BI Norwegian Business School - campus Oslo

GRA 19703

Master Thesis

Thesis Master of Science

Selecting characteristics using the Adaptive Group Lasso on U.S. industries

Navn:	Henrik Andreas Greve, Ivar Gjerstad Maseng
Start:	15.01.2020 09.00
Finish:	01.09.2020 12.00

Selecting characteristics using the Adaptive Group Lasso on U.S. industries

Ivar G. Maseng Henrik A. Greve

Supervisor: Patrick Konermann. MSc Business with Major Finance.

01.09.20

Abstract

Throughout the years, hundreds of factors have been proposed to forecast stock returns. Cochrane (2011) referred to these factors as the "zoo of new factors." In this thesis, we consider 62 of these factors and analyze which of them provide incremental value when forecasting stock return in 12 U.S industries. We apply the Adaptive Group Lasso (AGL) method for model selection described by Freyberger, Neuhierl, and Weber (2018), and use the Classical Linear Regression Model (CLRM) as a benchmark. The AGL selects, on average, approximately three characteristics, while the linear approach selects 24. The results indicate that the AGL approach generates more accurate predictions when the sample size increases compared to the CLRM. Our analysis indicates that there is no superior method for model selection in our samples.

Acknowledgments

This thesis ends our journey at the Master of Science program in Business with major in Finance at BI Norwegian Business School. We want to take this opportunity to thank BI and Patrick Konermann for facilitating and guiding us through our thesis. Finally, we would like to thank Freyberger, Neuhierl, and Weber for being a great inspiration.

List of Abbreviations

- AGL: Adaptive Group Lasso
- CAPM: Capital Asset Pricing Model
- CLRM: Classical Linear Regression Model
- MSE: Mean Squared Error

List of Figures

List of Tables

Contents

1 Introduction

How to predict stock return has always been one of the biggest conundrums within asset pricing. An extensive number of researchers, investors, academics, mathematics, and financial professionals have tried to answer this question by creating hundreds of factors. In the past decades, academics have faced a crossroads, where some deviate from the linear approach following the path of nonparametric methods for model selection. Historically, the majority of asset pricing theories have applied some variation of the Classical Linear Regression Model (CLRM) when attempting to forecast stock return. Since most of these factors are combinations of the companies' balance sheet and trading data, a potential problem with CLRM occurs when looking at many explanatory variables were some are highly correlated. This issue is known as (near) multicollinearity. The likelihood that hundreds of factors have a significant impact on security prices is rather slim. There is a high possibility that most of these factors are redundant and do not provide incremental value.

We address a topic of particular interest for investors, funds, or investment banks, as it allows them to identify characteristics that provide incremental information. The ability to recognize factors that drive return will help broaden the understanding of the industry's underlying mechanics and the market movements. This analogy also applies to academics trying to examine industries or attempting to tame the zoo of factors. Equally important, this thesis evaluates statistical methods that offer professionals across industries insight that can lead to more precise forecasts.

We follow the method of Freyberger et al. (2018) and use the cross-sectional model designed by Lewellen (2015) as a framework, combined with the Classical Linear Regression Model and the Adaptive Group Lasso approach described by Huang, Horowitz, and Wei (2010). We employ the proposed methods on 62 characteristics in order to answer the following research questions:

- Which characteristics provide incremental value when forecasting return in US industries using the Adaptive Group Lasso?
- How does the Adaptive Group Lasso approach for model selection perform out-of-sample, compared to the Classical Linear Regression Model?

The first step of our analysis is to obtain equal results as Freyberger et al. (2018). We achieve more or less identical results. The only distinction from the article we replicate is the difference in selecting BEME as described in Chapter 5. We are confident that this minor deviation does not affect our computations. This might be a consequence of the data collection process or the difference in the number of simulations. We are, therefore, convinced that the approach is correct.

The second step of our thesis is to utilize the same methodology to analyze industries sorted by the Fama-French 12 Industry Classification (Figure 2, Appendix). This gives us valuable insight into which characteristics that describe stock return. We observe that the Adaptive Group Lasso selects three characteristics on average, while the Classical Linear Regression Model selects 23. We have compared the two models output using the mean square error, presented in Chapter 6.

2 Literature

There have been numerous attempts to construct the best model when forecasting stock returns. Perhaps the most prominent attempt is the model constructed by Sharpe (1964), Lintner (1965), and Mossin (1966), the Capital Asset Pricing Model (CAPM).

$$R_i = R_f + \beta \left(E \left(R_m \right) - R_f \right) \tag{1}$$

The model argues that an asset's return is determined by the degree of exposure to systematic risk, scaled by its beta. Fama and Macbeth (1973) examined the CAPM's validity in a systematic review, testing the cross-sectional return on all assets listed on NYSE from 1926-1968. Their findings supported that expected returns tend to increase with the beta, as well as the fact that non-systematic risk does not affect the excess returns. However, they found evidence disputing the model, arguing that the proposed Security Market Line was too flat, and the intercept was non-zero. This resulted in Fama and Macbeth rejecting the theory.

In the turmoil of the CAPM, the Arbitrage Pricing Theory (APT) was formulated by Ross (1976, 1982), and later extended by Connor (1981), Huberman (1982), and Ingersoll (1982). The APT proposes a linear approximation of pricing relationship among assets, arguing that an asset's expected return can be linearly described through its sensitivity to variations in theoretical factors. As the APT gives no guidance in which factors to use, hundreds of papers have attempted to construct the best predicting factor models. Harvey, Liu, and Zhu (2016) provide an overview of over 300 previously published factors. The result of the review suggests that approximately 150 of these are significant, even after the problem of multiple comparisons is taken into consideration. Cochrane (2011) refers to the numerous attempts to construct explanatory factors as "a zoo of new factors."

Chen, Roll, and Ross (1986) found evidence supporting that industrial production, expected inflation, unanticipated inflation, excess return on long-term bonds over short-term government bonds, and excess return on long-term government bonds over T-bills are the best predictors for stock return. Fama and French (1992) found that future stock return could be predicted based on the market return, the return of a portfolio of small stocks in excess of the return on a portfolio of large stocks, and the return of a portfolio of stocks with a high book-to-market ratio in excess of the return on a portfolio of stocks with a low book-to-market ratio. Other noteworthy factors are Momentum (Carhart, 1997), Stock Market Liquidity (Pastor & Stambough, 2003; Acharya, 2005), Stock Market Volatility (Hodrick et al., 2006), Betting Against Beta (Frazini & Pedersen, 2013), Quality Minus Junk (Asness, Fazzini & Pedersen, 2013), and Dealers banks' Financial Constraints (Adrian, Eutela & Muir, 2014).

