International Science Journal of Management, Economics & Finance
2024; 3(1): 15-23
https://isg-journal.com/isjmef/
doi: 10.46299/j.isjmef.20240301.02
ISSN: 2720-6394

INTERNATIONAL SCIENCE GROUP

What are the determinants of stock returns? A comparison of shrinkage methodologies

Olena Hurina
Department of Economics, Management and Finance, Faculty of Natural Sciences, V.O. Sukhomlynskyi National University of Mykolaiv, Mykolaiv, Ukraine
ORCID: 0000-0002-6315-6067

Natalia Kornieva
Department of Economics, Management and Finance, Faculty of Natural Sciences, V.O. Sukhomlynskyi National University of Mykolaiv, Mykolaiv, Ukraine
ORCID: 0000-0002-7309-8673

Liydmyla Dombrovska
Department of Education, Mykolaiv Institute of Human Development of University «Ukraine», Mykolaiv, Ukraine
ORCID: 0000-0001-5089-950X

Karyna Lisianska
Finance, Banking and Insurance, V.O. Sukhomlynskyi National University of Mykolaiv, Mykolaiv, Ukraine
ORCID: 0009-0007-1192-753X

To cite this article:
Hurina Olena, Kornieva Natalia, Dombrovska Liydmyla, Lisianska Karyna. What are the determinants of stock returns? A comparison of shrinkage methodologies. International Science Journal of Management, Economics & Finance. Vol. 3, No. 1, 2024, pp. 15-23. doi: 10.46299/j.isjmef.20240301.02.
Received: 12 19, 2023; **Accepted:** 01 24, 2024; **Published:** 02 01, 2024

Abstract: One reason why the selected topic is significant is that investing in the stock market is one of the most popular ways for individuals and institutions to grow their wealth. However, determining future stock returns remains a challenging task, as stock prices are influenced by a wide range of factors, such as economic indicators, company financials, and global events. Over the years, researchers have used various methodologies to identify the determinants of stock returns, with the goal of improving investment decisions.

Keywords: stock, stock returns, regression models, return on stocks, interest rates, inflation, economic growth, exchange rates, effect of multicollinearity.

1. Introduction

One reason why the selected topic is significant is that investing in the stock market is one of the most popular ways for individuals and institutions to grow their wealth. However, determining future stock returns remains a challenging task, as stock prices are influenced by a wide range of factors, such as economic indicators, company financials, and global events. Over the years, researchers have used various methodologies to identify the determinants of stock returns, with the goal of improving investment decisions.

2. Object and subject of research

Our study uses a sample of stocks return predictors that are known to influence stock returns. We will then estimate different regression models using the methodologies mentioned above and compare their performance in terms of goodness of fit, variable selection, and prediction accuracy. We will also investigate the robustness of our results.

3. Target of research

The results of our study will provide valuable insights into the determinants of stock returns and the effectiveness of different methodologies in identifying them. These insights can be useful for investors, financial analysts, and policymakers who are interested in improving their understanding of the factors that drive stock returns. Overall, our study aims to contribute to the literature on stock market prediction and empirical finance by providing a comparative analysis of different regression techniques.

Stock returns are one of the most widely studied topics in finance. Understanding the factors that influence stock returns is essential for investors, financial analysts, and policymakers. There are various theories proposed to explain the determinants of stock returns. In the table 1, we will discuss some of the most prominent stock return theories.

Table 1. A brief description of the main directions of theories of return on stocks

