ERASMUS UNIVERSITY ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS
MSc Economics & Business
Specialization Financial Economics

Machine Learning for Dimension Reduction in Factor Zoo

Author: Aysegul Karakas
Student number: 694130
Thesis supervisor: Assistant Professor Amar Soebhag
Second reader: PhD Bart Van Vliet
Finish date: 07/2024

PREFACE AND ACKNOWLEDGEMENTS

This thesis furthered my interest in the finance and economics literature and made me realized how much I have learned in this master's program at ESE. I am deeply grateful to my supervisor, Assistant Professor Amar Soebhag, who introduced me to the exciting world of machine learning with his seminar. His mentorship has encouraged me to think one step ahead and aim higher. I also like to take this opportunity to express my gratitude to my family, whose unconditional support have been my strength throughout this journey.

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

ABSTRACT

This study aims to investigate whether conducting a nested model with the right-hand side (RHS) approach is an accurate way to span the factor zoo or if it is suffering a local optimum problem. If so, the study also investigates which techniques in the latest Machine Learning literature could be better candidates to span factor zoo and can yield closer results to the global optimum. I first ran a full-sample nested model with spanning regressions where I replicated Swade et al. (2023)'s methodology. Compared to other techniques in the literature, this model differentiates itself with its alpha perspective in spanning instead of using a covariance structure like PCA does. To control for the robustness, I run the same model with a rolling window and an expanding window. I observed different selected factors and metrics from the three different nested models which suggests a local optimum in the full-sample nested model's results. I further used a different approach to span the factor zoo, where I regressed the market excess return on the factor zoo. This means that I adopted a left-hand side (LHS) approach and took the market excess return as my LHS dependent variable. I ran several Machine Learning models for dimension reduction: Lasso Regression, PCA, PLS, and PLS with Elastic Nets to see which model presents better accuracy in compressing the factor zoo. According to the results, Lasso Regression demonstrated the highest accuracy, whilst PCA and PLS models presented lower accuracy. The results from these models indicate that implementing regularization increases the accuracy level compared to explaining the covariance matrix. Moreover, implementing Elastic Nets into the PLS model also augmented the model's accuracy; thus, this result also supports the conclusion that regularization improves models' performances. Despite a robust comparison between nested models and other ML techniques couldn't be made due to their different regression structures, the inaccuracy between three different nested models' selected factors and metrics suggests that we should look for further methods and utilize the innovative novelties of Machine Learning such as regularization.

Keywords: Factor Zoo, Dimension Reduction, Nested Model, Lasso Regression, PLS, PCA, Elastic Nets, Machine Learning, GRS Statistic, Regularization

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 Introduction

The world of finance offers hundreds of factors to investors. While the abundance of factors provides various diversification opportunities, many factors likely exhibit overlapping features. Managing such a vast array of factors can also lead to investor distraction, a significant topic in behavioural finance. Therefore, the finance literature must identify a minimal set of factors that effectively summarize all factors in the Factor Zoo. As Feng et al. (2017) suggested, there is a prevailing belief that true asset pricing models are inherently low-dimensional. To tackle the high-dimensionality problem in Factor Zoo, various literatures took place in the world of finance. The study by Swade et al. (2023) appears to be a notable paper addressing this issue and the authors propose a nested model with spanning regressions for dimension reduction. The results claim that the selected 15 factors capture the entire factor alphas in the Factor Zoo. This nested model adopts a right-hand side methodology which was previously proposed by Barillas and Shanken (2016). Furthermore, the nested model which spans the factor zoo within an alpha perspective, was adopted by Fama & French (2018) too. The key component of this approach is using Gibbons, Ross, and Shanken (hereafter, GRS) test statistics to determine the best factor model. GRS, developed by Gibbons, Ross, and Shanken (1989), is a popular metric for this kind of analysis as it is used in testing the hypothesis which implies that the intercept, thus the alpha, in each model is equal to zero (Gibbons, Ross, and Shanken, 1989). However, this methodology, similar to 'Forward Selection' in ML, can stuck in a local optimum and pick a different path that may significantly affect the robustness of the results. This means the model might include variables with significant positive incremental impacts locally, while excluding combinations of variables with greater global explanatory power. Hence, this research aims to assess the accuracy of the nested model approach and explore which alternative dimension reduction techniques in Machine Learning can yield more accurate models. For robustness checks in nested models, three different approaches are adopted: a full-sample model, a rolling window model, and an expanding window model. In analysing the nested models, I observed notable deviations in the selected factors between the full sample, rolling window, and expanding window models. Specifically, the number of factors required to span the factor zoo increased when using rolling and expanding windows. Especially, the rolling window nested model required significantly more factors to minimize the number of factors explaining the factor zoo. These observations suggest that the full-sample nested model may be omitting some significant factors present in the factor zoo. To compare Machine Learning techniques, I applied Lasso Regression, one of the most popular regularization methods, as well as PCA and PLS, to evaluate the impact of using the covariance structure of factors alongside the alpha perspective and regularization. Additionally, I ran a PLS model incorporating Elastic Nets to combine covariance structure and regularization. Among all ML techniques, Lasso presented highest robustness.

CHAPTER 2 Theoretical Framework

2.1 Factor Zoo

In the first section of the theoretical framework, I review the existing literature on factor investing. With the foundation of the Capital Asset Pricing Model (CAPM), Sharpe (1964) introduced the underlying methodology of Factor Investing. Factor investing enables investors to obtain higher expected returns by bearing an additional amount of risk on their investments. In other words, it means choosing the price of risk over the price of time. This revolutionary model of the finance world was first improved by Fama & French (1993, hereafter F&F) with the introduction of the F&F 3-Factor Model where the three factors are: Size (small minus big), Value (high minus low), and the Market Risk (Portfolio Return minus the Risk-Free rate return). Furthermore, extended to the F&F 5-Factor Model by including Profitability (excess returns of stocks with high profitability minus those with low profitability) and Investment (excess returns of companies with conservative investment policies over those with aggressive investment policies) themes (F&F, 2015). Following these initiative principles on factor investing, hundreds of other risk factors have been provided in the asset pricing literature, for instance, the database used in this paper contains 153 different factors from 13 themes (Jensen, Kelly, and Pedersen (2023, hereafter JKP)). This is where the high dimensionality problem starts. In this sense, implying dimension reduction possesses a high importance as it shortens computation times while simultaneously increasing classification accuracy by eliminating redundant features (Somol et al., 2004). The following sections will review machine learning techniques for dimension reduction and examine the existing literature on their application in the context of the Factor Zoo.

2.2 Dimension Reduction for Factor Zoo

To tackle the high-dimensionality problem in Factor Zoo, various literatures took place in the world of finance. Harvey et al. (2015) evaluate the statistical significance of a new factor through a benchmarking technique claiming that a new factor must overcome a t-statistic larger than 3.0. Another literature which is highly related to my paper is also about dimension reduction in the Factor Zoo: to distinguish efficient factors among the redundant ones, Feng et al. (2017) developed a model-selection method where they evaluate the explanatory power of any new factor compared to the previously existing ones. By eliminating the redundant ones, Feng et al. (2017) spanned a factor zoo containing 250 factors to 99 factors; however, they highlight the low probability of perfect variable selection. This is due to the possible erroneous inferences resulting from the locally insensitive nature of high-dimensional nuisance parameter estimation to minor inaccuracies (Chernozhukov et al., 2015).

While trying to identify the minimum number of factors explaining factor zoo, there are different determinants that can be used such as covariance structure or factor alpha. If we ask the question “how

do asset prices move together?", then we need to do a Principal Component Analysis (PCA) or Partial Least Squares (PLS) analysis demonstrating the covariance matrix. As explained by James et al. (2023) in their book "An Introduction to Statistical Learning: With Applications in Python.", PCA demonstrates the low-dimensional equivalent of the observations that account for a significant portion of the variance. James et al. (2023) also describe PLS, a dimension reduction technique, that detects linear combinations of the original variables, creates a new set, and finally fits these new features into a linear model by using their least squares. Both PCA and PLS take advantage of the covariance structure of the data; however, whereas PCA maximizes the variance among only the predictors, PLS optimizes the covariance between predictors and responses, not just predictors. An example of a recent paper using PCA, Bessembinder et al. (2021), assess the out-of-sample Sharpe ratios of portfolios using the principal components and authors show how additional significant factors are not-redundant by demonstrating their relevance to economic conditions. Giglio et Xiu (2021) also use PCA by using a risk-premium approach. Their paper suggests that knowing all true factors' identities isn't necessary to estimate one's risk premium, as long as we can reconstruct the entire factor space. On the other hand, Swade et al. don't explain the covariance structure of factor returns in their nested model. So, methods such as PCA and PLS are not aligned with their method. Swade et al. explain factors' contribution to the model via an alpha perspective. What we need to ask to be aligned with Swade et al.'s methodology is "How much does a new factor contribute to reducing the alpha in the augmented model compared to the existing models?". In conclusion, the nested model methodology lies in determining the key alpha contributors instead of explaining covariances. This paper will cover different Machine Learning techniques to explain both perspectives. Furthermore, regularization methods such as Lasso Regression and Elastic Nets will be covered too.

One example from the latest literature is about identifying the factors that are most relevant for explaining the cross-section of expected returns by using the Reduced-Rank Approach (RRA) (He et al., 2023). According to their results, the RRA 5-Factor model performs better than many popular models such as the F&F 5-Factor model, principal component analysis models, least absolute shrinkage models, partial least squares models, and selection operator models; however, even though a shrinkage restriction was used to avoid overfitting, pricing errors couldn't be avoided.

Another recent literature introduces a novel solution: Lettau (2023) uses a tensor factor model (TFM), an extension of principal component analysis (PCA), where assets are sorted by TFM characteristic factors rather than individual attributes. Instead of using dimension-reduction techniques on a wide range of portfolio returns derived from sorting stocks by various characteristics. Lettau (2023) constructs factors that encapsulate the information within these characteristics across assets. Then, he groups factors into portfolios based on these characteristic factors, rather than in the return space.

Even though classical econometrics techniques provide various dimension reduction techniques to be used for Factor Zoo, recent literature was enriched by certain Machine Learning techniques. One of the most common ML techniques providing dimension reduction opportunity is LASSO Regression which minimizes the residual sum of squares while enforcing a constraint that the total absolute value of the coefficients remains below a certain limit which ensures some coefficients are zero. (Tibshirani, 1996). Feng et al. (2020) adopt Lasso Regression but also account for model selection mistakes, so the research goes beyond just assuming a perfect model selection. This means that either by minimizing the bias in the model caused by the omitted variable or by explaining the cross-section of expected returns, they don't assume a perfect model. To avoid inaccuracy occurring from a non-perfect model, they use the double-selection LASSO by Belloni et al. (2014b) together with two-pass regressions like Fama-Macbeth. Feng et al. (2020) test the marginal contribution of a factor, beyond the explanatory power of another group of factors. For instance, they examined that the seasonality factor of Heston et Sadka (2008) which has a high correlation to the momentum factor. It is quite interesting how they demonstrated that it has a significant alpha against the F&F 3-factor model, while it has a quite low and insignificant alpha against a model including the momentum factor (Feng et al., 2020).

Another highly cited paper, Kozak et al. (2020), converts LASSO into an elastic nets technique by incorporating an additional penalty to the total of Stochastic Discount Factor (SDF) coefficients by combining L1 to L2 penalty. It follows the elastic net methodology where the cross-sectional out-of-sample R2 is maximized. The results of Kozak et al. (2020) demonstrate that a small number of characteristics-sparse SDF components are not effective in summarizing the cross-section of expected stock returns. However, if the SDF is formed using a small number of principal components, the model performs well. What makes this paper different is that it estimates the Stochastic Discount Factor (SDF) by using characteristic-sorted factors. Considering one of the SDF theories which implies that SDF depends on all available information, and thus on a high-dimensional set of variables (Chen et al., 2024), it can be said that this approach may be providing insightful explanations in factor zoo. This thesis study will also adopt elastic nets in one of the tested models, please see Section 4 for methods and 5 for results.

A very recent paper, by Bryzgalova et al. (2023a), outperforms the previously mentioned model of Kozak et al. (2020) in characterizing SDF by implementing a Bayesian Model Averaging (BMA). Bryzgalova et al. (2023a) also provide a further object of interest such as alphas, R2s, Sharpe ratios, confidence intervals, etc. Under their methodology, firstly, a very basic Bayesian estimator was developed by the authors for linear SDFs. With this method, weak factors were effectively identified in a finite sample and reasonable conclusions were drawn about the price of risk associated with the dominant factors, cross-sectional fit metrics, and other relevant factors. For factor selection,

Bryzgalova et al. (2023a) do not depend on marginal likelihoods; instead, they use suitable priors that are uninformative rather than flat priors.

It was inevitable that the latest advancements in machine learning, such as neural networks, would be applied to asset pricing to deal with complex functional dependencies (high dimensionality). Chen et al. (2024) use deep neural networks to identify primary factors driving asset prices as this method appears to be strong in modeling complex and non-linear relationships within large datasets. By implementing a no-arbitrage pricing theory condition in building test assets, Chen et al. (2024) improve the risk premium signal and end up with a higher signal-to-noise ratio which improves the accuracy of Machine Learning methods.

2.3 How to evaluate alpha: LHS or RHS Approach?

The evaluation of various models of literature has two prominent approaches to assess the contribution of potential factors. The first one, the Left-Hand Side Approach (LHS), assesses the model by examining the intercept (alpha) in time-series regressions, where significant alpha means the model does not fully capture the portfolio's returns as the alpha represents the portion of the portfolio's returns not explained by the factors (Fama & French, 2018). Here we ask, "How well do individual assets' returns align with the expected performance based on risk factor models?" In other words, the dependent variables of the LHS portfolios are the ones researchers are attempting to explain. An example is the stock portfolios which can contain sorting characteristics such as size, book-to-market ratio, profitability, investment, cash profitability, momentum, and anomalies (Fama & French, 2018). In this paper when I implement Lasso Regression, PCA, PLS, and PLS with Elastic Nets, my LHS dependent variable is Market Excess Return. On the other hand, the second approach is the Right-Hand Side Approach (RHS) which is built on spanning regressions for nested models where it is possible to observe how well a factor model is performing in pricing the other factors not included in the model (Swade et al., 2023). Here we ask, “Does the new factor provide additional explanatory power beyond the existing factors in the model?”. Hence, it means evaluating potential (non-nested) factor models by regressing new candidate factors against existing model factors to see if they provide additional information. LHS and RHS approaches will be explained in detail as part of the methodology in Section 4: Method.