The previously mentioned authors generally isolate the return predictor in their respective models, with the absence of conditioning based on already discovered return predictors. Haugen and Baker (1996) and Lewellen (2015) are two expectations: they do not isolate the return predictors. The introduction of these two was instrumental in discovering findings questioning the Efficient Markets Hypothesis's plausibility, which is a criterion for the APT. They both used the regressions from Fama and Macbeth (1973) to gather information on multiple characteristics. Haugen and Baker (1996) discovered conclusive evidence that stocks with low returns will have lower risk than stocks with higher expected and realized rates of return. They also found that the most crucial determinants of expected stock returns are unexpectedly equal to the world's major equity markets. Lewellen (2015) created a cross-sectional model to estimate how 15 characteristics and the possible composition of these could represent a stock's expected return. The result was that only a small number of the predictors of expected return were considered significant when analyzing the jointly predictive power of these 15 characteristics.

In more recent years, several authors propose model selection based on various statistical and economic theories using penalized regressions and a nonparametric model approaches (Horowitz 2016; Huang et al., 2010). Huang and Shi (2016) used the supervised Adaptive Group 'Least Absolute Shrinkage and Selection Operator" (Lasso) for model selection to test determinants of bond risk premia. They found that they could discover a single macro factor that is far more significant and relevant than macro factors from already existing literature. This is consistent with the paper written by Chinco, Clark-Joseph, and Ye (2018), which concludes that their model constructed through the Lasso approach, increased the forecast-implied Sharpe ratios. It also improves the out-of-sample fit, which can be explained by the fact that the "identifying predictors are unexpected, short-lived and sparse" (Chinco, Clark-Joseph & Ye, 2018). Li and Chen (2015) tried to forecast macroeconomic time series using Lasso, where they concluded that the Lasso approach reduced the mean square error. On the other hand, Zou and Hastie (2005) found that Lasso tends to have problems when the characteristics are highly correlated. They also criticize Lasso in cases where the variables are structured in clusters. In such a case, the model selects only one variable from each group, while ignoring the others. Even though Lasso was initially developed as a statistical tool in geophysical analysis, the approach seems to recognize stock predictors based on fundamental news.

Several papers have examined the impact of industry affiliation and expected return. Among them, Fama and French (1988) created an industry classification based on Standard Industry Classification (SIC) codes to create 17 industry portfolios, which was later extended in 1997. They also created numerous other industry classifications, ranging from 5 up to 49. All of these classifications contains distinct industry portfolios generated through the use of four-digit SIC codes (Fama & French, 1997). We use the Fama & French 12 industry classification, due to its size, transparency, and academic recognition.

3 Methodology

3.1 Model selection using Adaptive Group Lasso

In our thesis, we will extend the nonparametric method for model selection applied in the paper "Dissecting Characteristics Nonparametrically" written by Freyberger, Neuhierl and Weber (2018)[1]. They combine fundamental theory related to asset pricing and the Adaptive Group Lasso procedure described by Huang et al. (2010). Lasso is a regression analysis method used for regularization and variable selection (Tibshirani, 1996). Lasso's main advantage is that it helps reduce overfitting and is particularly useful for the selection of characteristics, especially in cases where we have several characteristics that do not contribute to the prediction. Lasso is almost identical to Ridge regression, but the motivation of using Lasso instead of Ridge Regression is that the penalty term is not squared. In other words, it can only include varying functions while eliminating constant and irrelevant functions by setting them equal to 0.

The computations in this thesis are written in R due to its ability to handle significant amounts of data using minimal storage memory. To use the functions, which we will describe in the following sections, we are required to install the packages' data.table', 'metrics', 'OEM' and 'stringr'.

Before we dive into the analysis, we create our characteristics (Table 9, Appendix). We transform the characteristics into normalized and orthonormal splines on an even quantile grid. Friedman (1991) describes splines as a function that is defined piecewise as a polynomial function, between prede-

[1]Since we are replicating the method used by Freyberger et al. (2018), all formulas in this section is retrieved or inspired by the original article.

termined knots[2]. There is no theory to support the use of a specific number of interpolation points. Anyhow, research suggests that a larger sample requires a larger number of splines, contrary to a small sample where fewer interpolation points are needed. (Wang & Tian, 2017). To determine the optimal number, we run the regression with 5, 10, 15, and 20 interpolation points to test the number of splines which estimates the most consistent selection of characteristics.

In order to categorize them as orthonormal, all splines have length 1 and are 0 when multiplied with another characteristic spline. This allows us to create and manage composite forms and surfaces through an extensive number of points (Talebitooti et al., 2015). There are two main reasons we normalize the characteristics; (1) We assume the characteristics might be exposed to skewness as a result of the inflation, and (2), due to Cochrane (2011), the sample will be less reactive to outliers. Freyberger et al. (2018) suggest a procedure to normalize the characteristics, which rank transform the characteristics from absolute sizes to relative sizes in the interval $C_{s,it-1} \in [1, 0]$ by using the following formula:

$$F_{s,t}(C_{s,it-1}) = \frac{rank(C_{s,it-1})}{N_t + 1} \tag{2}$$

In this case, $R\left[min_{i=1,\ldots,N_t}, C_{s,it-1}\right] = 1$ and $R\left[max_{i=1,\ldots,N_t}, C_{s,it-1}\right] = N_t$ (Freyberger et al., 2018). Freyberger et al. (2018) uses this transformation for portfolio sorting.

After normalizing the characteristics, the next step is to model the expected return. Freyberger et al. (2018) formulate return as an expression of

[2]These knots are predetermined actual numbers, with an equal number of observations between each knot. The higher number of knots gives a more realistic picture but doesn't necessarily describe the characteristics' overall trends .

the rank-transformed characteristics from the previous period, $\tilde{C}_{s,it-1}$, and the unknown function, $\tilde{m}_s(\cdot)$:

$$R_{it} = \sum_{s=1}^{S} \tilde{m}_{ts}\left[\tilde{C}_{s,it-1}\right] + \varepsilon_{it}, \qquad i = 1, 2, ..., n. \tag{3}$$