Types of stock return theory	Founder, period	The essence of the theory	Criticism of the theory
Efficient Market Hypothesis (EMH)	The hypothesis was formulated by the American economist Eugene Fama in his article for the Journal of Business, published in 1965	The market is efficient with respect to any information if it is immediately and fully reflected in the price of an asset.	There are a few critical theoretical and practical considerations against the EMH: Grossmann-Stiglitz Paradox, Trade Volume Paradox, Volatility Paradox, Market Bubbles
Capital Asset Pricing Model (CAPM)	The model was developed by Jack Traynor (1961, 1962), William Sharp (1964), John Litner (1965a, b) and Jan Mossin (1966) independently in the 60s. The model is based on the portfolio choice theory of Harry Markowitz.	The model is used to determine the required rate of return for an asset that is expected to be added to an already existing well-diversified portfolio, taking into account the market risk of this asset.	The CAPM model is based on the assumption of the existence of risk-free assets. However, these assets are conditionally risk-free. At the same time, the risk of losses when investing in these assets is not taken into account.
Fama-French Three-Factor Model	The model was developed in 1992 by Eugene Fama and Kenneth French to describe stock returns. In 2013, Fama received the Nobel Prize in Economics for his empirical analysis of asset prices.	In asset pricing and portfolio management, the Fama-French three-factor model is a statistical model. The three factors are excess market returns, the superiority of small companies over large ones, and the superiority of high balance sheet/market ratios compared to low balance sheet/market companies.	Several studies have reported that when the Fama–French model is applied to emerging markets the book-to-market factor retains its explanatory ability, but the market value of equity factor performs poorly.

Macroeconomic factors such as interest rates, inflation, economic growth, and exchange rates have been extensively studied as potential determinants of stock returns. The intuition is that changes in these factors affect the overall economy, which in turn affects the performance of companies and their stocks. For example, when interest rates rise, companies may face higher borrowing costs and reduced investment opportunities, which could lead to lower earnings and lower stock prices.

4. Literature analysis

Numerous studies have investigated the relationship between macroeconomic factors and stock returns. For instance, **Fama and French (1989) and Chen et al. (1986)** found that the market risk premium, the difference between the expected return on the market and the risk-free rate, is positively related to stock returns. On the other hand, macroeconomic variables such as inflation and interest rates were found to be less consistently related to stock returns.

As discussed above stock returns are influenced by several factors, such as economic conditions, interest rates, company performance, and market sentiment. However, traditional methods for estimating the relationship between these factors and stock returns suffer from several drawbacks, including multicollinearity, overfitting, and unstable parameter estimates. Shrinkage methodologies have emerged as an alternative approach to address these issues and improve the accuracy and robustness of stock return prediction.

Shrinkage methods are statistical techniques that borrow information across variables to improve the estimation of the relationship between the predictors and the response. In the frequentist framework, shrinkage methods such as ridge regression and Lasso have been widely used to reduce the variance of the estimated coefficients and improve the prediction accuracy. These methods impose a penalty on the size of the coefficients, effectively shrinking them towards zero, which can help to mitigate the effect of multicollinearity and overfitting.

Several studies have compared the performance of frequentist and Bayesian shrinkage methods in predicting stock returns. For example, **Chen et al. (2016)** compared the performance of Bayesian Lasso and Lasso in predicting monthly stock returns for the S&P 500 index. They found that Bayesian Lasso outperformed Lasso in terms of prediction accuracy and model selection, especially when the number of predictors was large. Similarly, **Li and Chen (2017)** compared the performance of Bayesian ridge regression and ridge regression in predicting daily stock returns for Chinese stock markets. They found that Bayesian ridge regression had a higher out-of-sample prediction accuracy and was more robust to data perturbation and model specification.

5. Research methods

In conclusion, shrinkage methodologies have shown promising results in improving the accuracy and robustness of stock return prediction. Both frequentist and Bayesian shrinkage methods have their advantages and limitations, and the choice of method depends on the specific context and objectives of the analysis.

6. Research results

Description of independent variables: The table shows the results of an OLS regression model with the dependent variable returns and 34 independent variables. The table provides information on the estimated coefficients, standard errors, t-statistics, p-values (P>|t|), and the 95% confidence intervals.

A statistically significant p-value (less than 0.05) indicates that there is evidence of a relationship between that independent variable and the dependent variable, holding all other independent variables constant.