2.4 Overcoming the Issue of Local Optimum

It is crucial to consider that a nested model may potentially be stuck with a local optimum. This means that the model could include variables that have a large positive incremental impact locally, but it may exclude combinations of variables that might have greater explanatory power on a global scale (Addis et al., 2005). This section investigates the literature on this specific topic to suggest several improvements. As the nested model method is aligned with the forward selection method from ML, I address this issue from a forward selection angle. For instance, Chotchantarakun (2023) recommends using a random algorithm during forward selection to avoid local optima. However, it is essential to ensure that the contribution of features does not vary randomly across distinct subgroups, as this could negatively impact the prediction mechanism with unstable and noisy observation patterns (Somol, 2004). According to Chotchantarakun (2023), Genetic Algorithms effectively avoid suboptimal solutions through crossover operations. Consequently, Chotchantarakun (2023) proposes a novel method called Forward Selection with Genetic Algorithms (FS-GA), which enhances sequential forward feature selection by combining it with Genetic Algorithms and the weak feature replacement technique.

Overall, the nested model approach (a type of forward selection) used by Swade et al. (2023) may suffer from a local optimum problem. This approach aims to build a factor model with the minimum number of factors needed to explain stock returns in the factor zoo, evaluating each factor incrementally to determine if it adds significant explanatory power to the existing model. However, once the most significant factor is included, subsequent factors are regressed against the updated model, which differ from the initial model. This process may miss important combinations of factors that provide a better overall fit, akin to the interaction terms discussed by Rosnow and Rosenthal (1989).

To address these issues, the latest literature offers improved techniques, such as those mentioned previously. Alternatively, using another Machine Learning technique instead of the nested model could help avoid local optima. To explore better methodologies for identifying the minimum number of factors explaining the factor zoo, this paper will implement not only the full-sample nested model but also the Rolling Window Nested Model, Expanding Window Nested Model, Lasso Regression, PCA, PLS, and PLS with Elastic Nets.

CHAPTER 3 Data

This research aims to replicate the nested model approach of Swade et al. and investigate the accuracy of this methodology. Furthermore, different dimension reduction techniques in Machine Learning will be compared to see if there are more accurate models compared to the previous methodology. The dataset was delicately gathered and analyzed to make sure it was aligned with the data used by Swade et al. Therefore, the factor data is collected from the same source, global factor data from (JKP, 2023). The construction of the JKP database comprises 153 factors and 13 themes. For the U.S. dataset, the monthly excess returns of U.S. capped value-weighted factors with a one-month holding period are gathered. According to JKP construction, for capped value-weighted factors, stocks are grouped into three categories each month, and the returns are computed within a limit at (capped at) the 80th percentile of the NYSE market equity weighted portfolios (Jensen et al. 2023). Soebhag et al. (2023) demonstrated how different weighting schemes impact the overall outcomes in a model. Therefore, to develop tradable and balanced portfolios, the structure selected in this study makes sure that small stocks will be able to keep their small weights and won't be dominated by any other big stocks within the portfolio (Jensen et al. 2023). The sample period starts from November 1971 and ends in December 2021. The factor return definition by JKP is highlighted as high- minus low-tercile return and all returns are excess returns in USD.

Figure 2 shows the number of factors per theme. The data construction for the factors is perfectly aligned with Swade et al.'s scheme. The 153 factors are categorised by 13 themes which were previously defined by JKP by using the Ward Hierarchical Clustering method of Murtagh and Legendre (2014). As can be seen from Figure 1, we have 6 factors in the Accruals theme, 7 in Debt Issuance, 22 in Investment, 11 in Low Leverage, 18 in Low Risk, 8 in Momentum, 12 in Profit Growth, 12 in Profitability, 17 in Quality, 11 in Seasonality, 6 in Short-Term Reversal, 5 in Size, and finally 18 in Value.

Figure 1: Number of Factors by Theme

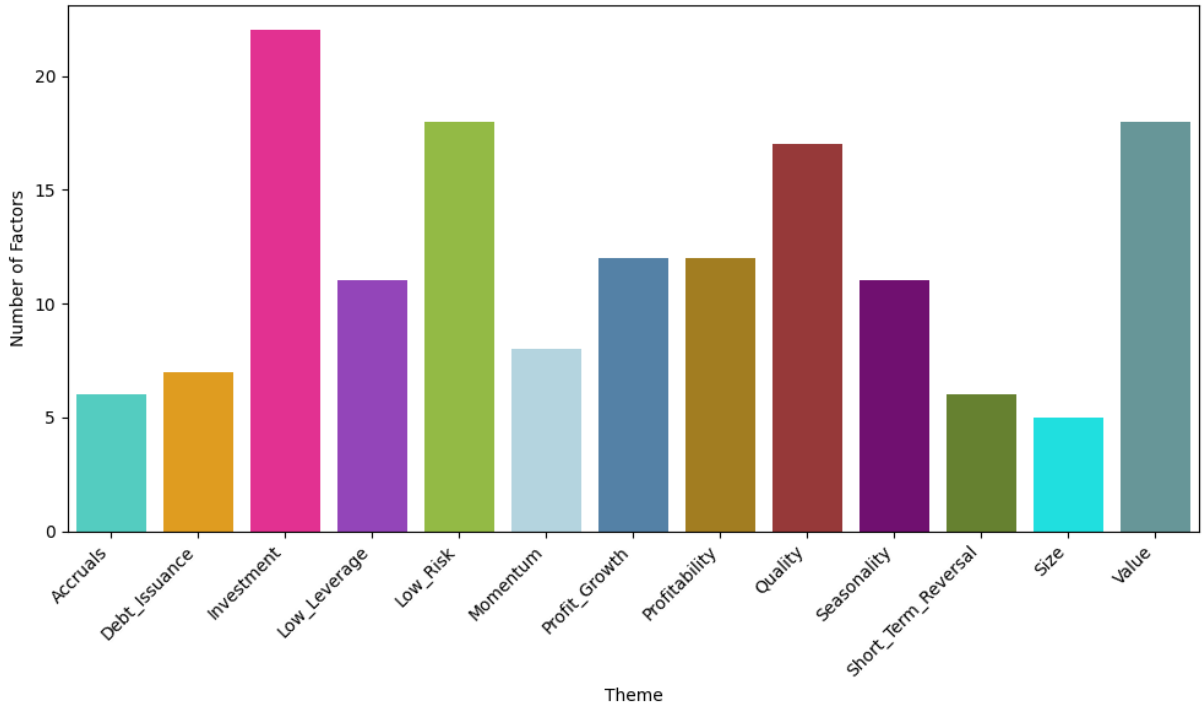


Figure 2 displays the average annualized return per each factor over the time period we have. The highest return was observed from the Operating cash flow-to-market factor (ocf_me): 7.38%, followed by the Price momentum t-12 factor (ret_12_7): 7.16% and Ebitda-to-Market Enterprise Value factor (ebitda_mev): 7.11%. Meanwhile, the lowest average returns were observed from Liquidity Scaled by Lagged Market Assets (aliq_mat), 21 Day Bid-Ask High-Low (bidaskhl_21d), and Firm Age (age) factors: -3.77%, -2.84%, and -2.43%, respectively. All numbers are rounded to 2 digits and a detailed description of all factors and themes can be accessed from the documentation by JKP (Jensen et al., 2021).

Figure 2: Average Annualized Return per Factor

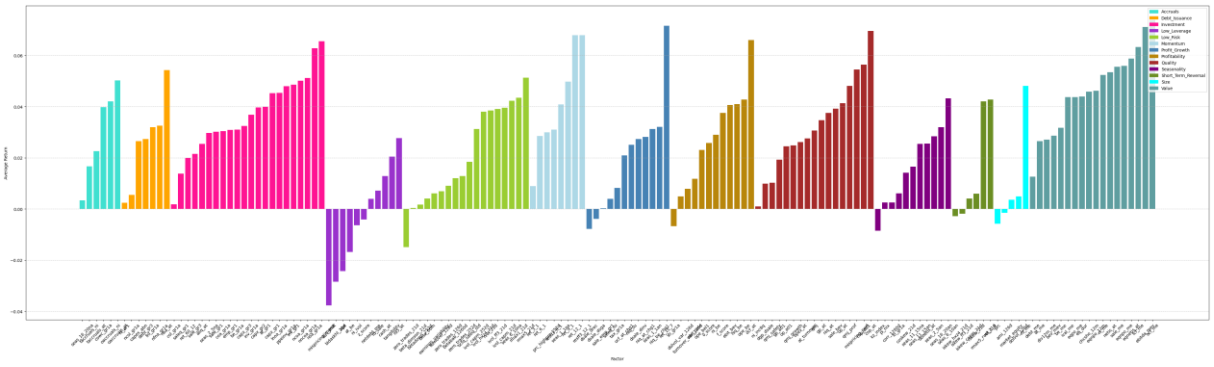
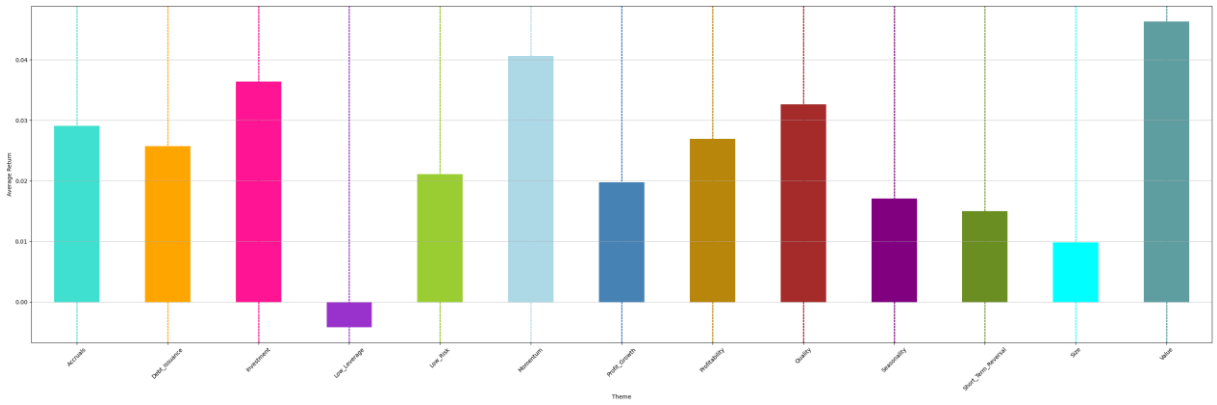


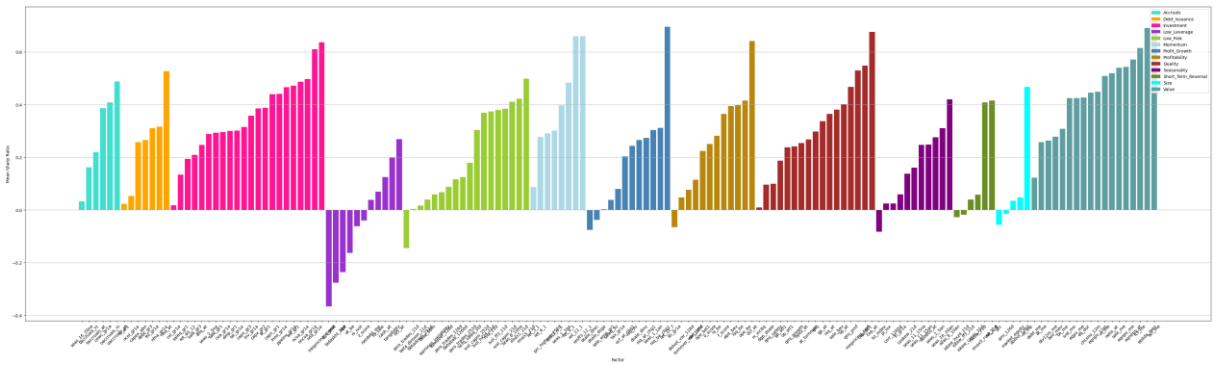
Figure 3 presents average annualized return per theme. Among 13 themes, Value is the theme presenting the best performance with an average return of 4.63% followed by Momentum and Investments themes (4.06% and 3.64%, respectively). On the other hand, Low Leverage is the only theme loser demonstaring an average return of -0.042%. Meanwhile, the second lowest average return was observed from Size theme: 0.098%.

