Proceedings of ICFTBA 2026 Symposium: Digital Government, Finance, Accountability, and Sustainable Economic Development
DOI: 10.54254/2754-1169/2026.AB38913

LASSO Regression on Asset Pricing Factor Selection

Shuangqi Li

School of Finance, Renmin University of China, Beijing, China
lishuangqi@ruc.edu.cn

Abstract. This paper investigates the problem of redundant factors in asset-pricing models. The excessive inclusion of factors may result in model overspecification and multicollinearity, thereby increasing estimation uncertainty and reducing the accuracy of asset-return predictions. To address this issue, the study proposes a factor-selection procedure based on the least absolute shrinkage and selection operator (LASSO) regression model. The LASSO objective function imposes a penalty on model complexity, discouraging the inclusion of redundant factors and shrinking the coefficients of irrelevant or weakly informative factors toward zero. The proposed procedure was evaluated using monthly data from the Chinese A-share market and successfully selected seven factors from a set of 29 commonly used factors. Nevertheless, in predicting the absolute levels of individual stocks' excess returns for the following month, the model did not significantly outperform the naïve historical-mean benchmark, as indicated by an out-of-sample R-squared value slightly below zero. However, the positive mean rank information coefficient (Rank IC) suggests that the LASSO model retains some ability to identify cross-sectional differences in expected stock returns. Therefore, although its point forecasts remain limited, the model may still provide useful information for stock ranking, portfolio construction, and investment decision-making.

Keywords: Redundant factor, multicollinearity, overspecification, factor selection, LASSO regression

1. Introduction

As the central topic for researchers of asset pricing in the past decades, the factor model has already gained much attention and development as researchers in both academic and industry field contributing new factors to the existing pool throughout the years. Mining for new systematic risk factors which are significantly correlated with the asset returns has already become a topic too well-discussed.

However, with such a tremendous factor bank that previous works have identified, it naturally proposes the question of whether all these known factors are really significant in all cases and the issue of factors selection in real-world quantitative investment practice. Researchers in asset pricing have already pointed out that although hundreds of factors appear to explain the cross-sectional asset returns, the validity of them is still established by a limited set of portfolio returns [1].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.AB38913

Previous researchers have worked on finding a global or versatile risk factors set that could price all kinds of assets including equities, bonds and funds, only to find that the pricing factors of different assets are not perfectly compatible, concluding that each market or asset retains its unique risk factors that determine its return [2]. However, few systematic works have been conducted on formulating a scheme of selecting factors in different investment scenarios so far.

Therefore, now the work of conceiving an effective factors selection framework that can locate an optimal factor subset among the whole factor pool which best matches the specific investment circumstance may be more important than adding new regressors to the model.

This study primarily aims at proposing a factors selection procedure by integrating the least absolute shrinkage and selection operator (LASSO) regression model with asset pricing and test the model's performance in prediction accuracy using a monthly A-stock market dataset spanning from Feb, 2016 to Nov, 2025, which incorporates data of 29 common factors and the target excess return of a specific basket of stocks.

This paper first gives a review of related literature and then clarifies certain theoretical mechanisms including the basic idea of return prediction model and the multicollinearity problem caused by the linear relationship between redundant factors. Afterwards, a factor selection procedure incorporating LASSO regression is proposed and would be tested using the aforementioned A-stock market data, which hopefully would present some key investment insights for individuals and corporations.

2. Literature review

2.1. Factor in asset pricing

Technically, factor in the context of econometrics is the variable or regressor in asset pricing models that explains variations in expected returns across assets by capturing both cross-sectional and time-series dynamics [3]. It usually represents a systematic risk that could influence the prices of all assets within the market and is often regarded as the risks for investors that need to be compensated or priced appropriately in the returns. This compensation, often described as risk premium in asset pricing literature, can be measured by the parameters in common asset pricing models.

2.2. Traditional multi-factor model

To date, numerous classical factor models have been developed. Previous research has evaluated the out-of-sample predictive performance of several widely used multifactor models [4]. These include the three-factor model (FF3), which extends the CAPM by incorporating size and book-to-market factors; the four-factor model, which adds a momentum factor to the FF3 model [5]. The betting-against-beta model, which augments the CAPM with the return spread between low- and high-beta portfolios [6]. Other prominent models include the six-factor model and the q-factor model [7, 8]. The findings indicate that newer models exhibit significantly better in-sample performance than the original CAPM. However, their out-of-sample performance does not demonstrate a statistically significant improvement [4].