Based on the p-values, we can see that *Beta, BetaTailRisk, BM, Cash, High52, IndMom, Size, OPLeverage, Mom12m and STreversal* are statistically significant at the 5% level, indicating that these variables are likely to have a significant impact on returns. The coefficients associated with these variables show the direction and magnitude of the relationship between the independent variable and the dependent variable. *Beta, Size, IndMom, STreversal and High52* have negative coefficients, indicating that they are negatively related to returns. *BetaTailRisk, BM, OPLeverage, Mom12m and Cash* have positive coefficients, indicating a positive relationship with returns. For example,

OPLeverage has a positive coefficient, indicating that higher operating leverage is associated with higher returns. You can see it from the Figure 1 that visualizes the relationship between OPLeverage and the dependent variable ret. Each point on the plot represents an observation in the randomly sampled dataframe. The plot presents 5000 observations. The x-axis shows the values of OPLeverage, and the y-axis shows the values of ret.

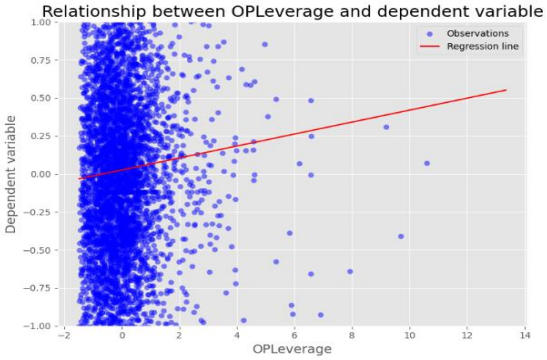

Figure 1. Scatter Plot with Regression Line for OPLeverage and Dependent Variable in an OLS Regression Model.

The variables that are not significant (i.e., have p-values greater than 0.05) are *Accruals, AM, BidAskSpread, EBM, AssetGrowth, NOA, Coskewness, DivInit, EntMult, EP, ExchSwitch, GrLTNOA, Herf, IdioRisk, Illiquidity, IntMom, NetDebtPrice, zerotrade, MaxRet, FirmAge, Price, Mom6m, MomOffSeason*. This means that there is insufficient evidence to conclude that these variables have a significant impact on returns. However, it is possible that they may have an indirect effect on the dependent variable through their correlation with other variables in the model.

The "Confidence Interval" column shows the confidence interval for each coefficient estimate. It represents the range of values within which we can be confident that the true population coefficient lies, with a certain level of confidence. A wider interval indicates more uncertainty in the estimate, while a narrower interval indicates greater precision. In our case, the 95% confidence interval is presented, which means that we can be 95% confident that the true population coefficient lies within the given range. For example, the Betas estimate is -0.0212, with a standard error of 0.004. The t-statistic is -5.917, which has a p-value of 0.000. Therefore, we can conclude that Beta is statistically significant in explaining ret. The 95% confidence interval for the coefficient is [-0.028, -0.014], which means that we can be 95% confident that the true coefficient lies within this range. Similarly, for other coefficients, we can interpret their confidence intervals in a similar way.

Overall, the results of the OLS regression with HC-robust standard errors and Breusch-Pagan test for heteroscedasticity suggest that *Beta, BetaTailRisk, BM, Cash, High52, IndMom, Size, OPLeverage, Mom12m and STreversal* may have a significant effect on stock returns.

Model Validity: The F-statistic is 17.18, with a very low p-value of 2.79e-101, which indicates that the overall model is significant. This result implies that the independent variables, taken together, have a statistically significant relationship with the dependent variable. However, the model's goodness-of-fit is evaluated using the R-squared (uncentered), which is 0.005, indicating that the independent variables explain only 0.5% of the variance in returns. From the Figure 3 it is evident that the predictive power of OLS model is limited. The red line of the Figure 2 in the plot represents perfect correlation between the predicted and actual values, i.e. a hypothetical scenario where the predicted values match the actual values exactly. Ideally, the blue points representing the actual observations should be close to this red line, indicating that the model predictions match the actual observations well, but this is not observed.