As an opposition to classical linear portfolio sorting, where $\tilde{m}_t$ are assessed with an constant (Chen, Roll & Ross, 1986; Fama & French, 1992; Carhart, 1998), Freyberger et al. (2018) estimates $\tilde{m}_t$ by using quadratic splines[3] over the interval of $\tilde{I}_l$. To obtain an unique estimation, Freyberger et al. (2018)[4] assumes that $0 = t_0 < t_1 <, ..., < t_l = 1$ is a series of ascending numbers in the interval of $[0, 1]$, equal to the portfolio breakpoints. $\tilde{I}_l$ for $l = 1, ..., L$ is a parition of the unit interval, that is; $\tilde{I}_l = [t_l, t_1]$ for $l = 1, ..., L-1$ and $\tilde{I}_L = [t_{L-1}, t_L]$. $t_0, ..., t_{L-1}$ are knots, and select $t_l = l/L$ for $l = 0, ..., L-1$. Hence, approximation of the unknown function, $\tilde{m}_{ts}$, is done by the following:

$$\tilde{m}_{ts} \approx \sum_{k=1}^{L+2} \beta_{tsk} p_k\left(\tilde{c}\right) \tag{4}$$

Both the numbers of intervals L and portfolios are user-specified, while $P_k(c)$ is a known basis function[5]. The Adaptive Group Lasso in nonparametric additive models has a two-step framework, based on spline representations of the factors in the underlying model (Huang et al., 2010). The first step consists of using the standard Group Lasso and allows us to attain an initial estimator of the nonparametric components. To estimate the coefficients, the

[3]Spline degree: $k-1$, where k is the number of variables in the spline function. Quadratic splines is splines of second degree.

[4]This assumption is built on the findings by Stone (1985), that was reformulated by Huang et al. (2010).

[5]A basis function is an element of the given splines.

model solves the following Lagrangian function in order to minimize BIC:

$$\breve{\beta}_t = \underset{b_{sk}:s=1,\ldots,S;k=1,\ldots,L+2}{arg\,min} \sum_{i=1}^{N}\left(R_{it} - \sum_{s=1}^{S}\sum_{k=1}^{L+2} b_{sk}p_k\left(\tilde{C}_{s,it-1}\right)\right)^2 + \lambda_1 \sum_{s=1}^{S}\left(\sum_{k=1}^{L+2} b_{sk}^2\right)^{1/2} \tag{5}$$

where λ_1 is the penalty parameter, that is, the amount of shrinkage towards the central point (Fang & Tang, 2013). We choose the λ_1 that minimizes the Bayesian Information Criterion (BIC) (Yuan & Lin, 2006),

$$BIC\,(\lambda) = log(RSS_\lambda) + (degrees\ of\ freedom) * \frac{log\ n}{n} \tag{6}$$

given the constraints of:

$$\sum_{k=1}^{L+2} b_{sk}p_k\left(\tilde{C}_{s,it-1}\right) = 0, \qquad 1 \leq s \leq S \tag{7}$$

At this point, we have created a Group Lasso model. What differentiates the Group Lasso and Adaptive Group Lasso is the extension described in the remaining part of this section. The first part of the extension is to use the Group Lasso estimator $\breve{\beta}_t$ to attain weights using:

$$w_{ts} = \begin{cases} \left(\sum_{k=1}^{L+2} \tilde{\beta}_{sk}^2\right)^{-1/2} & if \quad \sum_{k=1}^{L+2} \tilde{\beta}_{sk}^2)^{-1/2} \neq 0 \\ \infty & if \quad \sum_{k=1}^{L+2} \tilde{\beta}_{sk}^2)^{-1/2} = 0 \end{cases} \tag{8}$$

These weights prevents characteristics that were not selected in the Group Lasso, to be added in the next step (Huang et al., 2010).

In the second step, the Adaptive Group Lasso is applied to obtain consistent selection of characteristics.

$$\breve{\beta}_t = \underset{b_{sk}:s=1,\ldots,S;k=1,\ldots,L+2}{arg\,min} \sum_{i=1}^{N}\left(R_{it} - \sum_{s=1}^{S}\sum_{k=1}^{L+2} b_{sk}p_k\left(\tilde{C}_{s,it-1}\right)\right)^2 + \lambda_2 \sum_{s=1}^{S}\left(w_{ts}\sum_{k=1}^{L+2} b_{sk}^2\right)^{1/2} \tag{9}$$

where we choose λ_2 that minimizes BIC.

3.2 Model selection using Classical Linear Regression Model

We apply the Classical Linear Regression Method for model selection to create a benchmark for the Adaptive Group Lasso approach. We run the two regressions to achieve comparable results, as we wish to determine which model selects the best-fitting number of characteristics. The characteristics are normalized using the same procedure as the Adaptive Group Lasso, as described in 3.2 (2). The first step of the Classical Linear Regression Model is to run the following linear regression.

$$R_i = \alpha + \sum_{s=1}^{S} \beta_s C_{s,i} + \epsilon_i \tag{10}$$

After that, we conduct a step-wise regression using backward elimination. We use the "step" function in combination with the specification "backward elimination" in R. The approach begins with a regression including all 62 variables, proceeding to test if the removal of one of the characteristics increases or reduces the information criterion (AIC). The end goal is to achieve a final state where any characteristics' removal or change will increase AIC.

There are several potential pitfalls when dealing with CLRM. First, all the data is extracted from the company's balance sheet and trading data. This data is most likely influenced by many of the same underlying factors; increasing the probability of multicollinearity among the factors. Further, the linear regression is sensitive to outliers. This issue is combated when utilizing splines in the AGL approach. Lastly, Freyberger et al. (2018) found that a linear approach can be prone to overfitting during model selection. In the event of overfitting, characteristics that does not necessarily provide incremental value to the forecast of stock returns are included.

3.3 Measuring the performance of the models

Before the analysis, we divided the samples into two subsets; train sample and test sample (in-sample and out-of-sample); to avoid any bias in the samples. The train samples are applied when creating the models, and the test samples are used to validate the models' performance. 80% of the samples are utilized in model construction, and the remaining 20% of the samples are devoted to cross-validation.

To correctly select the model of highest relevance, we estimate the Mean Squared Error (MSE) for the CLRM and AGL for the test sample on the 12 industries. The MSE describes the mean squared difference between the actual and the estimated value. This estimate provides us a measure of how accurate our model selection is. We use the following function to compute this measurement:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{11}$$

4 Data

We retrieve our data from Wharton Research Data Services (WRDS), within the time-frame July 1965 to June 2014. We apply filters, common US stocks traded on NYSE, Amex, or Nasdaq. We will account for survivorship bias, including active and inactive companies listed in a time period of a minimum of two years. This criterion is created to obtain a representative sample of the market (Garcia & Gould, 1993). Our data file is a merged result of the following files:

Security Monthly	CRSP/Compustat	Monthly
Fundamentals Annual	CRSP/Compustat	Annual
Beta Suite	WRDS (Beta)	Daily
Financial Ratios Firm Level	WRDS (Beta)	Annual
12 Industry Classfication	Kenneth R. French	

We apply the same data as Freyberger et al. (2018) in our 12 industries analysis, with the corresponding time frame, 1965-2014. We aim to obtain an identical and coherent sample to correctly compare results from the full market, with the industries. The stock return is the dependent variable, while the characteristics are the independent variables. The characteristics are either product of trading data, balance statements, or a combination of both. We follow the framework presented by Hou, Xue, and Zhang (2015). A simple overview of all the characteristics with an explanation is presented in Table 9 (Appendix), with the descriptive properties in Figure 1 (Appendix). The four-digit SIC codes are categorized using the Fama & French 12 Industrial Classification (1997).