2.3. Negative impacts of model misspecification and redundant factors

Regarding this out-of-sample insufficiency of existing models, multiple researches have pointed to factor choices as a potential solution. A previous study highlighted the need to detect and address

DOI: 10.54254/2754-1169/2026.AB38913

model misspecification, arguing that few model specifications would be immune to omitted factors given the growing number of factors identified in the literature [9]. Besides evading missing factors, some has also pointed out the possibility of factor redundancy when factors capturing similar information upon economic risks are included in the same model, leading to multicollinearity and overfitting [10].

2.4. Application of LASSO regression in factor selection

On the topic of identifying redundant factors, several insightful researches have been done, especially in the field of econometrics and statistics. One team constructed a 3-factor model specifically regarding the Chinese stock market, which sheds light on the perspective that factor selection should be specific to the context, while another team employed double-selection LASSO under the framework of stochastic discount factor to identify independent factors carrying incremental information among 100 risk factors in Chinese stock market [11, 12]. Similarly, some have also proposed the Specification LASSO method and applied it in examining the interaction effects between firm sizes and other asset-specific characteristics as well as their explanatory power on cross-sectional returns [13].

Nevertheless, most researches related to LASSO regression's application in asset pricing have not directly responded to the issue of factor selection or did not empirically quantify the capacity improvement of the model before and after pruning redundant factors by LASSO. This gap between these prevailing statistics or machine learning models and the factor reduction need in asset pricing field is exactly what this study aims to bridge.

3. Theoretical mechanisms and model assumptions

3.1. Multi-factor model and return prediction

To get started, the multi-factor model constitutes the basis of this research, which directly originates from the classical CAPM model. The multi-factor model assumes that the return of assets can be predicted by certain risk factors in a linear form, the general formulation of the model is shown in equation:

$$r_{i,t+1} = \alpha_i + \beta_{i,1} F_{1,t+1} + \beta_{i,2} F_{2,t+1} + \ldots + \beta_{i,K} F_{K,t+1} + \varepsilon_{i,t+1}. \quad (1)$$

Specifically, people only used the market factor to conduct the regression before models like FF-3 introduced the company size and book-to-market value factors, since then, more than 200 factors have been found and have been regarded, at least by those who introduced them, as capable of significantly explaining the cross-sectional variation of asset returns [10].

3.2. Mechanisms behind redundant factors

As discussed in the Introduction section, this trend of factor explosion inevitably brings about the risks of incorporating too much variables in the regression, which could affect the accuracy and robustness of the model prediction. If the extra variables are uncorrelated, it would mean overspecification of the model, or if some variables are strongly correlated, it induces multicollinearity, both cases are undesirable. Furthermore, including more variables requires the

DOI: 10.54254/2754-1169/2026.AB38913

estimation of additional parameters and increases computational costs. The resulting complexity may also reduce the model's ability to generalize, leading to what is commonly known as overfitting.

3.3. Mechanisms of LASSO regression

A model which is particularly suited to tackle this issue is the LASSO regression, it has a different optimization objective from the classical OLS minimum squared errors objective as shown in equation :

$$\widehat{\beta} = arg\{\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\}, \tag{2}$$

the parameter λ serves as a penalty parameter, once the sum estimated absolute value of coefficients is large, the objective function would be penalized by a larger second terms. Compared with the classical OLS regression, this penalization terms can efficiently screen out unnecessary variables by rendering some estimations of β_i to be zero.

3.4. Model assumptions

According to the above theoretical discussion, several assumptions should be made:

• H1: The relationship between the asset returns and the risk factors is linear.

• H2: Investors are rational and construct the investment portfolios based on the multi-factor model.

• H3: The investor has full access to all the historical data of common factors and the excess returns for the assets in the portfolio, the investor makes decisions solely based on these data.

• H4: Certain collinearities do exist between different factors, so the investor has to resort to certain factor selection method to mitigate the prediction error caused by redundant factor inclusion.

4. Model construction and application

4.1. Model setup

The basic principle for the following factor selection practice is to maximally avoid two negative effects of excess factors inclusion. First, overspecifying the model by including factors that are irrelevant or only weakly correlated with returns would increase the complexity of the regression and reduce the model's ability to generalize. Second, including too many factors could also lead to multicollinearity. The LASSO regression is specifically designed to address these two issues.

Now assume a representative stock investor who adjusts his portfolio with n stocks at the end of each month based on f factors.

Assume the investor has the historical data of these f factors and wants to predict the excess returns of the n stocks in his portfolio ($r_1, r_2, \ldots r_n$) in the next month. To be clear, here the excess return means the return of a stock minus the risk-free rate in next month.