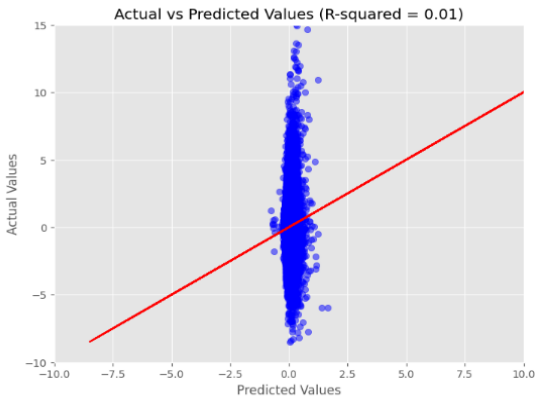

Figure 2. Comparison of Actual and Predicted Values for OLS Regression Model.

The log-likelihood is a measure of the fit of the model. The value of -2.7609e+05 suggests that the model is not a good fit for the data. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are used to compare different models. According to the table, the AIC is 5.522e+05, and the BIC is 5.526e+05. Lower values of AIC and BIC indicate a better model fit, and the values obtained suggest that the model may not be a good fit for the data.

The omnibus statistic is 71438.839, with a probability of 0.000, which means that the residuals are not normally distributed. This means that the residuals are unlikely to be the result of chance alone, and it raises the possibility that the regression model may not be a good fit for the data.

The skewness of the residuals is 1.108, indicating that the model tends to underestimate larger values. The kurtosis is 20.436, indicating that the heavy-tailed distribution of the residuals implies that there are more extreme values than would be expected under a normal distribution.

The Durbin-Watson statistic is a test for autocorrelation in the residuals. The value of 2.000 suggests that there is no significant autocorrelation in the residuals.

Finally, we used the Lagrange multiplier test to test for the presence of omitted variables that are correlated with the included independent variables. The p-value obtained is 0.0, indicating that there may be omitted variables in the model that are significantly correlated with the included independent variables.

In summary, the model may not be a good fit for the data, as indicated by the low R-squared and the values of the AIC and BIC. The normality of residuals is also in question due to the high omnibus statistic, skewness, and kurtosis values. The Lagrange multiplier test also suggests the possibility of omitted variables that are correlated with the included independent variables. As a result, the coefficients of the model may not be entirely reliable. Caution should be taken when interpreting the coefficients of the model.

The dependent variable in this regression is stock returns, and the table displays the estimated coefficients for each independent variable. A credible interval represents the range of values within which the true parameter is likely to lie with a certain degree of probability. To determine the significance of each variable, we look at the 95% credible intervals for each coefficient. If the credible interval does not include zero, then we can conclude that the variable is significant at the 5% level.

Based on the table provided, it appears that *BM, Beta, BetaTailRisk, High52, Cash, STreversal, IndMom, Mom12m and OPLeverage* are significant at the 5% level, since their credible intervals do not include zero. *Similarly, AM, Accurals, AssetGrowth, BidAskSpread, Coskewness, DivInit, DivOmit EBM, EntMult, ExchSwitch, FirmAge, GrLTNOA, Herf, IdioRisk, Illiquidity, NOA, NetDebtPrice, MaxRet Price, Size, zerotrade, IntMom, MaxRet, Mom6m, and MomOffSeason* are not significant at 5% level, since their credible intervals include zero, indicating that they are not significantly different from zero at the 95% level.

The coefficient for BM is 0.05, which means that a one-unit increase in BM results in a 0.05 unit increase in returns. This variable is a measure of the firm's book-to-market ratio and is commonly used as a proxy for value. Figure 3a shows the posterior distribution of the samples for the coefficient

of BM. This plot displays the estimated probability density function for the parameter, indicating the range of likely values for the coefficient and the relative likelihood of each value. As can be seen from the plot, the distribution is centered around a positive value, consistent with the interpretation that value firms tend to have higher returns. Figure 3b shows the changes in the coefficient's estimated value over time, with each line representing a different Markov Chain Monte Carlo (MCMC) sample. This plot also shows that the coefficient for BM convergence to its mean value.