Our industry classification is the only segmentation we conduct on our data. Freyberger et al. (2018) create categories, where they exclude firms with a size below 10th and 20th percentile of NYSE firms. The 12 industry average of observations is approximately 150 000, and the article we replicate has, as previously mentioned, approximately 1.6 million observations. This substantial difference in sample size is why we do not divide our sample any further than into industries.

5 Validation of the model

We are confident that our sample is consistent with Freyberger et al. (2018), due to the similar sample size and characteristics statistics (Figure 1, Appendix). Furthermore, we followed their approach step-by-step when extracting data and utilized the same source (WRDS). To ensure that our model is correct, we compare model selection for five outputs reported by Freyberger et al. (2018);

Table 1: **Outputs reported by Freyberger et al. (2018)**

Firms	All	All	All	All	All
Sample	Full	Full	Full	1965-1990	1991-2014
Knots	20	15	25	15	15
Sample size	1,6m	1,6m	1,6m	0.6m	1m
# Selected	13	16	13	11	14

We achieve identical results with both 20 and 25 interpolation points as Freyberger et al. (2018) for the longest sample period. We found that $\Delta Shrout$, ΔSO, $Investment$, LME, $Lturnover$, PM_{adj}, r_{2-1}, r_{12-2}, r_{12-7}, $Rel2high$, ROC, SUV and $Totalvol$, provides incremental value. When allowing for a wider grid, with 15 knots, our model does not select $BEME$, as opposition to Freyberger et al. (2018). We obtain identical results as Freyberger et al. (2018) for both the half-samples when using 15 knots.

Table 2: **Our validating results**

Firms	All	All	All	All	All
Sample	Full	Full	Full	1965-1990	1991-2014
Knots	20	15	25	15	15
Sample size	1,6m	1,6m	1,6m	0.6m	1m
# Selected	13	15	13	11	14

6 Results

This section will report the selected characteristics for each industry and the out-of-sample mean error for the obtained models. There are no explicit theories related to the correct number of interpolation points, but there is consensus amongst academics that the optimal number of knots depends on the sample size. Hence, we apply four distinct variations in interpolation points; 5, 10, 15, and 20. We observe a clear correlation between the number of observations and the number of selected characteristics. Accordingly, we divide the industries into three subcategories determined by sample size:

- Small industries (0 - 100 000 observations)
- Medium industries (100 000 - 200 000 observations)
- Large industries (200 000 + observations)

Table 3: **Out-of-sample:** Adaptive Group Lasso; Small industries

Industry	Knots	Sample Size	Avg. No of Characteristics
2. Consumer Durables	5, 10, 15, 20	52 214	3
4. Energy Oil	5, 10, 15, 20	71 560	2.5
5. Chemicals and Allied Products	5, 10, 15, 20	49 468	1.25
7. Telephone and TV	5, 10, 15, 20	32 891	1
8. Utilities	5, 10, 15, 20	67 537	1.75
Total average		54 734	1.9

In the small industries, we obtain a sample with an average of 54 734 observations. We see that the Adaptive Group Lasso model selects an average of 2.1 characteristics, which is 18.7 less than the Classical Linear Regression Model that selects 20.8 (Table 3-4). An overview of the most significant characteristics obtained from the AGL approach is presented in Figure 3 (Appendix).

Table 4: **Out-of-sample:** Classical Liner Regression Model; Small industries

Industry	Sample Size	No. of Characteristics selected
2. Consumer Durables	52 214	18
4. Energy Oil	71 560	24
5. Chemicals and Allied Products	49 468	22
7. Telephone and TV	32 891	15
8. Utilities	67 537	25
Average	54 734	20.8

The apparent trend is that the CLRM model quite consistently out-performs the AGL model when observing smaller samples. This argument's basis is that the CLRM has a better MSE in 2 of 5 industries and better than two or more interpolation points in the other three industries. We see that the AGL chooses between four characteristics, where the lagged one-month return (r_{2-1}) and market capitalization (LME) appears as the most significant.

2 Consumer Durables		**4 Energy Oil**		**5 Chemicals and Allie**		**7 Telephone and TV**		**8 Utilities**	
Model	*MSE*	*Model*	*MSE*	*Model*	*MSE*	*Model*	*MSE*	*Model*	*MSE*
5 knots	0,011992	5 knots	0,012489	5 knots	0,008631	5 knots	0,014558	5 knots	0,009159
10 knots	0,011370	10 knots	0,011145	10 knots	0,008670	10 knots	0,010424	10 knots	0,009615
15 knots	0,011439	15 knots	0,011030	15 knots	0,009203	15 knots	0,010214	15 knots	0,009373
20 knots	0,009560	20 knots	0,009894	20 knots	0,009319	20 knots	0,009375	20 knots	0,012188
LM	0,008317	LM	0,010671	LM	0,008074	LM	0,013707	LM	0,009486

The table above reports the out-of-sample MSE for the small industries, where we notice an evident disparity between strong MSE values, appropriate model, and the number of knots. Chemicals and Allied Products has the second-lowest number of observations. This industry is particularly interesting as the CLRM selects 22 characteristics, whereas the AGL only chooses a maximum of two. Comparing the two models, none of the characteristics selected are identical. The models have identified completely different characteristics that provide incremental information to the forecast of stock returns. The CRLM has a lower mean squared error than the AGL approach, regardless of the number of knots. In all essence, this heavily implies that the CLRM is the correct model for this specific industry to obtain an accurate forecast. Inspecting 5 and 10 knots, we observe a close to equal MSE between the two models. The mentioned knots only select one characteristic, namely the lagged one-month return (r_{2-1}). This might raise the question of overfitting due to the considerable difference in chosen characteristics.