Figure 3: Average Annualized Return per Theme



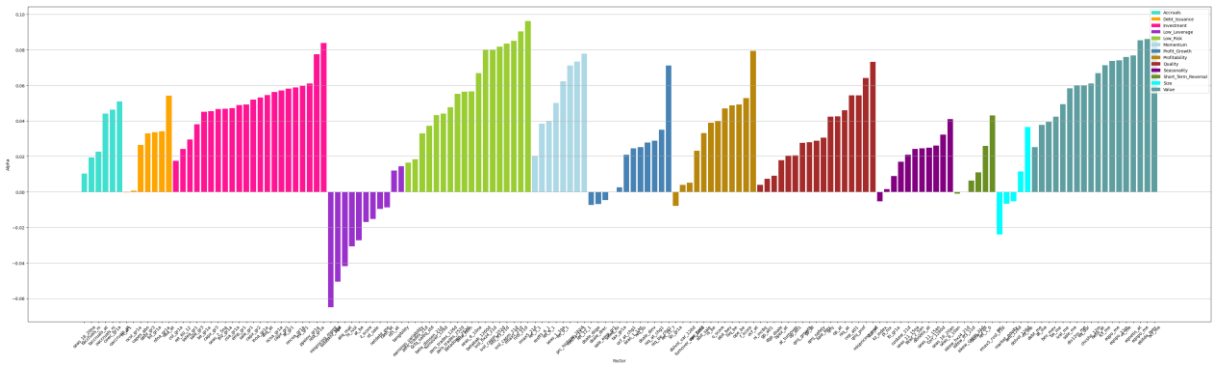
Average sharpe ratios per factor are displayed in Figure 4. While the Operating Cash Flow-to-Market (ocf_me) factor presents the highest sharpe ratio with a value of 71.69%, Price Momentum t-12 (ret_12_7) and Ebitda-to-Market Enterprise Value (ebitda_mev) factors have slightly lower rates and demonstrate the second and third highest ratios, (69.61% and 69.05%, respectively). Conversely, the lowest sharpe ratio was observed from Liquidity Scaled by Lagged Market Assets (aliq_mat) factor: -36.64%, and the second lowest sharpe ratio was from 21 Day Bid-Ask High-Low (bidaskhl_21d) factor: -27.64%.

Figure 4: Average Annualized Sharpe Ratio by Factor



As the goal of the nested model is identifying the few factors explaining all factor alphas, understanding factor alphas plays a crucial role. The alpha calculation methodology of Swade et al.'s was followed. To achieve this, the data for Market Excess Returns was collected from Kenneth R. French - Data Library (2023). As the dataset of Kenneth R. French expresses the data for Risk-free rate and Market Excess Return in percentage terms, the data collected from this source was divided by 100 before the data processing. To calculate alphas; firstly, betas are calculated by running a Linear Regression model where the monthly excess factor returns are regressed against monthly excess market returns, aligned with CAPM regression methodology, by using the scikit-learn library in Python. Secondly, alphas are calculated by subtracting the expected returns, determined by previously calculated betas times the market's excess returns, from the actual returns. To annualize alpha values, previously calculated alphas were multiplied by 12. Thirdly, all betas and alphas were assigned to the corresponding rows in the final data frame by matching with the relevant factors. Finally, average annualized alphas per factor were displayed in Figure 5. By presenting 9 negative factor alphas, the low-leverage theme is seen as the worst-performing theme. The lowest alpha value was observed from 21 Day Bid-Ask High-Low (bidaskhl_21d) factor with an alpha of -6.50% followed by Firm Age and Liquidity Scaled by Lagged Market Assets factors: -5.05% and -4.19%, respectively. On the other hand, the two highest alpha values are in the Low-Risk theme: The first one is "the highest 5 days of Return" factor (rmax5_21d) with an alpha of 9.62% and the second one is "return Volatility" factor (rvol_21d) with an alpha of 9.04%. The third highest alpha value is from the Value theme: "operating Cash Flow-to-Market" (ocf_me) factor, presenting an alpha of 8.78%. Overall, the average annualized alpha across all factors is calculated to be 3.53%. This is significantly aligned with the value determined by Swade et al., which is 3.51%, indicating strong consistency between our data sets.

Figure 5: Average Annualized Factor Alphas



CHAPTER 4 Method

4.1 Left-Hand-Side (LHS) or Right-Hand-Side Approach (RHS)?

As mentioned in Chapter 2, Theoretical Framework, there are two prominent approaches to assess the contribution of a potential factor: 1) the Left-Hand Side Approach (LHS) and 2) the Right-Hand Side Approach (RHS).

1) The Left-Hand Side (LHS) approach involves examining the intercept in time-series regressions, where a significant alpha indicates that the model does not fully capture the portfolio's excess returns (Fama & French, 2018). Stock portfolios used as LHS-dependent variables, which can include sorting characteristics, are: size, book-to-market ratio, profitability, investment, cash profitability, momentum, and anomalies (Fama & French, 2018). In this paper, I use market excess return as the LHS dependent variable and regress it on the existing factors in the sample within a single regression. For this analysis, I will utilize Lasso regression, PCA, PLS, and PLS with elastic nets. See the equation below where Fm is the factor of market excess return, αi is the intercept (alpha) for portfolio i, βFk is the factor loading for factor k, and ϵi is the error term for portfolio i.

$$Fm = \alpha i + \beta 1.F1 + \beta 2.F2 + ... + \beta k.Fk + ei \quad (1)$$

Swade et al. (2023), inspired by the findings of Barillas and Shanken (2016) which do not use the LHS approach due to its high dependency on the test assets used to gather excess returns and its limitation to only observe the contribution of factors included in the model. It is argued by Barillas and Shanken (2016) that interpreting traded factors is more efficient for pricing returns; thus, they recommend a right-hand side (RHS) approach. This methodology has been adopted by Swade et al. (2023) and Fama & French (2018) in developing their respective models.

2) RHS approach built on spanning regressions for nested models where it is possible to observe how well a factor model is performing in pricing the other factors not included in the model and it enables observing each factor's contribution step by step. See the equation below where Fnew is a new candidate factor being tested, αi is the intercept of the regression, βk is the regression coefficient of Fk, Fk is the existing factor k from previous iteration, and ϵ is the error term. Under this methodology, by using GRS statistics, the new factor's contribution to minimizing the model's alpha is being tested. Among all tested factors, the best-performing factor gets to be included in the permanent model and

the remaining factors will be left out to be tested in the next iterations till all remaining alphas are significantly equal to zero. The model conducted with the RHS approach is called a "Nested Model".

$$F\ new = \alpha i + \beta m.Fm + \beta 1.F1 + \beta 2.F2 + ... + \beta k.Fk + ei \quad (2)$$

As can be seen from equation 2, the RHS approach nested model seems to be a type of forward selection method, but it differs with its other characteristics. If we consider a basic forward selection, not the nested model, it is a method where the first step is a null model that does not contain any predictors but an intercept. In the following steps, while fitting linear regressions, the variable with the lowest residual sum of squares (RSS) would be included in the model (James et al., 2023). However, the Nested Model methodology compares each augmented model by their GRS test statistics to determine the best model. GRS, developed by Gibbons, Ross, et Shanken (1989), appears to be a popular metric for this kind of analysis as it is used in testing the hypothesis which implies that the intercept, thus the alpha, in each model is equal to zero (Gibbons, Ross, et Shanken, 1989). As mentioned by Kamstra and Shi (2020), in the empirical finance world, it is a golden standard for asset pricing. Please see section 4.2.2 for model evaluation with GRS. To conclude, Swade et al. (2023) constructed a nested model just like forward selection and they adopted the RHS approach. To span the factor zoo, they begin with the CAPM as the initial model and then they progressively apply spanning regressions. In this research, as my first method, I will be using the same methodology where the results will be presented in Section 5.

4.2 Building a Nested Model with RHS Approach

For Step 1, firstly, the iteration counter *l* is set to 0 (Set *l*: = 0) and the CAPM Regression is taken as the initial model where Fm is the excess market return, fi is the excess return of portfolio i (the factor being tested in the factor zoo) and N is the size of the factor zoo (Swade end al., 2023):

$$Fi = \alpha i + \beta m.Fm + ei \qquad i=1,...,N \quad (3)$$

For Step 2, the following spanning regressions will be built upon CAPM, and N - *l* different models will be augmented in each iteration where each model will be testing one of the previously excluded factors from the model (*l* also equals to number of factors already added to the model from previous iterations). Each newly included factor will be labeled as f selected and their coefficient will be labeled as β selected.

To be permanently added into the current model, a tested factor must be indicating better explanatory power than the other tested factors (step 3). See equation 4 summarizing step 2:

$$Fi = \alpha i + \beta m.Fm + \sum k=1 \text{ to } l\ (\beta \text{ selected}.F \text{ selected}) + ei \quad i=1,\ldots, N\text{-}l \tag{4}$$

For step 3, all models with different tested factors will be sorted to be compared by their explanatory powers. The model with the lowest GRS test statistic will be selected to be included into the model permanently and the remaining tested factors will be excluded from the current model to be tested again in the next iteration (if the stopping criteria is met at the next steps, excluded factors will be permanently removed from the model, see step 4 and 5).

For step 4, by using the newly constructed factor model as the base, the number of remaining significantly non-zero factor alphas will be identified where *x* is set to be the significance threshold. Now the iteration count will be increased by one: set $l := l + 1$.

$$n(\alpha)_{t>x} = |\{ ai \mid t(ai) > x \}| \qquad i= 1, \ldots, N\text{-}l \tag{5}$$

For step 5, if the number of remaining significant alphas is equal to zero (if $n(\alpha)t>x = 0$), the stopping criterion will be met, and the nested model will be completed. If not, for each iteration, a loop between step 2 and step 5 will continue to be conducted until the stopping criterion is met. The t-statistic threshold was determined as $t > 3$ for a less conservative criterion and for a more conservative approach it is also set to be $t > 2$. Meaning that if the number of significantly zero alphas is equal to zero and if the p(GRS) is higher than significance level (significance level is set to 0.05 in this study), stopping criterion will be met.

4.2.1 Example Iteration for the Nested Model Approach

To provide a better explanation for this process, iterations are illustrated.

For l=0, run each factor on CAPM.

$$F1 = \alpha 1 + \beta m.Fm + e1$$
$$F2 = \alpha 2 + \beta m.Fm + e2$$
$$\ldots$$
$$F153 = \alpha 153 + \beta m.Fm + e153 \tag{6}$$

First, calculate each model's GRS statistic. Following this, determine which factor model minimizes the GRS statistic to most. The tested factor from this model will be the selected factor which will be included into the model permanently in the next iteration. Finally, examine alphas to see if the

remaining factor models' alphas are significantly equal to zero. If this is the case, end the iteration; if not, move on to the next one iteration.

For l=1, run a two-factor Model: For each of the remaining 152 factors, run the following models (assuming factor 2 was selected in previous iteration):

$$F1 = \alpha 1 + \beta m * Fm + \beta 2 * F2 + e1$$
$$F3 = \alpha 3 + \beta m * Fm + \beta 2 * F2 + e3$$
$$\dots$$
$$F153 = \alpha 153 + \beta m * Fm + \beta 2 * F2 + e153 \quad (7)$$

Repeat the same next steps described before and if all remaining factor alphas are significantly equal to zero, stop iteration. Otherwise continue the iterations by augmenting the model with the next selected factor, running the model for the remaining factors, computing the GRS statistic, selecting the factor with the minimum GRS, and checking for remaining alphas.

4.2.2 Model Evaluation with GRS Test Statistic

GRS test is a common metric used in asset pricing as it is used in testing the hypothesis which implies that the intercepts, thus the alpha of each test asset, are jointly equal to zero (Gibbons, Ross, et Shanken, 1989). Under the GRS test, accepting the null hypothesis means the factors in the model explain all excess return and no significant alpha remains to explain the rest of the excess return. In contrast, rejecting the null hypothesis means the model's factors fail to explain all the excess returns.

While calculating GRS statistics, by using a different kind of test asset set, we assess the mean-variance efficiency of the portfolios (Kamstra et Shi, 2024). Swade et al. (2023) adopt F&F (2018)'s approach to calculate GRS statistics, while F&F (2018) follows Barillas et Shanken (2016)'s method. Barillas et Shanken (2016) assume that an accurate judgement of a model must be based on the maximum squared Sharpe ratio for the intercept where the intercept is the vector of intercepts from regression i on fi and Σi as the regression residuals' covariance matrix:

$$Sh^2(a_i) = a_i' \Sigma_i^{-1} a_i, \quad (8)$$

According to Barillas and Shanken (2016), the model with the smallest Sh^2 (αi) will outperform the remaining models. They supported this approach with the following relationship provided by GRS (1989), implying that Sh^2 (αi) is equal to the difference between the maximum squared sharpe ratio of fi along with regression i and the maximum sharpe ratio of fi by itself. Therefore, as can be seen from equation 9, to minimize Sh^2 (αi), Sh^2 (fi) must be maximized. Aligned with this approach, later, Fama & French (2018) suggested that instead of focusing on minimizing Sh^2 (αi), one must focus on maximizing Sh^2 (fi). These two claims of course support the same approach, but via different perspectives.

$$Sh^2(a_i) = Sh^2(\Pi_i, f_i) - Sh^2(f_i). \tag{9}$$

In this context, the maximum squared sharpe ratio of fi is defined as follows: **Ω** is the covariance matrix of factors and f is the model's average factor returns (Swade et al., 2023). See equation 10:

$$Sh^2(f) = \overline{f}^{\mathsf{T}} \Omega^{-1} \overline{f} \tag{10}$$

In conclusion, for a model with N test assets, K factors, and τ return observations, following GRS test statistic formula is being conducted where F GRS ~ F (N, τ −N −K):

$$F_{GRS} = \frac{\tau(\tau - N - K)}{N(\tau - K - 1)} \frac{Sh^2(\alpha)}{(1 + Sh^2(f))} \tag{11}$$

One should note that the data set collected from JPK (2023), same dataset used in Swade et al. (2023), my sample has monthly observations of 153 factors between 11/1971 to 12/21; thus, τ (tau) is a constant number of observations: 602, N is the number of test assets which is set to be 153 on the first iteration and then decreases by 1 after each iteration, and lastly, K is the number of selected factors in the model which is set to be 1 when l=0 and then increases by 1 after each iteration.

As detailed empirical explanations of GRS statistics fall outside the scope of this research, mathematical derivation of the GRS Test Statistic is excluded from this chapter. For further explanation, please see Barillas and Shanken (2016) and F&F (2018). Moreover, for the latest empirical explanation of GRS statistics, please see Kamstra and Shi (2024) which provides insights into confusion on the GRS Formula and proposes a novel method to evaluate competing models using the GRS statistic p-value.