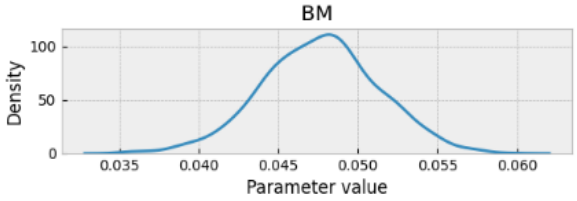

Figure 3a. Posterior Distribution of the Coefficient of BM in a Bayesian Linear Regression Model.

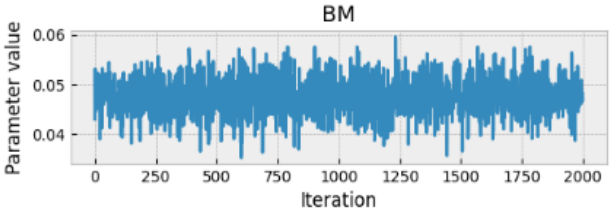

Figure 3b. Trace Plot of the Posterior Distribution for the Coefficient of BM in a Bayesian Linear Regression Model.

To gain a better understanding of the estimated coefficient for Beta, we plot its posterior distribution and trace plot. The posterior distribution plot, shown in Figure 5a, displays the estimated probability density function of Beta based on the observed data. As indicated by the negative coefficient of -0.02, a one-unit increase in Beta results in a 0.02 unit decrease in returns. This variable serves as a measure of the firm's market risk and is commonly used as a proxy for systematic risk. The negative coefficient suggests that firms with higher market risk tend to have lower returns, which is supported by the posterior distribution plot. The trace plot, shown in Figure 5b, further confirms the reliability of the estimated coefficient by displaying the chain of sampled values during the Markov Chain Monte Carlo simulation.

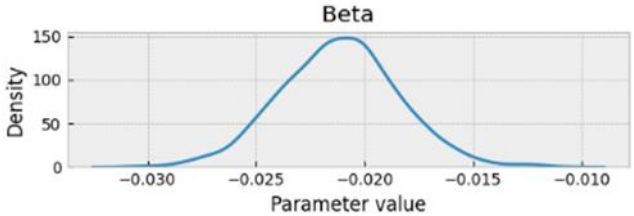

Figure 4a. Posterior Distribution of the Coefficient of Beta in a Bayesian Linear Regression Model.

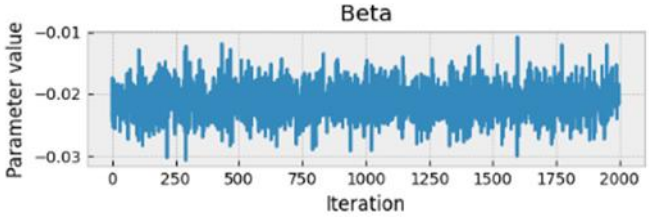

Figure 4b. Trace Plot of the Posterior Distribution for the Coefficient of Beta in a Bayesian Linear Regression Model.

To further investigate the estimated coefficient for *Mom12m,* we present its posterior distribution and trace plot. The posterior distribution plot, depicted in Figure 5a, displays the estimated probability density function of *Mom12m* based on the observed data. As indicated by the positive coefficient of 0.02, a one-unit increase in *Mom12m* results in a 0.02 unit increase in returns. This variable is commonly used as a measure of momentum, or the tendency of stocks that have performed well in the past to continue performing well in the future. The positive coefficient suggests that stocks with positive momentum tend to have higher returns, which is supported by the posterior distribution plot. The trace plot, shown in Figure 5b, further reinforces the reliability of the estimated coefficient by presenting the chain of sampled values during the Markov Chain Monte Carlo simulation.

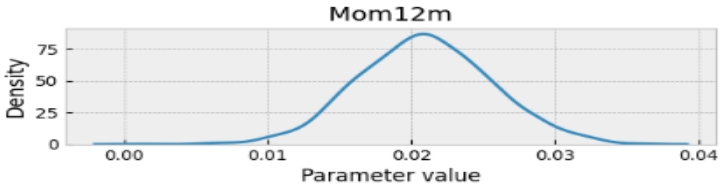

Figure 5a. Posterior Distribution of the Coefficient of Mom12m in a Bayesian Linear Regression Model.