Table 5: **Out-of-sample:** Adaptive Group Lasso; Medium industries

Industry	Knots	Sample Size	Avg. No of Characteristics
1. Consumer Nondurables	5, 10, 15, 20	121 134	3.75
9. Wholesale and retail	5, 10, 15, 20	178 114	3.75
10. Healthcare	5, 10, 15, 20	130 898	3.5
12. Other	5, 10, 15, 20	180 352	3.5
Total average		152 624.5	3.625

Table 6: **Out-of-sample:** Classical Liner Regression Model; Medium industries

Industry	Sample Size	No. of Characteristics selected
1. Consumer Nondurables	121 134	18
9. Wholesale and retail	178 114	23
10. Healthcare	130 898	25
12. Other	180 352	29
Total average	152 624.5	23.75

The medium industries have an average of 3.6 characteristics when estimated through the AGL model. The CLRM model selects 23.75 observations on average, with an mean sample size of 152 624.5. Figure 4 (Appendix) shows an overview of the 9 characteristics chosen by AGL in the medium industries. The most frequently selected characteristics are the standard unexplained volume (SUV), the lagged one-month return (r_{2-1}) and market capitalization (LME).

1 Consumer NonDurable		9 Wholesale and Retai		10 Healthcare		12 Other	
Model	*MSE*	*Model*	*MSE*	*Model*	*MSE*	*Model*	*MSE*
5 knots	0,008820	5 knots	0,011794	5 knots	0,010305	5 knots	0,009643
10 knots	0,009602	10 knots	0,012098	10 knots	0,010381	10 knots	0,009493
15 knots	0,009671	15 knots	0,010085	15 knots	0,010719	15 knots	0,011439
20 knots	0,011377	20 knots	0,008826	20 knots	0,011377	20 knots	0,009013
LM	0,009996	LM	0,011971	LM	0,010148	LM	0,009202

We observe that the CLRM outperforms the AGL approach for all knots in the Healthcare industry, selecting 25 characteristics. These results affirm that the superior model in this industry is the CLRM. The AGL model obtains a lower MSE in 58 % of the three remaining industries. Despite this, we cannot identify a definite trend for medium industries.

Table 7: **Out-of-sample:** Adaptive Group Lasso; Large industries

Industry	Knots	Sample Size	Avg. No of Characteristics
3. Manufacturing Machinery	5, 10, 15, 20	240 537	4
6. Business Equipment	5, 10, 15, 20	257 930	4.5
11. Money Finance	5, 10, 15, 20	225 793	3.5
Total average		241 420	4

Table 8: **Out-of-sample:** Classical Liner Regression Model; Large industries

Industry	Sample Size	No. of Characteristics selected
3. Manufacturing Machinery	240 537	24
6. Business Equipment	257 930	25
11. Money Finance	225 793	31
Total average	241 420	26.67

The AGL selects, on average, four characteristics on a mean sample size of 241 420 observations in the large industries, while the CLRM selects 26.67. In addition to the three previously mentioned characteristics, closeness to the 52 weeks high (rel_to_high_price) appears to be of significance in most industries.

3 Manufacturing Machir		6 Business Equipment		11 Money Finance	
Model	*MSE*	*Model*	*MSE*	*Model*	*MSE*
5 knots	0,008946	5 knots	0,011982	5 knots	0,011552
10 knots	0,009917	10 knots	0,010954	10 knots	0,012043
15 knots	0,009897	15 knots	0,011509	15 knots	0,009671
20 knots	0,009013	20 knots	0,011215	20 knots	0,009560
LM	0,009228	LM	0,011346	LM	0,013280

For two industries, Manufacturing Machinery and Business Equipment, neither the AGL nor the CLRM seems to exceed one another when considering the MSE. In the Money Finance industry, we observe that the AGL outperforms the CLRM, as it achieves lower MSE value for all of the knots in the entire sample. This, combined with the fact that the AGL model selects 27.5 fewer characteristics, indicates that the CLRM is prone to overfitting in this industry.

The analysis is conducted to obtain a more detailed understanding of the fundamental characteristics of each industry. We initially believed that the characteristics that describe capital structure would appear of significance when analyzing industries separately. This turned out not to be accurate, despite that Brealey, Myers & Allen (2019) found that banking services have four times higher debt-to-value ratio than pharmaceutical companies. We also notice that characteristics based on return and market capitalization appear to be of higher significance when analyzing the industries in separation.

Another aspect of the analysis and the corresponding results is that the characteristics selected are coherent with the factors chosen by Freyberger et al. (2018). In total, eight of the nine characteristics selected by the AGL approach are identical. Further, the average number of characteristics selected by CLRM compared to Freyberger et al. (2018) are in proximity to our results, with only 2.84 characteristics separating them. When running the CLRM analysis, we obtain an average of 23.74 characteristics for all the industries, while Freyberger et al. (2018) obtain 26.58 characteristics for the entire market.

As a general remark, we see that the number of observations heavily influences the number of characteristics selected. When the sample size grows, the number of characteristics selected increases. This might be one potential explanation behind the apparent trend in the model selection of the industries. When analyzing small industries, it becomes apparent that the CLRM eclipse the AGL approach, with some notable exceptions. This might be because the linear model selects more characteristics than the AGL approach regardless of sample size, which might again influence the model's performance. In medium industries, we observe more nuanced results. In two industries, the CLRM dominates and obtains a much better MSE than all the knots related to AGL. Contrarily, the two remaining industries in this selection is heavily dominated by a strong MSE (3/4 knots has a better MSE than CLRM in both industries) for the AGL model, which implies that the model selection in these industries, converges towards a more or less equal divide between the CLRM and the AGL. For the large industries', the results give an impression of a trend where the AGL is the predominant approach for model selection.

7 Conclusion

The likelihood that the entire "zoo of factors" has a significant impact on security prices is rather slim. We seek to answer which of 62 characteristics provide incremental value in the forecast of return using the Adaptive Group Lasso. There are a few dominant and recurring characteristics that are selected. Our analysis shows that the most frequently selected characteristics are the lagged one-month return (r_{2-1}), market capitalization (LME), and standard unexplained volume (SUV). This is coherent with the results obtained by Freyberger et al. (2018). Nonetheless, our model selects fewer characteristics than the article we replicate. The most likely explanation being sample sizes. When examining Table 3-8, this becomes evident as we observe a correlation between sample size and selected characteristics.

When assessing the quality of the out-of-sample model selection, we use the MSE to evaluate how well the AGL and CLRM performs. If we select all the best MSE values for the AGL, it will outperform the CLRM in 10 of 12 industries. This approach is not viable, since there is no theoretical framework highlighting the preferable amount of knots. We do not observe a consistently superior model as the MSE of the two methods fluctuate. On average, we see that the CLRM obtains a relatively consistent MSE for all the examined sample sizes. When the sample size is large enough, we observe that the Adaptive Group Lasso approach selects more characteristics with incremental value to the forecast of returns, which also have an enhanced mean squared error.