4.2.3 A Local Optimum Investigation: Does Nested Model provide suboptimal solutions?

As previously mentioned in Section 2, "Theoretical Framework", it is crucial to consider that the RHS nested model approach has the potential to create a local optimum problem. Therefore, as part of the goal of this paper, to span factor zoo with different models other than the full-sample nested model, I run the same nested model with a rolling window and expanding window to the same data set. Moreover, I implemented other ML methods such as Least Absolute Shrinkage and Selection Operator (LASSO) Regression, Principal Component Analysis (PCA), Partial Least Squares (PLS) Regression, and PLS with Elastic Nets to the same data set. Then I compared the results and robustness of different models. Please see Section 5 for results and Section 6 for Robustness results. The ML techniques used in this study have considerable potential to outperform the nested model. For instance, due to the regularization nature of the LASSO regression, this method may have the potential to avoid local optimum. Using an additional technique will highlight whether there is such a local optimum problem by comparing the effectiveness of those methods in producing more accurate statistics compared to the nested model.

4.3 LHS Approaches: Lasso Regression, PCA, PLS, PLS Elastic Nets

For further models, I adopted Lasso Regression as being one of the most popular regularization methods, PCA, and PLS too to observe how impactful is using the covariance structure of factors to span the factor zoo next to the alpha perspective and regularization. I also ran a PLS model by incorporating it with Elastic Nets to see the combination of covariance structure and regularization techniques.

4.3.1 Lasso Regression to Span Factor Zoo

LASSO Regression minimizes the residual sum of squares while enforcing a constraint that the total absolute value of the coefficients remains below a certain limit which ensures some coefficients are zero (Tibshirani, 1996). Please see Chapter 2, for a detailed explanation on the latest papers, Feng et al. (2020) and Kozak et al. (2020), using LASSO regression to span the factor zoo. The selection of the regularization parameter alpha is a crucial step in lasso regression. To select the best alpha, I used cross-validation and selected the best alpha as 0.00018. However, to adopt a more conservative approach and to decrease the number of factors explaining the factor zoo further, I slightly increased the best alpha and used an alpha of 0.00020.

4.3.2 PCA to Span Factor Zoo

A very popular unsupervised learning technique in the Finance Literature is Principal Component Analysis (PCA), which is used for dimension reduction in high-dimensional financial datasets. While dealing with the high-dimensionality problem, PCA also preserves the ideal level of variability. The most important feature of PCA is transforming the explanatory variables into a new collection of uncorrelated variables called principal components. PCA reduces the dimensionality of the data by concentrating analysis on its most crucial properties (Abdi and Williams, 2010).

4.3.3 PLS to Span Factor Zoo

Partial least squares (PLS) method is a recent supervised learning technique which investigates the underlying patterns that can explain the dependent variable and its predictors. One of the most featured qualifications of PLS is preventing the model to suffer from a multicollinearity issue (James et al., 2023). PLS identifies a set of principal components by creating linear combinations of the original features. This new collection of features is then fitted to a linear model using least squares regression (Abdi, 2010). This supervised method has the potential to reduce bias

4.3.4 PLS with Elastic Nets to Span Factor Zoo

To adopt a more conservative selection and increase the robustness of the PLS model, I included regularization into the PLS regression by incorporating Elastic Net Regression into PLS. With this inclusion, the model shrinks less important feature coefficients to zero before implying PLS. That means that I removed highly correlated features before applying PLS. Maintained cross-validation to determine the optimal number of components for PLS. This approach was inspired by two latest literatures on Factor Zoo. The first one is by Wan et al. (2024), a paper which claims that using covariance structure in the model might create bias, whilst the latter is unable to consider additional information. Therefore, they claim using a penalized reduced rank regression can give more robust outcomes. The second one is by Feng et al. (2020) where authors find that elastic net regression is a great alternative method in factor selection. Therefore, I adapted a model by assuming that a model combining PLS and Elastic Net would improve the model to span factor zoo when using a covariance structure. Please note that the PLS with Elastic Nets approach is a model assumption made by the author and it is open to discussion.

CHAPTER 5 Results

5.1 Spanning Factor Zoo with Nested Models (RHS Approach)

5.1.1) Spanning Factor Zoo with the Full-Sample Nested Model

I ran a nested model on the full sample to replicate the methodology of Swade et al. (2023), using the same dataset and time period. Please refer to Section 3: Data for details on the dataset and the time period used.

Table 1: Iterative Factor Selection with Full Sample Nested Model

Iteration	Selected Factor	GRS Value	p(GRS)	Sh^2(f)	Sh^2(alpha)	Remaining Alphas (t > 2)	Remaining Alphas (t > 3)
0	Market						
1	cop_at	3,88	0,00	0,15	1,52	105	86
2	noa_gr1a	3,25	0,00	0,27	1,39	98	78
3	saleq_gr1	2,86	0,00	0,33	1,28	65	34
4	ival_me	2,62	0,00	0,39	1,21	41	11
5	resff3_12_1	2,44	0,00	0,44	1,16	38	14
6	seas_6_10an	2,26	0,00	0,50	1,10	34	15
7	debt_me	2,09	0,00	0,54	1,04	25	9
8	seas_6_10na	1,97	0,00	0,58	1,00	35	7
9	zero_trades_252d	1,87	0,00	0,61	0,96	23	3
10	cowc_gr1a	1,79	0,00	0,65	0,93	13	1
11	nncoa_gr1a	1,68	0,00	0,70	0,89	14	3
12	ocf_me	1,58	0,00	0,73	0,84	6	1
13	zero_trades_21d	1,51	0,00	0,76	0,81	5	1
14	turnover_126d	1,41	0,00	0,81	0,78	11	1
15	rmax5_rvol_21d	1,30	0,02	0,85	0,73	8	2
16	seas_11_15na	1,23	0,06	0,87	0,69	3	0
17	o_score	1,19	0,09	0,89	0,67	2	0
18	niq_at	1,16	0,13	0,91	0,65	3	0
19	seas_16_20an	1,13	0,18	0,92	0,63	0	0
20	ni_ar1	1,11	0,22	0,93	0,62	0	0
21	ivol_ff3_21d	1,09	0,25	0,95	0,61	1	0
22	ni_me	1,07	0,30	0,96	0,60	2	0
23	dsale_dinv	1,04	0,38	0,98	0,58	0	0
24	ni_be	1,02	0,43	0,99	0,57	0	0
25	noa_at	1	0,47	1,00	0,55	1	0
26	age	0,98	0,55	1,01	0,54	0	0
27	ret_12_1	0,96	0,59	1,02	0,53	0	0
28	aliq_mat	0,95	0,63	1,03	0,52	0	0
29	nfna_gr1a	0,94	0,65	1,04	0,51	0	0
30	at_me	0,93	0,68	1,05	0,50	0	0

**Please note that this study faced some limitations: Due to lack of technical instructions on Sharpe Squared Alpha calculation, my GRS values slightly deviates from Swade et al. (2023)'s results (only with decimal points). Selected factors are identically same and in the correct order. This limitation caused to span factor zoo with 16 factors instead of 15 factors like it was done by Swade et al. (2023). However, according to F&F (2018), what matters is sharpe squared factors (Sh^2_(f)) and it is fully aligned with Swade et al. (2023).*

As demonstrated in Table 1, the factor with the lowest GRS statistic (3.88) is cop_at (Cash Based on Operating Profitability scaled by Assets), making it the first selected factor. This indicates that cop_at minimized alpha the most compared to other factors, leading to its permanent inclusion in the RHS of the model. Despite this, p(GRS) remained zero, and there were still 105 and 85 significant non-zero alphas (for (t > 2) and (t > 3), respectively), meaning the stopping criterion was not yet met. In the second iteration, the model regressed on "market + cop_at" and the factor noa_gr1a (Net Operating Assets Change 1yr) was selected with a GRS statistic of 3.25. This led to a moderate decrease in the number of remaining non-zero alphas. The iteration loop continued in this direction. Table 1 shows that in the less conservative approach (t > 3), the number of significantly non-zero alphas reduced to 1 by the 10th iteration and reached 0 by the 16th iteration. In the more conservative approach (t > 2), the number of significantly non-zero alphas reached 0 by the 19th iteration. Additionally, p(GRS) surpassed the 5% significance level after the 16th iteration. In conclusion, using a full sample nested model, the factor zoo was compressed to 16 factors. This means that 16 factors are required to span the factor zoo if a less conservative t-statistic and p(GRS) approach is adopted. According to the results in Table 1, the selected factors are categorized as follows: 4 factors belong to the Low Risk theme, 3 to the Investment theme, 3 to the Value theme, 2 to the Seasonality theme, 1 to the Short-term Reversal theme, 1 to the Accruals theme, 1 to the Quality theme, and 1 to the Momentum theme.

5.1.2) Spanning Factor Zoo with the Rolling Window Nested Model

[Table 2 about here.]

Please see Table 2 for the same nested model where it was built on a rolling window sample instead of a full sample. Firstly, Table 2 demonstrates that with a rolling window sample, the results highly deviate compared to the full-sample nested model. The first iteration captures a different factor with a much higher GRS statistic value: nfna_gr1a (Net Financial Assets Change 1yr) with a GRS value of 8.7. This factor was not even selected in the full-sample nested model. If we compare with the full-sample nested model, there are only 8 factors overlapping in the first 15 selected factors. On the other hand, with a stopping criterion of t>3, it takes 25 factors to span factor zoo, whereas in the full-sample nested model, it took 16 factors. Moreover, if we adopt the p(GRS) approach as the stopping criterion, it takes 76 factors to span the factor zoo and some of the selected factors deviate as the window size changes. Thus, please note that for comparison purposes, only results up to 30 iterations are presented in this study. The table with 76 iterations is not available.

5.1.3) Spanning Factor Zoo with the Expanding Window Nested Model

[Table 3 about here.]

Please refer to Table 3 for the results of the nested model built on an expanding window sample instead of a full sample. The expanding window nested model selects the same first 10 factors as the rolling window model, with nearly identical GRS statistics. However, it is observed that in the expanding window model, the GRS statistics decrease more rapidly than in the rolling window model. This fact suggests a better performance for expanding window. Adopting a non-conservative approach ($t > 3$), it takes 13 factors to span the factor zoo. In a more conservative approach ($t > 2$), it takes 23 factors to achieve the same span. If we use the p(GRS) perspective as the stopping criterion, the expanding window model claims to explain the entire factor zoo with 27 factors, as p(GRS) reaches 0.06 after the 27th iteration.

5.2 Spanning Factor Zoo with Latest ML Techniques (implying LHS Approach)

5.2.1 Using Regularization: LASSO

With the lasso regression where I regressed the market excess return on 153 factors, it takes 76 factors to span the factor zoo. Please see Table 4 in the Appendix for factor selection with Lasso Regression. From Table 4, it can be observed that factor ami-126d (Amihud Measure) is the biggest positive contributor to the model with a coefficient value of 0.022. On the other hand, factor dolvol_126d (dollar trading volume) is the biggest negative contributor with a coefficient value of -0.018. Overall, Lasso regression indicates that to cover a broad set of 153 factors, rather than assigning higher weights to fewer factors, it is more effective to select a larger number of factors (76 factors) with lower coefficients. Furthermore, Lasso Regression demonstrated strong robustness metrics, including an R^2 value of 0.71 (see Section 6: Robustness).

[Table 4 about here.]

The selection of the regularization parameter alpha is a crucial step in lasso regression. To select the best alpha, I used cross-validation and selected the best alpha as 0.00018. However, to adopt a more conservative approach and to decrease the number of factors explaining the factor zoo further, I slightly increased the best alpha and used an alpha of 0.00020. Please see Table 5 in the Appendix to see how R2 and MSE metrics deviate with different alpha values.

5.2.2 Using Covariance Structure: 1) PCA and 2) PLS

5.2.2.1 PCA

Table 6: Explained Variance with Top 3 PCA Components

Component	Explained Variance	Cumulative Explained Variance
1	0.362752	0.362752
2	0.146185	0.508937
3	0.077399	0.586336

I ran PCA regression where I regressed market excess return on 153 factors (LHS approach). According to the main results, it takes 38 components to explain a minimum 95% of the variance among independent variables. Refer to Table 6 for the explained variances associated with the top 3 principal components in the PCA analysis. The first principal component accounts for 36% of the variance by itself, while the top three components collectively explain 58% of the variance. Beyond the first three components, the cumulative explained variance increases at a slower rate, ultimately reaching 95% with the 38th component.

Please note that Table 6 details the variance explained by only the top three principal components. For a comprehensive view of how the explained variance progressively improves across all 38 PCA components, refer to Table 7 in the Appendix.

It is important to understand the contribution of each factor in the zoo to the 3 different PCA components. See Table 4 in the appendix for factor weighting scheme. It is observed that while the factor contributors in Component 1 mostly explains negative variance, the factor contributors in Component 3 explains mostly positive variance. On the other hand, features of Component 2 have a more balanced distribution.

5.2.2.2 PLS without Regularization

PCA is a commonly used method in finance literature for explaining the covariance matrix in predictions. However, a more recent method, PLS, not only explains the covariance of independent variables but also considers their relationship with the dependent variable. I ran PLS regression where I regressed market excess return on 153 factors (LHS approach). According to the main results, it takes 4 components to explain 100% of the variance.

Table 8: Explained Variance with PLS Components

Component	Explained Variance	Cumulative Explained Variance
1	0.593847	0. 593847
2	0.167734	0.761581
3	0.135553	0.897134
4	0.102866	1

Refer to Table 8 for the explained variances of the four PLS components. The first component alone captures 59% of the variance, and the top four components together explain 100% of the variance cumulatively. Unlike PCA components, which do not consider the relationship with the dependent variable, PLS components require significantly fewer components to capture the entire variance. See Table 4 in the Appendix for more information on the factor weighting scheme, which shows how various factors contribute to the four PLS components. I used cross-validation to find the number of components that produced the highest R2 score in order to calculate the ideal number of components. See Figure 6 for variations in R2 and MSE with varying numbers of components.

[Figure 6 about here.]