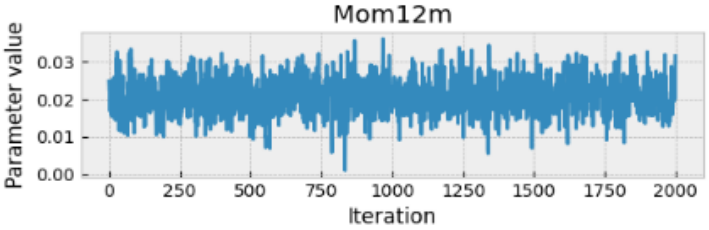

Figure 5b. Trace Plot of the Posterior Distribution for the Coefficient of Mom12m in a Bayesian Linear Regression Model.

Overall, the results of Bayesian OLS suggests that *BM, Beta, BetaTailRisk, High52, Cash, STreversal, IndMom, Mom12m and OPLeverage* have significant effect on stock returns.

Validity of the model: The acceptance probability of 0.92 demonstrates that the Markov Chain Monte Carlo (MCMC) algorithm utilized to estimate the posterior distribution is highly efficient and effective in exploring the parameter space. Moreover, the absence of any divergences confirms that the MCMC algorithm is well-behaved, and the posterior distribution is well-defined.

Furthermore, the R-hat statistic for all variables being 1 indicates that there is no evidence of lack of convergence or mixing in the MCMC algorithm. This finding strengthens the reliability of the model. Additionally, the effective sample size (n_eff) being greater than 980 signifies that the posterior samples contain highly informative data, and the estimates' uncertainty is low. Narrow credible intervals also support this.

However, our analysis of the log posterior predictive density revealed that the model was a poor fit to the data. The obtained result of -276037.5625 indicates that the model cannot be relied upon for accurate predictions. A higher value of the log posterior predictive density indicates a better fit between the model and the data, while a lower value indicates a poorer fit.

Moreover, we investigated sigma, which represents the amount of unexplained variation in the data and can help us assess the model's goodness of fit. As we can see from Figure 6a and Figure 6b the posterior distribution plot and the trace plot demonstrate that the value of sigma is relatively high. This finding indicates that the errors are relatively large, and the model may not fit the data as well.

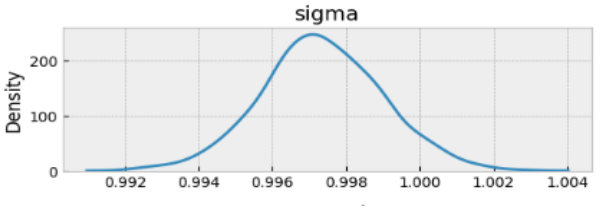

Figure 6a. Posterior Distribution for the Error Standard Deviation in a Bayesian Linear Regression Model.

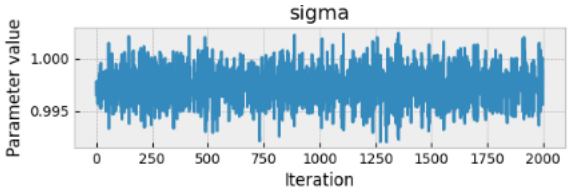

Figure 6b. Trace Plot of the Posterior Distribution for the Error Standard Deviation in a Bayesian Linear Regression Model.

7. Prospects for further research development

Framing interlocking stock returns as a strategic choice based on rational cost benefit considerations, we make some delimiting assumptions that open up prospects for future research.

8. Conclusions

In conclusion, while the MCMC algorithm and the posterior distribution estimate were effective and reliable, the model's poor fit to the data and the high value of sigma suggest that the model may not be the best fit for predicting new data.

This empirical thesis examined the determinants of stock returns by comparing three different shrinkage methodologies: OLS, Bayesian OLS, and Bayesian Lasso. In this thesis we did not only provide an overview of the theoretical foundations of statistical methodologies and statistical approaches, but also demonstrated the practical steps involved in running a regression analysis in Python and defining key parameters. By providing a detailed description of the data preparation process, variable selection, and model evaluation, this study should provide a clear and replicable guideline for future researchers in this field.