Our thesis can be viewed as a starting point for future research. One possible extension would be to compare full markets or industries from different countries (i.e., London Stock Exchange). In order to determine if the same characteristics are significant, regardless of country. This proposes a challenge since there are a few characteristics that are entirely based on the US market and require modification. An alternative approach would be to use a smaller or lager Industry Classification provided by Fama-French. This would potentially uncover even more industry-specific characteristics.

One limitation of our thesis is that we do not apply any filters based on market capitalization. Freyberger et al. (2018) exclude the lowest 10th and subsequent 20th percentile when conducting their out-of-sample simulations. A possible expansion of our thesis could be to analyze the industries small companies and large companies, before comparing their results. This topic has been analyzed using the CLRM, but not the AGL approach. Therefore it would be interesting to examine how the AGL approach of selecting characteristics compares to the CLRM, and examine if the approach diverges from extant theory, something that is highly plausible.

Lastly, it would be insightful to conduct an analysis with an extended number of industry-specific characteristics, i.e., spot prices on raw materials. Some industries might be driven by factors not present on a balance sheet, nor in the trading data.

8 Bibliography

References

[1] Acharya, V. V. & Pedersen, L. H. (2005). Asset pricing with liquidity risk. Journal of Financial Economics, pp.375–410.

[2] Adrian, T., Etula, E. & Muir. T. (2012). Financial intermediaries and the cross-section of asset returns. Journal of Finance, Forthcoming.

[3] Ang, A., Hodrick, R. J.,Xing, Y., & Zhang. X., (2006). The cross-section of volatility and expected returns. Journal of Finance, pp. 259–99.

[4] Balakrishnan, K., Bartov, E. & Faurel, L. (2010). Post loss/prot announcement drift. Journal of Accounting and Economics, pp. 20-41.

[5] Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. The Journal of Finance, pp. 507-528.

[6] Brealey, R., Myers, S. & Allen, A. (2019). Principles of Corporate Finance (13th Edition). New York: McGraw Hill

[7] Carhart, M. M. (1997). On persistence in mutual fund performance. Journal of Finance, pp. 57–82.

[8] Chen, N., Roll, R. & Ross, S. A. (1986). Economic Forces and the Stock Market. The Journal of Business, pp. 383-403.

[9] Chinco, A., Clark-Joseph, A. D., & Ye, M. (2018). Sparse signals in the cross-section of returns. Journal of Finance (forthcoming).

[10] Cochrane, J. H. (2011). Presidential address: Discount rates. Journal of Finance, pp. 1047-1108.

[11] Connor, G. (1981). A Factor Pricing Theory for Capital Assets. Unpublished working paper.

[12] Cooper, M. J., Gulen, H. & Schill, M. J. (2008). Asset growth and the cross-section of stock returns. Journal of Finance, pp. 1609–51.

[13] Datar, V. T., Naik, N. Y. & Radclie, R. (1998). Liquidity and stock returns: An alternative test. Journal of Financial Markets, pp. 203-219.

[14] Fama, E. F. & J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. Journal of Political Economy, pp. 607-636.

[15] Fama, E. F. & K. R. French (1992). The cross-section of expected stock returns. Journal of Finance, pp. 427-465.

[16] Fan, Y. & Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. Journal of the Royal Statistical Society, pp. 531-552.

[17] Frazzini, A. & Pedersen, L. H. (2013). Betting against beta. Working Paper.

[18] Freyberger, J., Neuhierl, A. & Weber, M. (2018). Dissecting Characteristics Nonparametrically, NBER Working Paper No. 23227.

[19] Friedman, J. H. (1991). Multivariate Adaptive Regression Splines, The Annals of Statistics, pp. 1-67.

[20] Gandhi, P. & Lustig, H. (2015). Size anomalies in US bank stock returns. The Journal of Finance, pp. 733-768.

[21] Garcia, C. & Gould, F. (1993). Survivorship bias. Journal of Portfolio Management, pp. 52-56.

[22] Harvey, C. R., Y. Liu, & H. Zhu (2016). ... and the cross-section of expected returns. Review of Financial Studies, pp. 5-68.

[23] Haugen, R. A. & Baker, N. L. (1996). Commonality in determinants of expected stock returns. Journal of Financial Economics, pp. 401-439.

[24] Horowitz, J. L. (2016). Variable selection and estimation in high-dimensional models.Canadian Journal of Economics, pp. 389-407.

[25] Huang, J., Horowitz, J. L. & Wei, F. (2010). Variable selection in non-parametric additive models. Annals of Statistics, pp. 2282-2313.

[26] Huang, J.Z. & Shi, Z., (2016). Determinants of bond risk premia. Unpublished Manuscript, Penn State University.

[27] Huberman, G. (1982). A Simple Approach to Arbitrage Pricing Theory. Journal of Economic Theory, pp. 183-191.

[28] Ingersoll, J. (1982). Some Results in The Theory of Arbitrage Pricing. Unpublished working paper.

[29] Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. Journal of Finance, pp. 881–98.

[30] Jegadeesh, N. & Titman, S. (1993). Returns to buying winners and selling losers: implications for stock market efficiency. Journal of Finance, pp. 65–91.

[31] Lewellen, J. (2015). The cross section of expected stock returns. Critical Finance Review, pp. 1-44.

[32] Li, J. & Chen, W. (2014). Forecasting macroeconomic time series: Lasso-based approaches and their forecast combinations with dynamic factor models. International Journal of Forecasting, pp. 996-1015.

[33] Lintner, J. (1965). The valuation of risky assets and the selection of risky investments in stock portfolios and capital budgets. The Review of Economics and Statistics, pp. 13-37.

[34] Mossin, J. (1966). Equilibrium in a capital asset market. Econometrica, pp. 768-783.

[35] Pastor, L. & Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. Journal of Political Economy, pp. 643–85.

[36] Pontiff, J. & Woodgate, A. (2008). Share issuance and cross-sectional returns. Journal of Finance, pp. 921–45.

[37] Roll, R (1977). A Critique of the Asset Pricing Theory's Tests Part I: On Past and Potential Testability of the Theory. The Journal of Financial Economics, pp. 129-176.

[38] Ross, S. (1976). The Arbitrage Theory of Capital Asset Pricing. Journal of Economic Theory, pp. 341-60.

[39] Ross, S. (1982). On the General Validity of The Mean-Variance Approach in Large Markets. In Sharpe and Cootner (eds.), Financial Economics: Essays in Honor of Paul Cootner.

[40] Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. The Journal of Finance, pp. 425-442.

[41] Sloan, R. (1996). Do stock prices fully reect information in accruals and cash ows about future earnings? Accounting Review, pp. 289-315.

