim (2004)	2004
Operating cash flow-to-market	ocf_me	Desai, Rajgopal, and Venkatachalam (2004)	2004
Change in current operating liabilities	col_gr1a	Richardson et al. (2005)	2005
Firm age	age	Jiang, Lee, and Zhang (2005)	2005

Change in noncurrent operating liabilities	ncol_gr1a	Richardson et al. (2005)	2005
Change in current operating assets	coa_gr1a	Richardson et al. (2005)	2005
Change in net noncurrent operating assets	nncoa_gr1a	Richardson et al. (2005)	2005
Change in common equity	be_gr1a	Richardson et al. (2005)	2005
Change in long-term investments	lti_gr1a	Richardson et al. (2005)	2005
Change in financial liabilities	fnl_gr1a	Richardson et al. (2005)	2005
Change in short-term investments	sti_gr1a	Richardson et al. (2005)	2005
Change in current operating workng capital	cowc_gr1a	Richardson, Sloan, Soliman, and Tuna (2005)	2005
Total accruals	taccruals_at	Richardson et al. (2005)	2005
Change in net financial assets	nfna_gr1a	Richardson et al. (2005)	2005
Downside beta	betadown_252d	Ang, Chen, and Xing (2006)	2006
Net debt issuance	dbnetis_at	Bradshaw, Richardson, and Sloan (2006)	2006
Number of zero trades with turnover as tiebreaker (12 months)	zero_trades_252d	Liu (2006)	2006
Equity net payout	eqnpo_12m	Daniel and Titman (2006)	2006
Net total issuance	netis_at	Bradshaw et al. (2006)	2006
Net debt-to-price	netdebt_me	Penman, Richardson, and Tuna (2007)	2007
Net payout yield	eqnpo_me	Boudoukh, Michaely, Richardson, and Roberts (2007)	2007
Payout yield	eqpo_me	Boudoukh et al. (2007)	2007
Year 1-lagged return, annual	seas_1_1an	Heston and Sadka (2008)	2008
Asset Growth	at_gr1	Cooper, Gulen, and Schill (2008)	2008
Return on net operating assets	ebit_bev	Soliman (2008)	2008
Year 1-lagged return, nonannual	seas_1_1na	Heston and Sadka (2008)	2008
Years 2-5 lagged returns, nonannual	seas_2_5na	Heston and Sadka (2008)	2008

Net stock issues	chcsho_12m	Pontiff and Woodgate (2008)	2008
Years 2-5 lagged returns, annual	seas_2_5an	Heston and Sadka (2008)	2008
Growth in book debt (3 years)	debt_gr3	Lyandres, Sun, and Zhang (2008)	2008
Change PPE and Inventory	ppeinv_gr1a	Lyandres et al. (2008)	2008
Profit margin	ebit_sale	Soliman (2008)	2008
Assets turnover	sale_bev	Soliman (2008)	2008
Asset tangibility	tangibility	Hahn and Lee (2009)	2009
Quarterly return on assets	niq_at	Balakrishnan, Bartov, and Faurel (2010)	2010
Percent total accruals	taccruals_ni	Hafzalla et al. (2011)	2011
Ebitda-to-market enterprise value	ebitda_mev	Loughran and Wellman (2011)	2011
Operating leverage	opex_at	Novy-Marx (2011)	2011
Percent operating accruals	oaccruals_ni	Hafzalla, Lundholm, and Matthew Van Winkle (2011)	2011
Tax expense surprise	tax_gr1a	J. Thomas and Zhang (2011)	2011
Price momentum t-12 to t-7	ret_12_7	Novy-Marx (2012)	2012
Inventory growth	inv_gr1	Belo and Lin (2012)	2012
Cash-to-assets	cash_at	Palazzo (2012)	2012
Gross profits-to-lagged assets	gp_at1	Novy-Marx (2013)	2013
Frazzini-Pedersen market beta	betabab_1260d	Frazzini and Pedersen (2014)	2014
Liquidity of book assets	aliq_at	Ortiz-Molina and Phillips (2014)	2014
Liquidity of market assets	aliq_mat	Ortiz-Molina and Phillips (2014)	2014
Robust Minus Weak	RMW	Fama and French (2015)	2015
Conservative Minus Aggressive	CMA	Fama and French (2015)	2015
Operating profits-to-lagged book equity	ope_bel1	Fama and French (2015)	2015
Quarterly return on equity	niq_be	Hou, Xue, and Zhang (2015)	2015
Operating profits-to-book assets	op_at	Ball et al. (2016)	2016
Operating profits-to-lagged book assets	op_atl1	Ball et al. (2016)	2016
Cash-based operating profits-to-book assets	cop_at	Ball, Gerakos, Linnainmaa, and Nikolaev (2016)	2016

Cash-based operating profits-to-lagged book assets	cop_atl1	Ball, Gerakos, Linnainmaa, and Nikolaev (2016)	2016
Mispricing factor: Management	mispricing_mgmt	Stambaugh and Yuan (2017)	2017
Mispricing factor: Performance	mispricing_perf	Stambaugh and Yuan (2017)	2017
Operating cash flow to assets	ocf_at	Bouchaud et al. (2019)	2019
Change in operating cash flow to assets	ocf_at_chg1	Bouchaud, Krueger, Landier, and Thesmar (2019)	2019
Quality minus Junk: Composite	qmj	C. S. Asness, Frazzini, and Pedersen (2019)	2019
Quality minus Junk: Safety	qmj_safety	C. S. Asness et al. (2019)	2019
Quality minus Junk: Profitability	qmj_prof	C. S. Asness et al. (2019)	2019
Quality minus Junk: Growth	qmj_growth	C. S. Asness et al. (2019)	2019
Market correlation	corr_1260d	C. Asness, Frazzini, Gormsen, and Pedersen (2020)	2020

Table 5. List of factors with corresponding publication years