Using Classical OLS methodology we identified 10 independent variables, that affect the stock returns, including *Beta, BetaTailRisk, BM, Cash, High52, IndMom, Size, OPLeverage, Mom12m, and STreversal*.

One of the main issues with classical OLS is its reliance on the assumption that the independent variables are uncorrelated and have a constant variance. This assumption is often violated, which leads to biased estimates of the coefficients and inflated standard errors, making it difficult to determine the true relationship between the independent variables and the dependent variable. As a result, the R-squared value of the classical OLS model for stock returns tends to be low, less than 1%, which indicates that the model does not explain much of the variation in the dependent variable.

The results of the Bayesian OLS regression provided evidence that same variables, except Size variable, as the variables which were detected using classical OLS, have a significant effect on stock returns. The validity of the model was supported by a high acceptance probability, the absence of divergences, and an effective sample size greater than 980. However, Bayesian OLS regression also fail to produce reliable estimates of the coefficients relevant to determining stock returns due to high sigma value, which indicates that the model is overfitting to the data.

References:

1) Baker, M., Wurgler, J., & Yuan, Y. (2012). Global, local, and contagious investor sentiment. Journal of Financial Economics, 104(2), 272-287.

2) Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. Journal of Financial Economics, 49(3), 307-343.

3) Barberis, N., & Thaler, R. (2003). A survey of behavioral finance. Handbook of the Economics of Finance, 1, 1053-1128.

4) Bauwens, L. and D. Korobilis (2013). Bayesian methods. in Handbook Of Research Methods And Applications In Empirical Macroeconomics, ed. by N. Hashimzade and M. A. Thornton, Edward Elger Publishing, Handbooks of Research Methods and Applications series, chap. 16, 363–380.

5) Bhattacharya, A., Pati, D., Pillai, N., and Dunson, D. (2016) Dirichlet-Laplace priors for Bayesian Lasso. Biometrika, 103(3), 623-640

6) Chen, Nai-Fu, Roll, Richard and Ross, Stephen A. (1986), Economic Forces and the Stock Market, The Journal of Business, Vol. 59, No. 3, pp. 383-403.

7) Chen, Andrew Y., and Tom Zimmermann (2021). Open-Source Cross-Sectional Asset Pricing, Finance and Economics Discussion Series 2021-037. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2021.037.

8) Fahrmeir, L., T. Kneib, and S. Konrath (2010). Bayesian regularisation in structured additive regression: A unifying perspective on shrinkage, smoothing and predictor selection," Statistics and Computing, 20, 203–219.

9) Fama, Eugene F. and French, Kenneth R. (1989), Business Conditions and Expected Returns on Stocks and Bonds, Journal of Financial Economics, Vol. 25, No. 1, pp. 23-49.

10) 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.

11) Feng, G. and N. G. Polson (2016). Regularizing Bayesian Predictive Regressions," arXiv:1606.01701 [stat].

12) Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis (Vol. 2). Boca Raton, FL: CRC press.

13) Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag.

14) Hou, K., Xue, C., & Zhang, L. (2020). Digesting anomalies: An investment approach. Review of Financial Studies, 33(11), 5118-5164.

15) Hogg, R. V., & Tanis, E. A. (2005). Probability and statistical inference. Pearson Education

16) James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R. Springer

17) Kyung, M., Gill, J., and Ghosh, M. (2010). Penalized regression, standard errors, and Bayesian Lassos. Bayesian Analysis, 5(2), 369-411.

18) Li, X., & Chen, C. (2017). Bayesian Ridge Regression for Stock Return Prediction. Applied Economics Letters, 24(6), 402-407. doi: 10.1080/13504851.2016.1180263

19) Park, T., & Casella, G. (2008). The Bayesian lasso. Journal of the American Statistical Association, 103(482), 681-686.

20) Ross, S. A. (1976). The arbitrage theory of capital asset pricing. Journal of Economic Theory, 13(3), 341-360.

21) Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

22) Wasserman, L. (2013). All of statistics: A concise course in statistical inference. New York: Springer Science & Business Media.