[42] Talebitooti, R. (2015). Shape design optimization of cylindrical tank using b-spline curves", Computers & Fluids, pp. 100-112.

[43] Tibshirani, R (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (methodological). pp. 267–88.

[44] Treynor, J. L (1962). Market Value, Time, and Risk. Unpublished Working Paper.

[45] Weiner, C. (2005). The impact of industry classification schemes on financial research, SFB 649 Discussion Paper, No. 2005,062, Humboldt University of Berlin, Collaborative Research Center 649 - Economic Risk, Berlin

[46] Xiong, S., Dai, B. & Qian, P. Z. (2016). Orthogonalizing Penalized Regression, Technometrics, pp. 285-293.

[47] Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. J. R. Statist. Soc. B, pp. 49-67.

[48] Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology, pp. 301-320.

9 Appendix

Tables:

Figures:

Table 9: **Description of the 62 Characteristics**

	Previous return
r_{2-1}	The lagged one-month return.
r_{6-2}	The cumulative returned obtained two months ago for a 6 months period.
r_{12-2}	The cumulative returned obtained two months ago for a 12 months period.
r_{12-7}	The cumulative returned obtained in the period between 12 and 7 months ago.
r_{36-13}	The cumulative returned obtained in the period between 12 and 7 months ago.
	Investment
$Investment$	The year-on-year % change in total assets (AT)
$\Delta SHROUT$	% change in outstanding shares.
ΔCEQ	% change in Book-Value of Equity
$\Delta PI2A$	change in Property, Plants and Equipment + Inventory divide on Total Lagged assets (TA)
IVC	change in Inventories (INVT) between t-2 and t-1 divide on average total assets (AT)
NOA	Net Operating Assets, (Operating assets – operating liablities * lagged total assets)
	Profitability
ATO	Sales to lagged net operating assets, $\frac{Sales}{Net\ operating\ assets\ _{t-1}}$
CTO	Capital Turnover (Ratio of net sales * lagged total assets (AT)
$\Delta(\Delta GM - \Delta Sales)$	% change in Gross margin and Sales (Gross margin = Difference in sale and costs of goods sold)
EPS	Earnings per share
IPM	Pre-tax profit margin (ratio of pre-tax income to sales)
PCM	Price-to-cost margin (Net sales – Costs of goods sold divided by net Sales)
PM	Profit Margin (Operating income after depreciation divided on Sales)
PM_adj	Adjusted Profit Margin ((Operating income after depreciation divided on Sales) – average profit margin)
$Prof$	Profitability (Gross prof divided by book value on Equity)
RNA	Return on net operating assets (operating income after depreciation * lagged net operating assets)
ROA	Return on Assets $\frac{Net\ Income}{Average\ total\ assets}$
ROC	Return on Capital
ROE	Return on equity, $\frac{Net\ Income}{Total\ Assets(AT) - Total\ Liabilities}$
$ROIC$	Return on invested Capital
$S2C$	Sales to cash, $\frac{Sales}{Cash}$
SAT	Asset Turnover (ratio of sales compared to total assets (AT))
SAT_adj	Adjusted asset turnover (ratio of sales compared to total assets – average asset turnover)
	Intagibles
AOA	Absolute value of operation accruals
OL	$\frac{\sum(cost\ of\ goods\ sold)\ (COGS) + administrative\ expenses\ (XSGA)}{Total\ Assets(AT)}$
Tan	Tangibility (0.715 * total receivables + 0.547 * inventories + 0.535 * property, plant and equipment + cash and short term investments divided on total assets
OA	$\Delta\ noncash\ working\ capital -\ depreciation\ (DP) \times\ lagged\ total\ assets\ (TA)$

Characteristics cont.

	Value
$A2ME$	Asset to market cap, $\frac{Total\ Assets(AT)}{Market\ Cap\ December_{t-1}}$. Market Cap = SHROUT $*$ Price.
$BEME$	Book value of equity
$BEME\_adj$	ratio of Book value of equity compared to market value of equity – average industry ratio of book value of equity compared to market value of equity using Fama etc 48 industry level
C	The CF to TA ratio
$C2D$	ratio (income and extraordinary items (IB), and dep and amor (dp) to tot liab (LT)
CTO	Capital turnover as the ratio of net sales (SALES) times total assets (AT)
ΔSO	Log change in the split adjusted SHARES OUTSTANDING (split adjusted shares are Compustat shares outstanding and adjustment factor (AJEX)
$Debt2P$	Debt to price (ratio of long-term debt and debt in current liabilities to market capitalization dec t-1, market cap is Shares outstanding * price
$E2P$	Earnings to price (ratio of income before extraordinary items to shares outstanding
FCF	Free Cash Flow = $(NI + DP - \Delta WC - CAPEX)/BEME$
LDP	Dividend price ratio (annual dividend divided by last months price
NOP	Net payout ratio (common dividends + purchase of common and preferred stock – sale of common and preferred stock divided by market cap
Q	Tobin's Q
$02P$	Payout ratio (common dividends + purchase of common and preferred stock – change in value of net number of preferred stocks outstanding divided by market cap
$S2P$	Sales to price, $\frac{Sales}{Price}$
$Sales\_g$	Sales growth
	Trading frictions
AT	Total assets
$Beta$	Correlation between the excess return of stock i and the market return (CAPM)
$Beta\ daily$	Sum of regression coefficients of daily excess returns on the market excess return and one lag of the market excess return
DTO	Turnover (Turnover is the ratio of volume (VOL) times shares outstanding (SHROUT))
$Idiovol$	Idiosyncratic volatility (std of residuals from regression of excess returns on three factor model FandF)
LME	Total Market Capitalization of the previous month (Price * Shares outstanding)
LME_adj	Industry-adjusted-size (Price * Shares outstanding – average market capitalization FandF 48 industry)
$Lturnover$	$\frac{Last\ Month's\ Volume(VOL)}{Shares\ Outstanding(SHROUT)}$
$Rel\_to\_high\_price$	Closeness to 52-week high (ratio of stock price (PRC) at the end of the previous calendar month and the previous 52 week high price
Ret_max	Maximum daily return in the previous month
$Spread$	Bid-Ask spread (average bid-ask spread in the previous month)
$Std\ turnover$	Standard deviation of the residuals from a regression of daily turnover on a constant (use one month of daily data and require at least fifteen non-missing observations)
$Std\ volume$	Standard deviation of the residuals from a regression of daily volume on a constant (one month of daily data and require at least fifteen non-missing observations)
SUV	Standard unexplained volume (diff between actual volume and predicted volume, previous month)
$Total\ vol$	Total volatility