However, the investor is also rational enough to realize the issues of overspecification and multicollinearity, so he decides to use only a subset of the total f factors in his prediction model, and this is where the LASSO regression comes in, whose objective is specially formulated to penalize redundant factor inclusion and in turn could help the investor appropriately choose one particular subset of factors to put into regression.

DOI: 10.54254/2754-1169/2026.AB38913

4.2. Model construction

To begin with, the input for the prediction model would be the next-month excess returns as dependent variable and the historical stock-month data of the f factors, which would constitute the design matrix of the regression.

Applying the LASSO regression on these sample data, and using certain rules to select the parameter λ of the penalization terms, then a fitted model is obtained including the optimized coefficients for a selected subset of the total f factors after trading off between the in-sample fit of the model and parsimony factor inclusion.

Then the fitted model could be used to conduct the prediction of next-month excess returns given new current data of the selected subset of factor. Related model diagnostics and assessment could also be conducted by calculating root mean squared error (RMSE) or mean absolute error (MAE) measures for the prediction. Furthermore, statistics like the out-of-sample R^2 could be used to compare the performance of the LASSO regression with the historical-mean benchmark prediction.

By applying the process described above to their own portfolios, investors could theoretically generate robust forecasts of next-month excess returns, provided that sufficient factor data have been collected. They could then use these forecasts to inform their investment strategies and portfolio-adjustment plans. Moreover, one can also catch a glimpse of the crucial factors that may strongly affected the next-month excess return of his portfolio by virtue of the factor selection mechanism of the LASSO regression.

4.3. Model application in real world investment scenario

The efficacy of LASSO regression in factor selection and return prediction could be tested using a real-world A-stock dataset.

The data collected are the monthly data of 29 common factors from 11 categories for about 50 individual stocks in the A-stock market, which could be regarded as an investment portfolio of an investor, the investor would like to predict the next-month excess returns using the factors information and the historical next-month excess returns data. The data spans from Feb, 2016 to Nov, 2025 with 6297 entries of data, which provides us with abundant samples for training, validating and testing the LASSO model. The basic information of the 29 factors is summarized in Table 1. The data is prepared from public source and is monthly winsorized, median-imputed, and standardized.

Table 1. Common factors used by the investor

Category	Factors
Size	log_total_assets
Value	book_to_price, earnings_to_price, cashflow_to_price
Momentum	momentum_12_2, return
Volitility	volatility_12m, max_monthly_return_12m, amplitude
Trade	log_turnover_amount, illiquidity_proxy
Profitability	return_on_assets, return_on_equity, net_margin
Growth	revenue_growth, net_profit_growth, equity_growth, asset_growth

DOI: 10.54254/2754-1169/2026.AB38913

Table 1. (continued)

Operation	receivables_turnover, inventory_turnover, total_asset_turnover
Solvency	quick_ratio, cash_ratio, current_ratio, interest_coverage
Leverage	equity_ratio, debt_to_assets
Cash Flow	operating_cashflow_to_sales, operating_cashflow_to_net_income

Figure 1 presents a scatterplot matrix of the four solvency factors, revealing clear linear correlations among factors within the same category. To some extent, this justifies the previously mentioned necessity to prune unnecessary factors and mitigate the negative effects of factor collinearity.

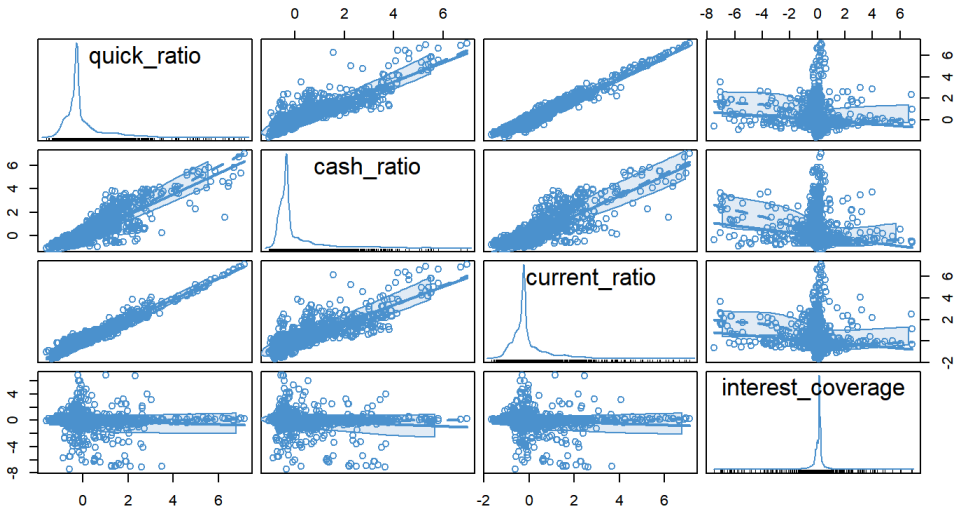


Figure 1. Scatter plot matrix of the four solvency factors

The 6297 entries of data are partitioned into three groups, 3533 for the training dataset, 1407 for the validation set, which is used for the selection of λ in LASSO regression, and 1357 for test set, which is untouched in the training of model and is used to evaluate the prediction accuracy of the resulting factor-selected LASSO regression.