5.2.2.3 PLS with Regularization: Elastic Nets in PLS

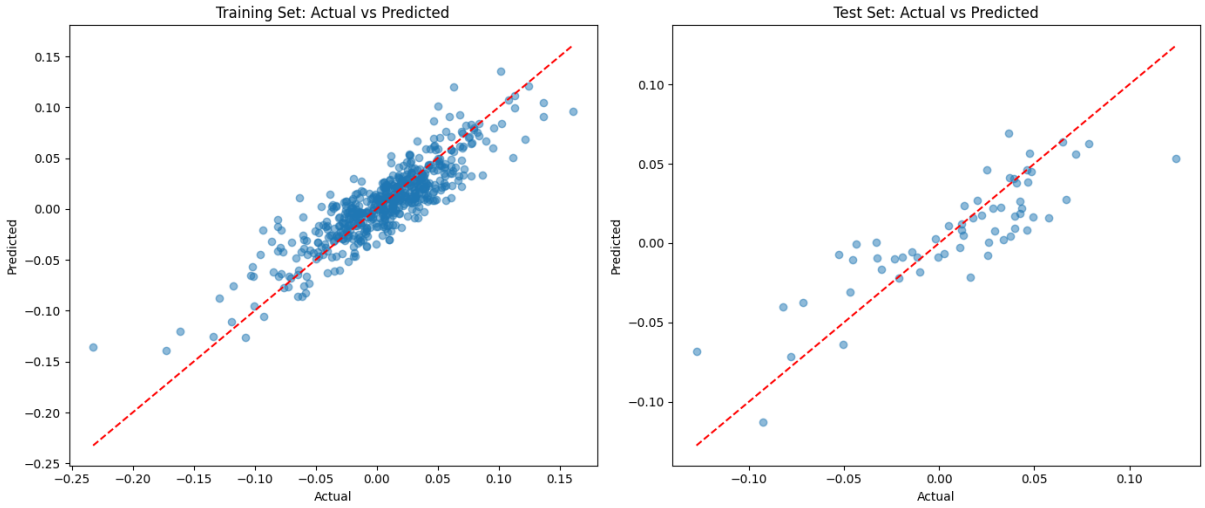
In this method, highly correlated features were removed prior to applying PLS, and Elastic Net regression was used to penalize large coefficients, thereby making the model more conservative. Using the PLS with Elastic Net model, 5 components were needed to capture 100% of the variance, compared to 4 components in the regular PLS model. Cross-validation was used once more to find the number of components that maximized the R2 value, in order to ascertain the ideal number of components. With Elastic Net, the model's robustness grew as the number of components rose (see Section 6: Robustness). For information on the factors that contribute to the PLS with Elastic Net model's components, see Table 4.

***Please note that the PLS with Elastic Nets approach is a model assumption made by the author and it is open to discussion.*

CHAPTER 6 Robustness

6.1 Lasso Regression – Robustness

Figure 7: Actual vs. Predicted Values for Training and Test Sets - LASSO



For the distribution of predictions in the Lasso regression model, please refer to Figure 7. Firstly, in the robustness analysis, it is important to set efficient rates for the training set and the test set. Considering the size of the sample, the training set is set to comprise 90% of the full sample, while the test set is set to cover the remaining 10%. While interpreting the scatter plot, please note that the Y-axis displays the predicted values produced by the Lasso model using the training set, and the X-axis reflects the actual target values from the entire sample. The ideal scenario is shown by the red line where the y and x axes are equal. Despite a few dispersed predictions, the Lasso regression successfully clusters its predictions around the red line overall, as seen from the figure, suggesting a strong overall fit. With an R2 of 0.71 and an MSE of 0.0006, Table 9—which is shown below—further exemplifies the Lasso model's resilience. It appears to be the most resilient model among all LHS models.

Table 9: Lasso Regression Metrics

LASSO	
Best Alpha Selected by Cross-Validation:	0,00018
Alpha used in the model:	0,0002
MSE - Test Sample	0,0006
R2 - Test Sample	0,71

6.2 PCA – Robustness

Figure 8: Actual vs. Predicted Values for Training and Test Sets - PCA

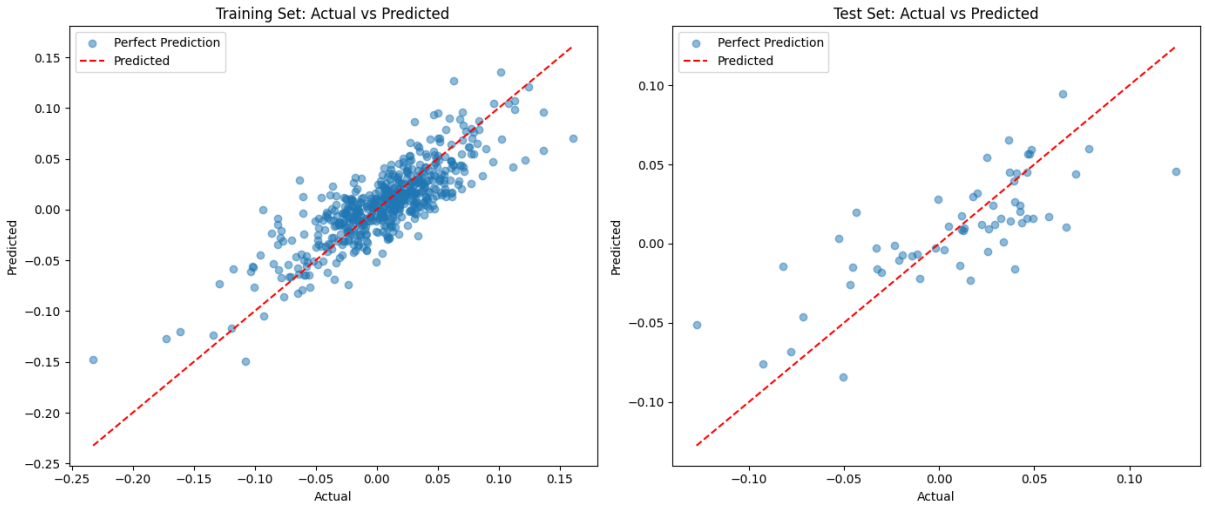


Figure 8 represents the distribution of predictions in the PCA regression model. The training set was selected as 90% of the full sample, whereas the test set was set to 10% of the full sample. Compared to lasso regression, the predictions are less clustered around the perfect prediction line where y=x. Therefore, as presented in Table 10, we observe moderately lower R2 (0.5940) and higher MSE (0.00086) in the PCA model compared to the Lasso model.

Table 10: PCA Metrics

PCA	
Number of Components min. 95% Variance	38
MSE - Test Sample	0,00086
R2 - Test Sample	0,5940

6.3 PLS – Robustness

Figure 9: Actual vs. Predicted Values for Training and Test Sets – PLS

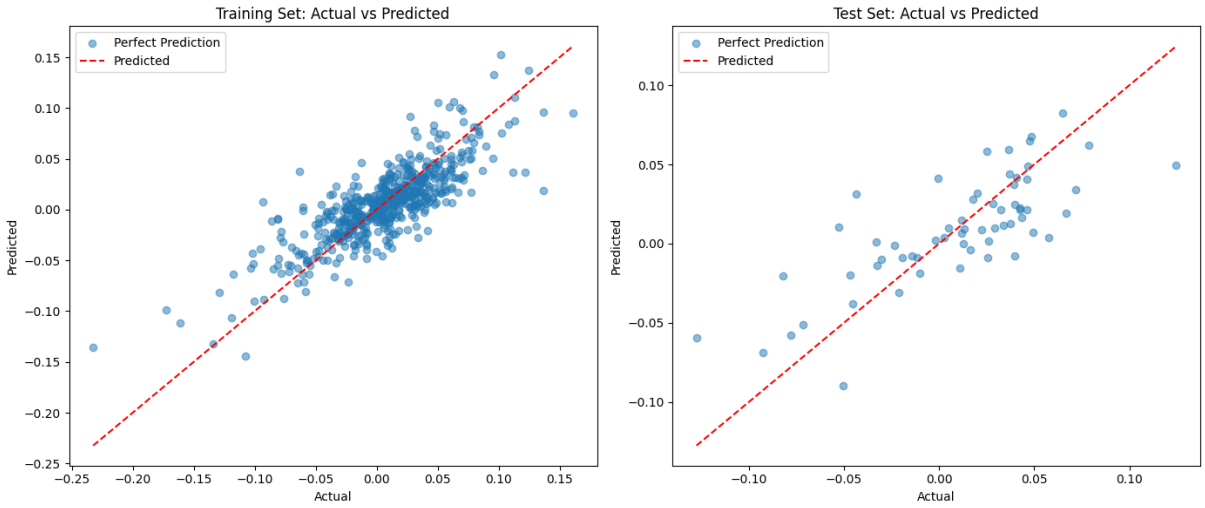


Figure 9 shows the distribution of predictions in the PLS regression model. The training set comprises 90% of the full sample, while the test set constitutes the remaining 10%. Compared to the Lasso regression, similar to the PCA model, the predictions in the PLS model are less clustered around the perfect prediction line. As shown in Table 11, the PLS model has an R^2 value of 0.5901 and an MSE value of 0.00087. This indicates significantly less robustness compared to the Lasso model and slightly less accuracy compared to the PCA model.

Table 11: PLS Metrics

PLS	
Number of Components Explaining Variance - based on R2	4
MSE - Test Sample	0,00087
R2 - Test Sample	0,5901

6.4 PLS with Elastic Nets – Robustness

Table 12: PLS with Elastic Nets Metrics

PLS with Elastic Nets	
Number of Components Explaining Variance - based on R2	5
MSE - Test Sample	0,00080
R2 - Test Sample	0,6191

To improve the PLS model's accuracy, regularization was implemented using Elastic Nets. This involved regularizing the covariance matrix with Elastic Nets before running the PLS model. As a result, the R^2 value increased from 0.5901 to 0.6191, and the MSE decreased from 0.00087 to 0.00080, as demonstrated in Table 12.

6.5 Metric Comparison between Nested Models - Robustness

Figure 10: Metric Evaluation in Different Nested Models

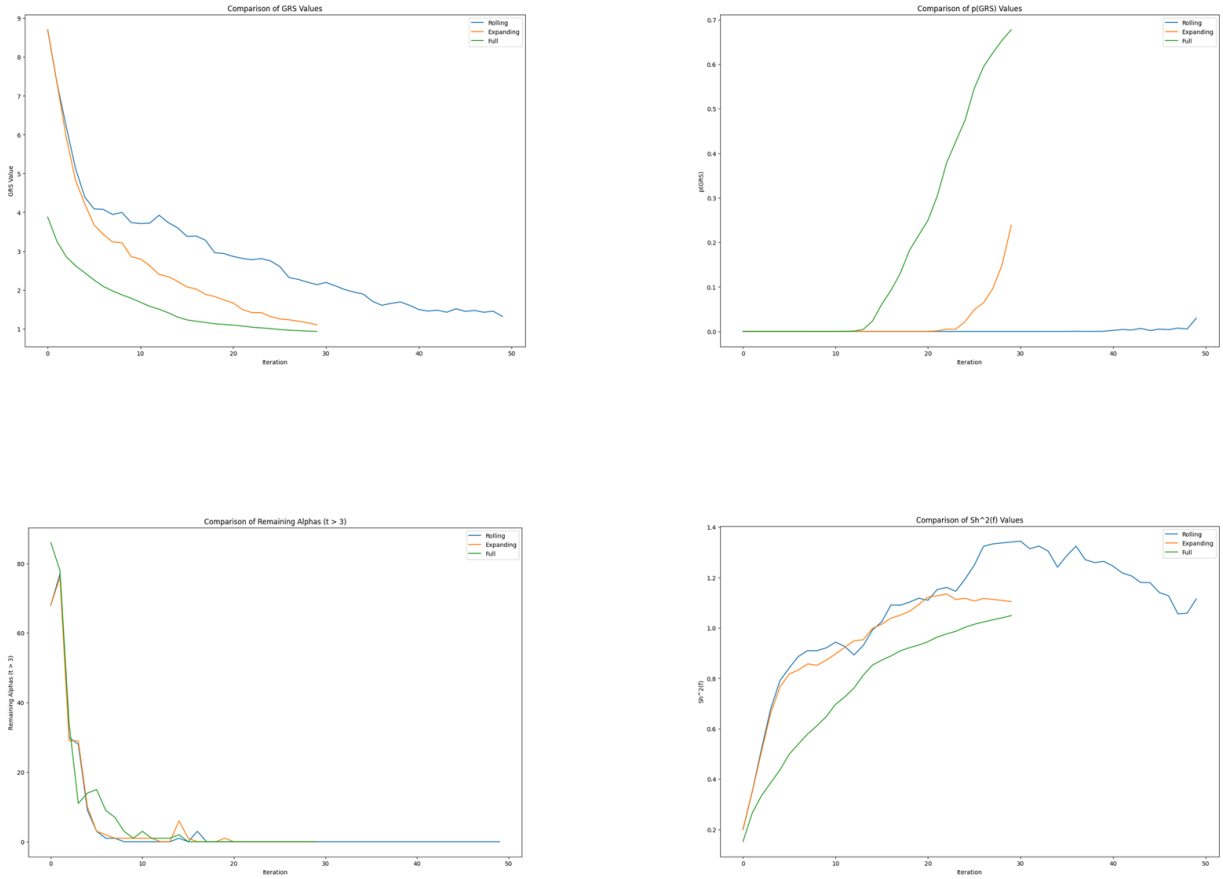


Figure 10 displays a comparison of metrics across three different nested models: 1) full sample, 2) rolling window, and 3) expanding window. The first impression is on the GRS Values: Rolling window and expanding window models present significantly higher GRS Vaues compared to the full sample nested model. Moreover, the rolling window model exhibits a slower decrease in GRS values compared to other models. On the other hand, it is observed that the p(GRS) value increases faster in the full sample nested model compared to the other nested models. According to these results, it is crucial to interpret that using p(GRS) as the stopping criterion may lead to a biased decision making, as the model may not have achieved to global optimum yet. Moreover, in the rolling window model, spanning the factor zoo requires considerably more iterations (77 iterations, as discussed in Section 5) to meet the stopping criteria for p(GRS). Instead of taking the p(GRS) as the stopping criterion, it is also possible to use the remaining alpha perspective. For instance, in terms of the less conservative stopping criterion of t>3, the three models exhibit similar behaviour for the number of remaining alphas, but with different factor selection. Another critical metric that should be examined carefully is the Sharpe squared factors (Sh^2(f)). Sh^2(f) increases more rapidly in the rolling window and expanding window models, indicating their superior performance in spanning the factor zoo despite their initially higher GRS values. According to Fama & French (2018), the Sharpe squared factor is the most prominent metric for model evaluation.