In the result, the model selected 7 factors that could have influenced the next-month excess returns of the stock. Among these factors, illiquidity_proxy, log_turnover_amount and equity_growth have the three highest absolute coefficients. The details of the 7 selected factors and its corresponding coefficients in regression are shown in Table 2.

Table 2. 7 factors selected by the LASSO model and their coefficients

Category	Factor	Coefficient
Trade	illiquidity_proxy	0.008568
Trade	log_turnover_amount	-0.003838
Growth	equity_growth	-0.001720
Momentum	return	0.001676

DOI: 10.54254/2754-1169/2026.AB38913

Table 2. (continued)

Profitability	net_margin	-0.001647
Volatility	amplitude	0.000931
Size	log_total_assets	-0.000324

In the selected 7 factors, the two trade factors both have a large coefficient, indicating that the trade factor groups may play an important role in pricing the stock asset. The illiquidity_proxy is defined as the monthly return of a stock divided by its approximate monthly turnover, at here a one-unit increase of the standardized illiquidity_proxy would lead to a 0.86% absolute increase of the next-month excess return. The log_turnover_amount is defined as the natural logarithm of one plus monthly turnover of the stock, at here a 100% increase in the standardized monthly turnover would approximately cause the next-month excess return to fall 0.0038, namely an absolute decrease of 0.38%.

The interpretations for other factors' coefficients are similar, a 100% absolute increase in the standardized equity growth, current monthly return and net margin would lead to a 0.172%, 0.168% and 0.165% absolute change to the dependent. A one-unit increase of the standardized spread between the highest and lowest stock price of the month divided by last-month closing price would imply a 0.093% absolute increase of the next-month excess return and a 100% increase in the standardized total assets of the company would cause a 0.032% absolute decrease to the next-month excess return.

The RMSE of the model on the test set is 12.47% while the MAE measure is 7.96%, these are both large prediction errors considering usual monthly percentage variation of stock prices. Indeed, the out-of-sample R^2 of the regression is -0.0071, meaning that the prediction performance of the model cannot even catch up with the simple historical-mean forecast, so it is clear that the monthly A-stock return data is severely impacted by market noise and uncertainty, whose exact value could not be accurately predicted or captured using company-level factor data.

However, the conclusion does not mean that the incorporation of LASSO in asset pricing practice is of no value, the mean Rank IC of the LASSO model is 0.0340 and 60.87% of the months have a positive Rank IC. This implies that although the model has limited capacity in predicting the absolute level of stock returns, it still has considerable capacity to recognize the relative level of the cross-sectional stock returns, in other words, the LASSO regression is more suitable to predict whether the excess return of stock A would be higher than that of stock B, instead of accurately predicting the specific next-month return of a single stock.

5. Insights and suggestions

Above all, in this A-stock market example, the LASSO model successfully selected 7 factors that could possibly influence the pricing of the stock out of 29 candidate factors. The selected factors distributed across different categories including trade, momentum, growth, profitability, volatility and size, indicating these fundamentals of stocks could have carried incremental information for next-month excess return prediction in this particular A-stock market scenario. Even though the prediction power of the model still need improvement, the LASSO model has its value in predicting the relative position between different stocks' future returns.

Naturally, the advice for investors who believe in factor model and rely on quantitative model in trading is that they should heed the risk of inaccurate prediction caused by excess factors in model,

DOI: 10.54254/2754-1169/2026.AB38913

deliberately check the correlation between factors in model and actively tackle this issue by relevant factor-selection method, like the one introduced in this paper. However, investors should also note that LASSO regression could not be relied upon in predicting the absolute level of future returns, but it can provide valuable information on the comparative relationship between future returns of different stocks and inform the investor about the essential factors that may hold more importance in predicting the stock returns.

For corporations with trading business, the advice is similar, traders working in corporations should follow the same principle of factor-parsimony in their daily trading practices, intentionally seeking solutions to the redundant factor issue and also be aware of the shortcoming of common statistics models like LASSO regression in accurately predicting the absolute values of future asset returns.