CHAPTER 7 Discussion and Limitations

This section contextualizes the results, discussion points and limitations of this study. Dimension reduction in Factor Zoo appears to be a highly popular topic, especially now that new approaches and procedures are being introduced. While examining the stability of current models, this study also looks for alternate strategies that offer better precision and durability. The existing literature asserts that the Right-Hand Side (RHS) approach has less sample dependency than the Left-Hand Side (LHS) approach, suggesting it should result in less biased outcomes (refer to Section 2 for further explanation). However, the varied outcomes from the three different nested models using different samples (full sample, rolling window, expanding window) open the discussion on the reliability of the RHS approach. The deviations mentioned above should be discussed by further studies on this specific topic as they show that the RHS approach may be suffering from local optima. Therefore, this study suggests further research to address the local optima. Additionally, this research cannot directly claim that the rolling window or expanding window method is superior to using a full sample. It is also crucial to consider the potential for overfitting in rolling windows, especially when smaller window sizes are used to generate more iterations to span factor zoo. In this context, the expanding window approach might emerge as the most accurate among the nested models due to its capacity to incorporate more data over time, potentially offering a more robust solution. There are several limitations in the nested models that should be acknowledged. First, this study encountered challenges in calculating the GRS statistics used in the nested model. Specifically, the full-sample nested model was conducted to replicate the approach from Swade et al. (2023). Due to insufficient technical instructions on sharpe squared alpha calculation, my GRS values deviated slightly from those reported by Swade et al. (2023) in terms of decimals. However, my sharpe squared factor values are fully aligned with Swade et al.'s results. Given this alignment, the small decimal variation in GRS values is considered acceptable as Sh^2 (f) values are fully matching. This claim can be supported by the suggestion of Fama and French (2018) which claims that the optimal model to minimize remaining alphas is the one with the largest Sh^2 (f). Furthermore, while my selected factors are identical and in the correct order as in the replicated paper, this limitation resulted in spanning a factor zoo with 16 factors which is just one factor more compared to results of Swade et al. (2023). Finally, since the inclusion of metrics of robustness such as MSE or R2 was not possible under the nested model framework, the robustness analysis was limited. To address this, the nested model was conducted with three different sample constructions (full-sample, rolling window, expanding window), and the results were compared to analyse robustness. Consequently, the metric evaluation scheme (Figure 10) illustrates different metric values for different selected factors, although there may be overlaps within selected factors. This scheme makes it possible for readers to follow the ways in which sample selection would affect other metrics and factor selections, thus building sample construction into the considerations of model evaluation. These limitations emphasize the necessity of careful interpretation

of the results and further research for methodology improvement used in nested models to make outcomes more reliable and accurate. The outcome suggests that a global optimum of nested modelling would not be possible, though the interpretation of Figure 10 is arguable. This also sets up an issue for further discussion on how a proper robustness test among nested models can be performed. As discussed earlier, one of the primary objectives of this study was to identify the most effective machine learning technique for spanning the factor zoo. To this end, the LHS approach was adopted, regressing the market on 153 factors in the zoo. Among the LHS models, Lasso regression outperformed the others. While both PCA and PLS models demonstrated comparable robustness, PCA exhibited slightly better accuracy by decimal points. However, PLS, which reduces the predictors to a smaller number of orthogonal components, was expected to yield more accurate results due to its efficient handling of multicollinearity among the predictors. This raises an important question: when spanning the factor zoo, is it more effective to explain the covariance structure of the independent variables alone, or should the relationship with the dependent variable be incorporated? Although this is open to discussion, the small difference observed in robustness of PCA and PLS models is insufficient to draw definitive conclusions. Furthermore, consistent with existing literature, Lasso regression emerged as the best-performing model among all LHS models. Feng et al. (2020) also identify Lasso regression as the most suitable method for spanning the factor zoo, noting the strong performance of the elastic nets model as well. This indicates that both my findings and the existing finance literature suggest that regularization techniques offer higher accuracy compared to models that use a covariance structure. Motivated by this insight, I incorporated regularization into the PLS model using elastic nets to determine if regularization could improve a covariance-based model. Indeed, among all models with a covariance structure (PCA, PLS, PLS with Elastic Nets), PLS with Elastic Nets exhibited better robustness metrics. However, the appropriateness of combining these two models remains open to discussion, and according to my findings, it is not yet supported by a robust mathematical foundation in the literature. Therefore, it is important to note that the PLS with Elastic Nets model was developed based on the author's assumptions. In drawing conclusions to compare all models in this study, including both RHS and LHS approaches, certain limitations were encountered. It should be highlighted that as robustness metrics such as R2 and MSE were not applicable for the nested models, a comparison among RHS and LHS models could not be conducted, and this limits the conclusion of this study. This limitation underlines the need for further research and discussion on how to evaluate different models together.

CHAPTER 8 Conclusion

This work had two main objectives. The first one was to determine whether doing a nested model with the RHS approach is good at spanning the factor zoo or, rather, it falls on problems of local optima. The second objective was to search for what techniques from the latest machine learning literature could be better candidates for spanning the factor zoo in the attempt to reach the global optimum. With regards to the nested models, in terms of the full sample, rolling window model, and expanding window model, the selected factors differ quite vehemently. By similar token, the number of factors necessary to span the factor zoo is higher under rolling and expanding windows. Especially in the case of the rolling window model, substantially more factors were required. This means that there might be some important factors in the zoo which are omitted by the full-sample nested model. Moreover, notice that the three of the nested models produce noticeably different values according to the GRS statistic, p(GRS), and Sharpe-squared factors. It suggests that the full-sample model is probably caught in a local optimum, and one may have difficulties to reach the global optimum using the RHS approach built on the alpha perspective. In conclusion, one big area of interest in the discussion is the kind of trade-offs that exist between different sample constructions and their effect on model performances. While a full-sample model can possibly give an overall view, it could lose very essential factors that only show up under the rolling and expanding window models. To obtain the second objective, the LHS approach was adopted, and the market excess return was regressed on the 153 factors of factor zoo with different ML techniques. Among all LHS models, the Lasso Regression returned a higher accuracy than the PCA and PLS models, as it has showed the best metrics for the R-squared and mean-squared errors. These results show that a regression model with regularization outperforms fitting a linear model with components describing the covariance matrix. Encouraged by this notion, I introduced regularization to the PLS model using Elastic Nets. Indeed, the PLS model with Elastic Nets also did better than the plain PLS model and showed better robustness metrics. In this study, it was challenging to carry out a proper comparison between the RHS and LHS models as these models had different regression structures. However, the inconsistencies of selected factors and metrics across different models seem to imply that the nested model may suffer from local optimum compared to other models. For example, Lasso regression was proven to be robust and accurate; however, the selected factors from Lasso Regression were not fully aligned with the selected factors in nested models. This further suggests an insight into areas that need more study in order to enhance the robustness and reliability of the RHS approach. Besides, the findings from this study will support further examination of innovative machine learning methods, like regularization techniques and their implementation on different regression models.

REFERENCES

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley interdisciplinary reviews: computational statistics*, *2*(1), 97-106.

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433-459.

Addis, B., Locatelli, M., & Schoen, F. (2005). Local optima smoothing for global optimization. Optimization Methods and Software, 20(4-5), 417-437.

Barillas, F., & Shanken, J. (2016). Which alpha?. The Review of Financial Studies, 30(4), 1316-1338.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. Review of Economic Studies, 81(2), 608-650.

Bessembinder, H., Burt, A., & Hrdlicka, C. M. (2021). Time series variation in the factor zoo. Aaron Paul and Hrdlicka, Christopher M., Time Series Variation in the Factor Zoo (December 22, 2021).

Bessembinder, H., Burt, A., & Hrdlicka, C. M. (2022). Factor returns and out-of-sample alphas: Factor construction matters. Aaron Paul and Hrdlicka, Christopher M., Factor Returns and Out-of-Sample Alphas: Factor Construction Matters (November 19, 2022).

Bryzgalova, S., Huang, J., & Julliard, C. (2023). Bayesian solutions for the factor zoo: We just ran two quadrillion models. The Journal of Finance, 78(1), 487-557.

Chen, Luyang, Markus Pelger, and Jason Zhu (2024) "Deep Learning in Asset Pricing," Management Science,20 (2), 714–750.

Chernozhukov, V., Hansen, C., & Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. Annu. Rev. Econ., 7(1), 649-688.

Chotchantarakun, K. (2023). Optimizing Sequential Forward Selection on Classification using Genetic Algorithm. Informatica, 47(9).

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.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. Journal of financial economics, 116(1), 1-22.

Fama, E. F., & French, K. R. (2018). Choosing factors. Journal of financial economics, 128(2), 234-252.

Feng, G., Giglio, S., & Xiu, D. (2017). Taming the factor zoo (pp. 1-56). Working paper.

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. The Journal of Finance, 75(3), 1327-1370.

Gibbons, Michael R., Stephen A. Ross, and Jay Shanken (1989). "A test of the efficiency of

a given portfolio." Econometrica, 1121–1152

Giglio, S., & Xiu, D. (2021). Asset pricing with omitted factors. Journal of Political Economy, 129(7), 1947-1990.

Global Factor Data. (n.d.). https://jkpfactors.com/Kenneth R. French - Data Library. (n.d.). https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html#Benchmarks

Harvey, C. R., Liu, Y., & Zhu, H. (2015). … and the cross-section of expected returns. The Review of Financial Studies, 29(1), 5-68.

Harvey, C., Liu, Y., 2016. Lucky factors. Unpublished working paper. Duke University, Fuqua School of Business, Durham, NC

He, A., Huang, D., Li, J., & Zhou, G. (2023). Shrinking factor dimension: A reduced-rank approach. Management science, 69(9), 5501-5522.

Heston, S. L., & Sadka, R. (2008). Seasonality in the cross-section of stock returns. Journal of Financial Economics, 87(2), 418-445.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An introduction to statistical learning: With applications in Python. Springer.

Jensen, T. I., Kelly, B., & Pedersen, L. H. (2021). Global Factor Data Documentation.

Jensen, T. I., Kelly, B., & Pedersen, L. H. (2023). Is there a replication crisis in finance?. The Journal of Finance, 78(5), 2465-2518.

Kamstra, M. J., & Shi, R. (2020). A Note on the GRS Test. Available at SSRN 3775089.

Kamstra, M. J., & Shi, R. (2024). Testing and Ranking of Asset Pricing Models Using the GRS Statistic. Journal of Risk and Financial Management, 17(4), 168.

Kozak, S., Nagel, S., & Santosh, S. (2018). Interpreting factor models. The Journal of Finance, 73(3), 1183-1223.

Kozak, S., Nagel, S., & Santosh, S. (2020). Shrinking the cross-section. Journal of Financial Economics, 135(2), 271-292.

Lettau, M. (2023). High-dimensional factor models and the factor zoo (No. w31719). National Bureau of Economic Research.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?. Journal of classification, 31, 274-295.

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. Psychological Bulletin, 105(1), 143.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. The journal of finance, 19(3), 425-442.

Si, Y., Zhang, Y., & Li, G. (2022). An efficient tensor regression for high-dimensional data. arXiv preprint arXiv:2205.13734.

Soebhag, Amar, Bart Van Vliet, and Patrick Verwijmeren (2023). “Non-Standard Errors in Asset Pricing: Mind Your Sorts.” Available at SSRN: https://ssrn.com/abstract=4136672.

Somol, P., Pudil, P., & Kittler, J. (2004). Fast branch & bound algorithms for optimal feature selection. IEEE Transactions on pattern analysis and machine intelligence, 26(7), 900-912.

Swade, A., Hanauer, M. X., Lohre, H., & Blitz, D. (2023). Factor zoo (. zip). Available at SSRN.

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

Wan, R., Li, Y., Lu, W., & Song, R. (2024). Mining the factor zoo: Estimation of latent factor models with sufficient proxies. *Journal of Econometrics*, *239*(2), 105386.