6. Conclusion

In summary, this paper reiterates the redundant factors issue in asset pricing and creatively introduces the LASSO regression into the field of asset pricing to prune the unimportant factors in model. The LASSO model was tested using Chinese A-share market data to predict next-month excess returns and selected seven factors from a total of 29 commonly used factors. Though the model's test set RMSE and MAE are far from satisfying, which indicates LASSO's incompetency in pinpointing the absolute level of the next-month excess returns, the mean Rank IC of the model shows that the regression can still recognize the relative level of different stocks' returns with considerable accuracy.

However, even if the LASSO regression could select a certain subset of factors from a given factor pool for a certain asset portfolio, it does not mean that all these selected factors would definitely be stable and significant factors that could consistently explain the pricing of the asset portfolio, further significant checks of the factors should still be conducted using techniques like the Fama-MacBeth regression.

Moreover, it is still hard to ignore the fact that the prediction power of LASSO on absolute values of returns is even worse than the naive historical mean prediction. Indeed, simply stuffing over 3500 entries of historical stock-month data into the regression model is rather too extensive for real-world practice, several modifications of the prediction process could be added in future, like conducting rolling regressions on a window of 3 months and discarding training data too far back in time to ensure the training and validating datasets both carries up-to-date factors information and avoid the market noise in data too old. These modifications may help improve the predicting power of the LASSO model and should be tried out in later research.

References

[1] Kleibergen, F., & Zhan, Z. (2025). Risk premia from the cross-section of individual assets. *Journal of Econometrics, 252*, Article 106108. https://doi.org/10.1016/j.jeconom.2025.106108

[2] Dang, T. D., Hollstein, F., & Prokopczuk, M. (2026). Factor pricing across asset classes. *Journal of Empirical Finance, 87*, Article 101688. https://doi.org/10.1016/j.jempfin.2026.101688

[3] Zhang, J., Lan, W., Feng, L., & Feng, G. (2026). Testing asset pricing factor models: An out-of-sample perspective. *Journal of Econometrics, 258*, Article 106335. https://doi.org/10.1016/j.jeconom.2026.106335

[4] Kan, R., Wang, X., & Zheng, X. (2024). In-sample and out-of-sample Sharpe ratios of multi-factor asset pricing models. *Journal of Financial Economics, 155*, Article 103837. https://doi.org/10.1016/j.jfineco.2024.103837

[5] Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance, 52*(1), 57–82. https://doi.org/10.1111/j.1540-6261.1997.tb03808.x

DOI: 10.54254/2754-1169/2026.AB38913

[6] Frazzini, A., & Pedersen, L. H. (2014). Betting against beta. *Journal of Financial Economics, 111*(1), 1–25. https://doi.org/10.1016/j.jfineco.2013.10.005
[7] Barillas, F., & Shanken, J. (2018). Comparing asset pricing models. *The Journal of Finance, 73*(2), 715–754. https://doi.org/10.1111/jofi.12607
[8] Hou, K., Xue, C., & Zhang, L. (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies, 28*(3), 650–705. https://doi.org/10.1093/rfs/hhu068
[9] Beaulieu, M. C., Dufour, J. M., & Khalaf, L. (2025). Identification-robust and simultaneous inference in multifactor asset pricing models. *Journal of Econometrics, 248*, Article 105915. https://doi.org/10.1016/j.jeconom.2024.105915
[10] Yu, J. (2026). Taming the factor zoo: A comprehensive review of methodologies for screening factor redundancy in asset pricing models. *Advances in Economics, Management and Political Sciences, 252*(1), 160–165. https://doi.org/10.54254/2754-1169/2026.LD31217
[11] Liu, J., Stambaugh, R. F., & Yuan, Y. (2019). Size and value in China. *Journal of Financial Economics, 134*(1), 48–69. https://doi.org/10.1016/j.jfineco.2019.03.008
[12] Meng, Q., Qi, J., Wang, S., & Zhao, X. (2026). Disciplining the factor zoo: Identifying pricing factors in the Chinese stock market. *Economic Modelling, 155*, Article 107370. https://doi.org/10.1016/j.econmod.2025.107370
[13] Dong, C., & Li, S. (2021). *Specification LASSO and an application in financial markets* (Cambridge Working Papers in Economics No. 2139). Faculty of Economics, University of Cambridge. Retrieved from https://www.econ.cam.ac.uk/sites/default/files/publication-cwpe-pdfs/cwpe2139.pdf