APPENDIX

Table 2: Iterative Factor Selection with Rolling Window Nested Model

Iteration	Window Start	Window End	Selected Factor	GRS Value	p(GRS)	Sh^2(f)	Sh^2(alpha)	Remaining Alphas (t > 2)	Remaining Alphas (t > 3)
0			Market						
1	0	400	nfna_gr1a	8,7	0,00	0,20	3,5	105	68
2	6	406	ival_me	7,3	0,00	0,35	3,3	100	77
3	12	412	resff3_12_1	6,0	0,00	0,51	3,0	56	27
4	18	418	cop_atl1	4,8	0,00	0,66	2,6	67	25
5	24	424	cowc_gr1a	4,6	0,00	0,76	2,7	26	9
6	30	430	seas_2_5an	4,2	0,00	0,80	2,5	15	3
7	36	436	debt_gr3	4,0	0,00	0,81	2,3	12	2
8	42	442	seas_6_10an	4,0	0,00	0,83	2,3	8	1
9	48	448	zero_trades_252d	4,0	0,00	0,80	2,3	23	0
10	54	454	zero_trades_21d	3,7	0,00	0,88	2,2	22	1
11	60	460	debt_me	3,6	0,00	0,91	2,1	24	0
12	66	466	ocf_me	3,5	0,00	0,97	2,1	14	3
13	72	472	tangibility	3,0	0,00	0,98	1,9	11	0
14	78	478	nncoa_gr1a	3,0	0,00	0,98	1,8	9	0
15	84	484	saleq_gr1	3,9	0,00	1,06	1,8	23	3
16	90	490	seas_11_15na	2,8	0,00	1,05	1,7	4	0
17	96	496	taccruals_ni	2,8	0,00	1,08	1,7	5	0
18	102	502	seas_6_10na	2,5	0,00	1,14	1,6	5	1
19	108	508	noa_gr1a	2,4	0,00	1,18	1,5	5	0
20	114	514	dbnetis_at	2,3	0,00	1,18	1,4	1	0
21	120	520	turnover_126d	2,3	0,00	1,18	1,5	2	0
22	126	526	rmax5_rvol_21d	2,2	0,01	1,14	1,3	3	0
23	132	532	oaccruals_at	2,1	0,25	1,14	1,3	4	0
24	138	538	capex_abn	2,0	1,08	1,08	1,2	1	0
25	144	544	emp_gr1	1,8	1,20	1,13	1,1	0	0
26	150	550	ebitda_mev	1,8	9,80	1,09	1,0	0	0
27	156	556	ni_me	1,8	1,80	1,09	1,0	0	0
28	162	562	qmj_safety	1,6	6,20	1,10	0,9	1	0
29	168	568	sti_gr1a	1,6	0,0003	1,08	0,8	0	0
30	174	574	lnoa_gr1a	1,6	0,0002	1,05	0,8	0	0

Table 3: Iterative Factor Selection with Expanding Window Nested Model

Iteration	Window End	Selected Factor	GRS Value	p(GRS)	Sh^2(f)	Sh^2(alpha)	Remaining Alphas (t > 2)	Remaining Alphas (t > 3)
0		Market						
1	400	nfna_gr1a	8,7	0,00	0,20	3,5	105	68
2	406	ival_me	7,3	0,00	0,35	3,3	98	76
3	412	resff3_12_1	5,9	0,00	0,51	2,9	55	29
4	418	cop_atl1	4,8	0,00	0,66	2,6	61	29
5	424	cowc_gr1a	4,2	0,00	0,77	2,4	22	10
6	430	seas_2_5an	3,6	0,00	0,82	2,1	17	3
7	436	debt_gr3	3,4	0,00	0,83	2	15	2
8	442	seas_6_10an	3,2	0,00	0,86	2	10	1
9	448	zero_trades_252d	3,2	0,00	0,85	1,8	28	1
10	454	zero_trades_21d	2,8	0,00	0,87	1,6	7	1
11	460	rmax5_rvol_21d	2,7	0,00	0,90	1,6	12	1
12	466	turnover_126d	2,6	0,00	0,92	1,5	10	1
13	472	ret_12_7	2,4	0,00	0,95	1,4	13	0
14	478	fcf_me	2,3	0,00	0,95	1,4	5	0
15	484	age	2,2	0,00	1,00	1,3	39	6
16	490	seas_11_15na	2	0,00	1,01	1,25	6	1
17	496	fnl_gr1a	2	0,00	1,04	1,22	4	0
18	502	seas_6_10na	1,8	0,00	1,05	1,1	3	0
19	508	nncoa_gr1a	1,8	0,00	1,07	1,1	5	0
20	514	noa_gr1a	1,7	0,00	1,09	1	11	1
21	520	op_at	1,6	0,00	1,12	1	7	0
22	526	zero_trades_126d	1,4	0,001	1,13	0,9	2	0
23	532	debt_me	1,4	0,00	1,13	0,8	0	0
24	538	ocf_me	1,4	0,01	1,11	0,8	3	0
25	544	ni_ar1	1,3	0,02	1,12	0,7	0	0
26	550	seas_16_20an	1,25	0,04	1,11	0,7	0	0
27	**556**	**saleq_gr1**	**1,23**	**0,06**	**1,12**	**0,7**	**0**	**0**
28	562	ni_me	1,29	0,09	1,11	0,6	0	0
29	568	qmj_safety	1,1	0,10	1,11	0,6	0	0
30	574	o_score	1,1	0,20	1,10	0,6	0	0

Table 4: Factor Weighting Scheme

	LASSO	Top 3 PCA Components			PLS Components				PLS with Elastic Nets Components				
Factors	**(1)**	**(1)**	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**	**(4)**	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
age		0,13	0,03	-0,01	-0,14	0,06	-0,03	-0,02	-0,20	0,12	0,02	-0,02	-0,02
aliq_at		-0,13	0,02	-0,04	0,14	-0,05	-0,05	-0,03	0,20	-0,08	-0,10	-0,06	-0,01
aliq_mat	-0,0018	0,08	-0,09	0,04	-0,08	0,11	0,07	0,13	-0,10	0,12	0,08	0,17	0,08
ami_126d	0,0221	0,04	0,11	-0,02	-0,05	-0,03	-0,16	-0,01	-0,08	0,02	-0,17	-0,05	-0,23
at_be	-0,0090	0,12	-0,02	-0,02	-0,12	0,12	-0,01	0,07	-0,17	0,18	0,03	0,07	-0,05
at_gr1		-0,12	0,06	-0,04	0,12	-0,09	-0,09	-0,06					
at_me		-0,11	0,08	0,07	0,10	-0,18	-0,06	-0,01	0,13	-0,23	-0,15	-0,06	0,01
at_turnover		0,01	-0,08	0,14	-0,01	-0,04	0,13	0,16	-0,02	-0,08	0,12	0,17	0,00
be_gr1a	0,0077	-0,11	0,10	-0,05	0,11	-0,10	-0,12	-0,12					
be_me		-0,10	0,10	0,09	0,09	-0,19	-0,11	0,03					
beta_60m		-0,10	-0,07	-0,10	0,13	0,10	-0,01	0,03	0,20	0,08	-0,03	0,05	0,01
beta_dimson_21d		-0,09	-0,06	-0,07	0,12	0,08	-0,03	0,06	0,18	0,07	-0,07	0,05	0,02
betabab_1260d	-0,0009	-0,09	-0,07	-0,10	0,12	0,12	-0,04	0,05					
betadown_252d	-0,0113	-0,11	-0,05	-0,07	0,13	0,05	-0,04	0,04					
bev_mev		-0,11	0,09	0,08	0,10	-0,19	-0,08	0,00					
bidaskhl_21d		0,11	0,10	0,03	-0,14	-0,03	-0,10	-0,02	-0,21	0,04	-0,08	-0,05	-0,04
capex_abn	0,0020	0,02	0,05	-0,14	-0,02	0,05	-0,07	-0,15	-0,03	0,11	-0,01	-0,15	-0,25
capx_gr1		-0,10	0,06	-0,07	0,11	-0,06	-0,07	-0,12	0,16	-0,08	-0,09	-0,14	-0,03
capx_gr2		-0,10	0,04	-0,09	0,11	-0,05	-0,05	-0,12	0,16	-0,08	-0,06	-0,12	-0,19
capx_gr3		-0,10	0,03	-0,08	0,11	-0,04	-0,04	-0,09					
cash_at		0,12	0,01	-0,03	-0,13	0,09	-0,03	-0,01	-0,19	0,16	0,02	0,00	0,05
chcsho_12m	0,0058	-0,12	-0,03	0,03	0,13	-0,07	0,03	0,02	0,18	-0,14	-0,01	0,02	0,00
coa_gr1a		-0,11	0,07	-0,08	0,11	-0,05	-0,12	-0,09	0,16	-0,06	-0,17	-0,13	-0,02
col_gr1a	0,0001	-0,12	0,07	-0,02	0,12	-0,09	-0,11	-0,04					
cop_at		0,00	-0,10	0,02	0,01	0,08	0,07	0,14	0,04	0,09	0,08	0,13	0,20
cop_atl1		0,07	-0,10	0,03	-0,06	0,12	0,07	0,15	-0,07	0,16	0,10	0,15	0,19
corr_1260d	-0,0021	-0,03	-0,02	-0,14	0,05	0,14	-0,10	0,01	0,08	0,20	-0,09	0,00	-0,10
coskew_21d		-0,01	-0,03	-0,09	0,02	0,07	0,02	-0,08	0,04	0,08	0,07	-0,06	0,00
cowc_gr1a	-0,0039	-0,01	0,04	-0,18	0,02	0,10	-0,12	-0,13	0,04	0,17	-0,08	-0,15	-0,02
dbnetis_at	0,0002	-0,03	-0,01	0,02	0,02	-0,06	0,08	-0,02	0,04	-0,11	0,11	-0,02	-0,33
debt_gr3	-0,0005	-0,02	-0,02	-0,09	0,03	0,07	-0,03	0,01	0,06	0,10	-0,02	0,01	-0,29
debt_me		-0,12	0,06	0,06	0,11	-0,16	-0,04	-0,02					
dgp_dsale	0,0023	0,02	-0,12	0,03	-0,01	0,08	0,14	0,10	0,00	0,08	0,18	0,13	0,12
div12m_me		-0,13	0,01	-0,01	0,14	-0,05	-0,05	0,01	0,19	-0,09	-0,12	-0,01	0,05
dolvol_126d	-0,0182	-0,03	0,09	-0,01	0,02	-0,05	-0,16	0,03	0,02	-0,03	-0,20	-0,02	-0,23
dolvol_var_126d	0,0021	-0,06	-0,09	0,11	0,06	-0,08	0,19	0,05	0,09	-0,17	0,16	0,08	0,10
dsale_dinv		0,02	-0,03	-0,10	-0,01	0,10	0,02	-0,07	-0,01	0,14	0,10	-0,07	-0,04
dsale_drec	0,0002	0,00	-0,05	-0,03	0,01	0,04	0,09	-0,04	0,02	0,00	0,12	0,01	-0,13
dsale_dsga		0,06	-0,07	0,01	-0,05	0,08	0,08	0,06	-0,07	0,10	0,12	0,09	-0,12
earnings_variability		-0,08	-0,07	0,00	0,10	0,01	0,06	0,05	0,15	-0,04	0,02	0,08	0,04
ebit_bev		-0,06	-0,16	0,11	0,07	0,00	0,20	0,17	0,12	-0,08	0,18	0,21	0,09
ebit_sale		-0,10	-0,10	0,02	0,12	0,01	0,11	0,06	0,18	-0,05	0,08	0,09	0,07
ebitda_mev		-0,12	0,02	0,11	0,11	-0,16	0,00	0,08	0,15	-0,22	-0,08	0,04	0,01
emp_gr1		-0,12	0,04	-0,07	0,13	-0,05	-0,06	-0,09	0,18	-0,08	-0,09	-0,11	-0,08
eq_dur		-0,11	0,06	0,08	0,11	-0,15	-0,07	0,04					
eqnetis_at		-0,12	-0,02	0,03	0,13	-0,08	0,04	0,01					
eqnpo_12m	-0,0084	-0,13	0,00	-0,01	0,14	-0,05	-0,03	0,02					
eqnpo_me		-0,13	0,00	0,00	0,14	-0,06	-0,03	0,01					
eqpo_me	-0,0097	-0,13	0,02	0,02	0,13	-0,08	-0,06	0,03					

f_score	-0,0005	-0,08	-0,12	-0,02	0,11	0,05	0,13	0,05	0,17	0,01	0,14	0,08	-0,06
fcf_me	0,0006	-0,10	-0,06	0,03	0,10	-0,07	0,12	-0,01	0,15	-0,15	0,11	0,02	-0,05
fnl_gr1a		0,00	-0,03	-0,05	0,01	0,03	0,04	0,01	0,02	0,04	0,08	0,02	-0,35
gp_at		0,05	-0,11	0,10	-0,05	0,05	0,13	0,17	-0,06	0,03	0,14	0,19	0,08
gp_atl1		0,09	-0,09	0,07	-0,09	0,08	0,11	0,15					
inv_gr1	0,0037	-0,09	0,07	-0,10	0,10	-0,04	-0,11	-0,13	0,14	-0,04	-0,12	-0,17	0,02
inv_gr1a	0,0040	-0,08	0,07	-0,12	0,09	-0,02	-0,11	-0,15	0,14	-0,01	-0,11	-0,19	0,01
iskew_capm_21d	0,0017	-0,06	-0,05	0,07	0,06	-0,04	0,07	0,08	0,09	-0,11	0,02	0,10	0,05
iskew_ff3_21d		-0,06	-0,04	0,10	0,06	-0,07	0,06	0,10	0,09	-0,14	0,00	0,12	0,00
iskew_hxz4_21d	-0,0024	-0,06	-0,03	0,11	0,06	-0,08	0,06	0,10	0,08	-0,16	-0,01	0,12	0,01
ival_me	-0,0053	-0,09	0,09	0,11	0,07	-0,17	-0,10	0,06					
ivol_capm_21d		-0,12	-0,07	-0,03	0,14	0,02	0,05	0,03					
ivol_capm_252d		-0,12	-0,07	-0,03	0,14	0,02	0,04	0,03					
ivol_ff3_21d	-0,0019	-0,12	-0,07	-0,03	0,14	0,01	0,04	0,03					
ivol_hxz4_21d	-0,0025	-0,12	-0,07	-0,03	0,14	0,02	0,05	0,02					
kz_index	-0,0101	-0,05	0,04	0,00	0,05	-0,07	-0,03	-0,02	0,07	-0,08	-0,04	-0,06	-0,14
lnoa_gr1a		-0,10	0,06	-0,05	0,10	-0,08	-0,06	-0,10	0,14	-0,11	-0,08	-0,12	-0,12
lti_gr1a	-0,0012	-0,07	0,01	-0,02	0,07	-0,05	0,00	-0,05	0,10	-0,08	-0,01	-0,06	-0,08
market_equity	-0,0082	0,05	0,10	0,04	-0,07	-0,07	-0,11	0,02					
mispricing_mgmt		-0,12	0,03	-0,04	0,13	-0,06	-0,05	-0,05					
mispricing_perf		-0,01	-0,19	-0,03	0,04	0,14	0,20	0,08	0,08	0,12	0,27	0,14	-0,01
ncoa_gr1a		-0,08	0,07	-0,06	0,08	-0,08	-0,07	-0,11	0,11	-0,10	-0,08	-0,14	-0,16
ncol_gr1a		0,06	0,07	-0,02	-0,07	-0,01	-0,07	-0,06	-0,11	0,03	-0,05	-0,07	-0,13
netdebt_me	-0,0103	0,12	-0,04	-0,03	-0,12	0,13	0,02	0,03					
netis_at		-0,12	-0,03	0,04	0,13	-0,09	0,06	0,03					
nfna_gr1a		0,00	-0,04	-0,07	0,00	0,06	0,04	-0,01	0,02	0,08	0,09	0,01	-0,31
ni_ar1	0,0024	0,03	0,00	0,00	-0,03	0,01	0,02	-0,02	-0,05	0,02	0,05	-0,01	0,09
ni_be	-0,0048	-0,08	-0,14	0,08	0,10	-0,01	0,20	0,11					
ni_inc8q	0,0023	-0,01	-0,17	0,04	0,03	0,09	0,19	0,12	0,05	0,05	0,22	0,18	0,07
ni_ivol		0,11	0,04	0,02	-0,13	0,00	-0,02	0,00	-0,19	0,05	0,02	-0,02	-0,10
ni_me		-0,12	-0,01	0,08	0,12	-0,10	0,00	0,08					
niq_at		-0,04	-0,18	0,05	0,06	0,07	0,22	0,14	0,10	0,01	0,23	0,20	-0,05
niq_at_chg1	-0,0004	0,02	-0,12	-0,10	0,00	0,13	0,15	-0,06	0,01	0,13	0,26	0,01	-0,21
niq_be	-0,0020	-0,06	-0,16	0,04	0,08	0,03	0,23	0,08					
niq_be_chg1	0,0014	0,00	-0,13	-0,10	0,02	0,13	0,16	-0,04					
niq_su		0,01	-0,12	-0,08	0,00	0,11	0,18	-0,06	0,03	0,10	0,29	0,00	-0,20
nncoa_gr1a		-0,10	0,06	-0,05	0,10	-0,09	-0,05	-0,10					
noa_at	-0,0026	0,02	0,04	-0,10	-0,01	0,05	-0,06	-0,12	-0,02	0,08	-0,03	-0,11	0,08
noa_gr1a	-0,0094	-0,11	0,06	-0,09	0,12	-0,05	-0,09	-0,11					
o_score	0,0014	-0,08	-0,13	0,06	0,10	0,01	0,15	0,12	0,16	-0,06	0,12	0,15	0,07
oaccruals_at		0,05	0,08	-0,10	-0,05	0,07	-0,16	-0,05	-0,07	0,16	-0,13	-0,11	0,09
oaccruals_ni		-0,05	0,06	-0,01	0,05	-0,08	-0,05	-0,06	0,07	-0,09	-0,06	-0,11	0,01
ocf_at		-0,07	-0,11	0,04	0,09	0,03	0,10	0,14	0,14	-0,01	0,08	0,14	0,11
ocf_at_chg1		0,05	-0,04	-0,09	-0,04	0,13	0,01	-0,03	-0,05	0,18	0,09	-0,02	-0,01
ocf_me	-0,0039	-0,12	0,01	0,08	0,11	-0,14	0,03	0,02					
ocfq_saleq_std	0,0051	-0,11	-0,07	0,04	0,12	-0,03	0,07	0,08	0,18	-0,10	0,03	0,09	0,07
op_at		0,01	-0,16	0,08	0,01	0,07	0,18	0,19	0,02	0,03	0,19	0,22	0,03
op_atl1	0,0037	0,05	-0,16	0,08	-0,04	0,10	0,18	0,19					
ope_be		-0,08	-0,12	0,10	0,09	-0,04	0,19	0,10					
ope_bel1		-0,04	-0,16	0,12	0,05	0,00	0,24	0,15	0,08	-0,08	0,24	0,20	0,06
opex_at		0,03	-0,04	0,10	-0,04	-0,03	0,07	0,11					
pi_nix	-0,0019	-0,02	-0,04	0,09	0,02	-0,06	0,07	0,09	0,03	-0,11	0,05	0,09	-0,07
ppeinv_gr1a		-0,11	0,05	-0,07	0,11	-0,06	-0,07	-0,10					
prc	0,0028	0,02	0,15	0,14	-0,05	-0,19	-0,13	0,05	-0,10	-0,20	-0,21	-0,02	-0,10
prc_highprc_252d		-0,05	-0,11	-0,19	0,09	0,19	0,05	-0,08	0,153	0,21	0,12	-0,04	0,03
qmj	-0,0090	-0,01	-0,18	0,00	0,04	0,16	0,14	0,18	0,077	0,15	0,16	0,22	0,16
qmj_growth	0,0041	-0,01	-0,13	0,02	0,03	0,10	0,12	0,11	0,055	0,09	0,13	0,14	0,22
qmj_prof		-0,03	-0,17	0,08	0,05	0,06	0,19	0,19	0,09	0,01	0,19	0,22	0,11

qmj_safety	-0,0062	0,00	-0,14	-0,08	0,03	0,21	0,05	0,13	0,064	0,23	0,06	0,17	0,06
rd5_at		0,10	0,00	-0,04	-0,10	0,10	-0,03	-0,01	-0,15	0,16	0,01	0,00	0,08
rd_me	0,0010	0,02	0,09	0,10	-0,04	-0,12	-0,09	0,04	-0,07	-0,12	-0,14	-0,01	0,12
rd_sale	-0,0014	0,11	0,02	-0,05	-0,12	0,10	-0,04	-0,03					
resff3_12_1	0,0021	-0,01	-0,06	-0,19	0,03	0,14	0,07	-0,18	0,064	0,16	0,18	-0,14	-0,09
resff3_6_1	-0,0007	-0,02	-0,04	-0,17	0,04	0,11	0,04	-0,17	0,082	0,12	0,12	-0,14	-0,02
ret_12_1		0,03	-0,10	-0,21	-0,01	0,21	0,08	-0,15	0,017	0,26	0,21	-0,09	-0,04
ret_12_7	0,0026	0,04	-0,08	-0,13	-0,03	0,15	0,10	-0,10	-0,02	0,17	0,21	-0,05	-0,11
ret_1_0	0,0056	0,01	0,04	0,13	-0,03	-0,12	0,00	0,05	-0,06	-0,16	-0,04	0,05	0,01
ret_3_1	-0,0010	0,01	-0,06	-0,18	0,01	0,16	0,04	-0,13	0,045	0,20	0,12	-0,10	0,08
ret_60_12		-0,04	0,12	-0,06	0,03	-0,08	-0,14	-0,12	0,038	-0,06	-0,16	-0,16	-0,22
ret_6_1	0,0020	0,02	-0,07	-0,21	0,00	0,19	0,04	-0,16	0,031	0,24	0,15	-0,11	0,04
ret_9_1		0,03	-0,09	-0,22	0,00	0,21	0,06	-0,15					
rmax1_21d		-0,12	-0,07	-0,02	0,14	0,02	0,03	0,04					
rmax5_21d		-0,12	-0,07	-0,02	0,14	0,02	0,03	0,05					
rmax5_rvol_21d	0,0011	-0,03	0,00	0,03	0,03	-0,03	0,00	0,02	0,032	-0,06	-0,02	0,03	0,05
rskew_21d		-0,08	-0,06	0,01	0,09	0,00	0,07	0,03	0,132	-0,06	0,06	0,05	0,04
rvol_21d		-0,12	-0,07	-0,04	0,14	0,03	0,03	0,04					
sale_bev	0,0105	0,06	-0,05	0,07	-0,06	0,02	0,07	0,11	-0,09	0,03	0,08	0,12	0,07
sale_emp_gr1		0,06	-0,04	0,01	-0,06	0,04	0,06	0,01	-0,08	0,07	0,11	0,02	-0,06
sale_gr1		-0,12	0,06	-0,02	0,12	-0,09	-0,09	-0,06					
sale_gr3	-0,0075	-0,11	0,05	-0,06	0,12	-0,06	-0,08	-0,08	0,168	-0,09	-0,11	-0,10	-0,13
sale_me	0,0036	-0,10	0,06	0,12	0,09	-0,20	-0,02	0,03					
saleq_gr1		-0,11	0,07	0,02	0,11	-0,13	-0,09	-0,02					
saleq_su	0,0020	0,08	-0,10	-0,07	-0,07	0,15	0,11	-0,02	-0,08	0,18	0,21	0,03	-0,14
seas_11_15an		0,01	-0,01	-0,02	-0,01	0,03	0,01	-0,01	-0,01	0,04	0,04	-0,02	0,02
seas_11_15na	0,0008	0,02	-0,01	-0,04	-0,01	0,07	-0,05	0,04	-0,01	0,11	-0,06	0,03	-0,05
seas_16_20an	-0,0035	0,00	0,02	0,02	-0,01	-0,03	-0,01	0,02	-0,02	-0,03	-0,02	0,01	-0,14
seas_16_20na	-0,0026	-0,01	0,01	-0,05	0,02	0,04	-0,08	0,03	0,031	0,09	-0,10	0,00	0,02
seas_1_1an	0,0018	0,03	-0,02	-0,01	-0,03	0,02	0,06	-0,04	-0,04	0,01	0,12	-0,02	-0,10
seas_1_1na	-0,0035	0,05	-0,08	-0,22	-0,02	0,22	0,06	-0,16					
seas_2_5an	0,0008	0,02	-0,02	0,06	-0,02	-0,01	0,04	0,06	-0,03	-0,01	0,03	0,06	0,17
seas_2_5na		-0,10	0,07	-0,05	0,10	-0,07	-0,11	-0,07	0,134	-0,09	-0,15	-0,11	-0,13
seas_6_10an	-0,0007	0,03	0,02	0,06	-0,04	-0,03	-0,03	0,08	-0,06	-0,01	-0,06	0,06	-0,06
seas_6_10na	-0,0002	-0,09	0,01	-0,06	0,10	0,00	-0,05	-0,03	0,147	0,00	-0,08	-0,04	0,06
sti_gr1a		0,02	-0,06	0,05	-0,02	0,02	0,10	0,07	-0,02	-0,01	0,12	0,09	-0,12
taccruals_at	-0,0016	0,03	0,12	-0,05	-0,03	-0,01	-0,19	-0,05	-0,06	0,07	-0,21	-0,12	0,21
taccruals_ni		-0,05	0,10	-0,01	0,04	-0,08	-0,13	-0,06	0,05	-0,06	-0,17	-0,13	0,19
tangibility	-0,0011	0,09	0,02	-0,08	-0,09	0,08	-0,04	-0,05	-0,13	0,16	0,02	-0,06	-0,11
tax_gr1a	-0,0048	0,07	-0,14	0,03	-0,06	0,10	0,20	0,09	-0,07	0,09	0,27	0,14	-0,13
turnover_126d		-0,12	-0,03	-0,03	0,14	0,01	-0,04	0,06					
turnover_var_126d	0,0083	-0,04	-0,09	0,09	0,04	-0,06	0,19	0,03					
z_score	-0,0091	0,11	-0,07	0,01	-0,11	0,12	0,06	0,09					
zero_trades_126d		-0,12	-0,03	-0,02	0,13	0,01	-0,04	0,07					
zero_trades_21d		-0,11	-0,03	-0,03	0,13	0,01	-0,04	0,06					
zero_trades_252d	-0,0012	-0,12	-0,03	-0,04	0,13	0,01	-0,04	0,06					

Table 5: Alpha Weighting Scheme

Alpha	Test_R2	Test_MSE
0,00001	0,6770	0,0007
0,00001	0,6788	0,0007
0,00001	0,6813	0,0007
0,00002	0,6847	0,0007
0,00002	0,6881	0,0007
0,00003	0,6913	0,0007
0,00004	0,6971	0,0006
0,00005	0,7016	0,0006
0,00006	0,7039	0,0006
0,00008	0,7049	0,0006
0,00011	0,7090	0,0006
0,00018	0,7135	0,0006
0,00026	0,7071	0,0006
0,00034	0,6981	0,0006
0,00045	0,6867	0,0007
0,00060	0,6804	0,0007
0,00079	0,6710	0,0007
0,00105	0,6572	0,0007
0,00139	0,6390	0,0008
0,00184	0,6112	0,0008
0,00244	0,5788	0,0009
0,00324	0,5444	0,0010
0,00429	0,5096	0,0010
0,00569	0,4700	0,0011
0,00754	0,4392	0,0012
0,01000	0,4000	0,0013

Table 7: Increase in Explained Variance – PCA

Component	Explained Variance	Cumulative Explained Variance
1	0,363	0,363
2	0,146	0,509
3	0,077	0,586
4	0,051	0,637
5	0,035	0,672
6	0,034	0,705
7	0,026	0,731
8	0,022	0,753
9	0,018	0,771
10	0,015	0,786
11	0,013	0,799
12	0,012	0,810
13	0,011	0,821
14	0,010	0,831
15	0,009	0,841
16	0,009	0,849
17	0,008	0,857
18	0,007	0,865
19	0,007	0,872
20	0,006	0,878
21	0,006	0,884
22	0,006	0,890
23	0,005	0,895
24	0,005	0,901
25	0,005	0,906
26	0,005	0,910
27	0,005	0,915
28	0,004	0,919
29	0,004	0,923
30	0,004	0,927
31	0,003	0,930
32	0,003	0,934
33	0,003	0,937
34	0,003	0,940
35	0,003	0,943
36	0,003	0,946
37	0,003	0,948
38	0,002	0,951

Figure 6: Number of PLS Components vs MSE & vs R2