Table 10: **Selected Characteristics using the Adaptive Group Lasso**

Firms		All	All	All	All	All
Sample		Full	Full	Full	1965-1990	1991-2014
Knots		20	15	25	15	15
Sample size		1,6m	1,6m	1,6m	0.6m	1m
# **Selected**		13	16	13	11	14
Characteristics	# **Selected**	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
BEME	1				•	
$\Delta SHROUT$	5	•	•	•	•	•
ΔSO	4	•	•	•		•
Investment	4	•	•	•		•
LDP	1				•	
LME	5	•	•	•	•	•
Lturnover	4	•	•	•		•
NOA	2		•			•
NOP	1				•	
PM_adj	4	•	•	•		•
r_{2-1}	5	•	•	•	•	•
r_{12-2}	4	•	•	•	•	
r_{12-7}	5	•	•	•	•	•
r_{36-13}	2		•			•
Rel_to_high_price	5	•	•	•	•	•
Ret_Max	1				•	
ROC	4	•	•	•		•
SUV	5	•	•	•	•	•
Total_vol	4	•	•	•		•

Figure 1: Characteristics Descriptive Statistics

	Mean	Median	Std. Dev
a2me	3,202	1,512	7,881
aoa	6,567	0,057	2557,564
at	3654,596	199,231	36743,900
at_adj	0,020	-0,054	0,745
ato	2,522	1,912	21,543
beme	0,900	0,679	0,961
beme_adj	0,012	-0,114	0,864
beta	0,983	0,904	0,614
beta_daily	0,846	0,777	1,744
c	0,139	0,067	0,176
c2d	0,103	0,137	2,001
cto	1,285	1,126	1,213
cum_return_12_2	0,150	0,063	0,693
cum_return_12_7	0,080	0,033	0,460
cum_return_1_0	0,013	0,000	0,167
cum_return_36_13	0,353	0,143	1,251
cum_return_6_2	0,067	0,029	0,412
d_ceq	0,221	0,082	4,028
d_dgm_dsales	-0,410	-0,002	40,844
d_shrout	0,009	0,000	0,137
d_so	0,038	0,005	0,145
debt2p	0,887	0,289	3,156
dpi2a	0,082	0,044	0,277
dto	0,000	0,000	0,013
e2p	-0,015	0,054	0,609
eps	1,598	0,852	55,491
free_cf	-0,340	0,044	58,001
idio_vol	0,030	0,022	0,028
investment	0,156	0,075	0,651
ipm	-1,397	0,064	104,790
ivc	0,013	0,001	0,065

Mean	Median	Std. Dev	
0,018	0,000	0,088	ldp
1960,191	138,957	11555,269	lme
335,925	-393,393	11319,633	lme_adj
0,097	0,046	0,206	lturnover
0,648	0,670	0,497	noa
0,004	0,006	0,168	nop
0,029	0,013	0,177	o2p
1,843	-0,032	2557,571	oa
1,050	0,914	0,933	ol
-0,914	0,327	104,504	pcm
-1,371	0,079	107,995	pm
0,510	0,088	104,387	pm_adj
1,097	0,632	45,993	prof
1,671	1,165	1,826	q
0,741	0,792	0,212	rel_to_high_price
0,073	0,051	0,088	ret_max
0,017	0,039	0,201	rna
0,017	0,039	0,201	roa
-12,085	-1,252	1817,103	roc
0,031	0,100	3,223	roe
0,041	0,058	0,144	roic
108,884	12,900	1854,536	s2c
2,615	1,244	5,202	s2p
0,457	0,092	30,991	sales_g
1,141	1,024	0,956	sat
0,035	0,019	0,064	spread_mean
0,380	0,176	1,215	std_turn
218,737	21,828	1708,369	std_volume
0,241	-0,190	3,025	suv
0,532	0,543	0,140	tan
0,032	0,025	0,028	total_vol

Figure 2: Fama and French Industry Classification - 12 Industries

1 NoDur Consumer NonDurables -- Food, Tobacco, Textiles, Apparel, Leather, Toys

0100-0999
2000-2399
2700-2749
2770-2799
3100-3199
3940-3989

2 Durbl Consumer Durables -- Cars, TV's, Furniture, Household Appliances

2500-2519
2590-2599
3630-3659
3710-3711
3714-3714
3716-3716
3750-3751
3792-3792
3900-3939
3990-3999

3 Manuf Manufacturing -- Machinery, Trucks, Planes, Off Furn, Paper, Com Printing

2520-2589
2600-2699
2750-2769
3000-3099
3200-3569
3580-3629
3700-3709
3712-3713
3715-3715
3717-3749
3752-3791
3793-3799
3830-3839
3860-3899

4 Enrgy Oil, Gas, and Coal Extraction and Products

1200-1399
2900-2999

5 Chems Chemicals and Allied Products

2800-2829
2840-2899

6 BusEq Business Equipment -- Computers, Software, and Electronic Equipment

3570-3579
3660-3692
3694-3699
3810-3829
7370-7379

7 Telcm Telephone and Television Transmission

4800-4899

8 Utils Utilities

4900-4949

9 Shops Wholesale, Retail, and Some Services (Laundries, Repair Shops)

5000-5999
7200-7299
7600-7699

10 Hlth Healthcare, Medical Equipment, and Drugs

2830-2839
3693-3693
3840-3859
8000-8099

11 Money Finance

6000-6999

12 Other -- Mines, Constr, BldMt, Trans, Hotels, Bus Serv, Entertainment

1000-1199
1400-1999
2400-2499
3800-3809
4000-4799
4950-4999
7000-7199
7380-7599
7700-7999
8100-9999

Figure 3: Characteristics chosen in the small industries

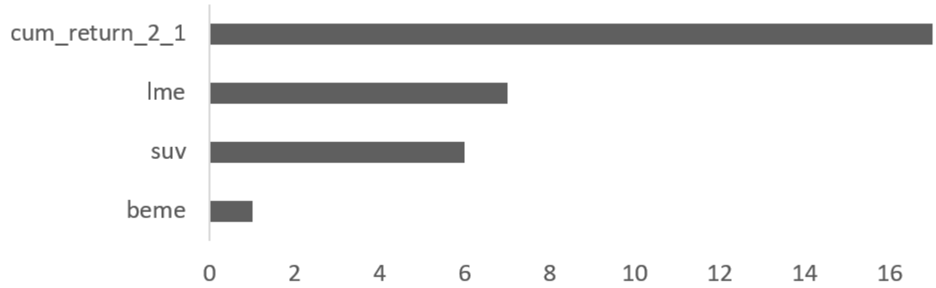


Figure 4: Characteristics chosen in medium industries

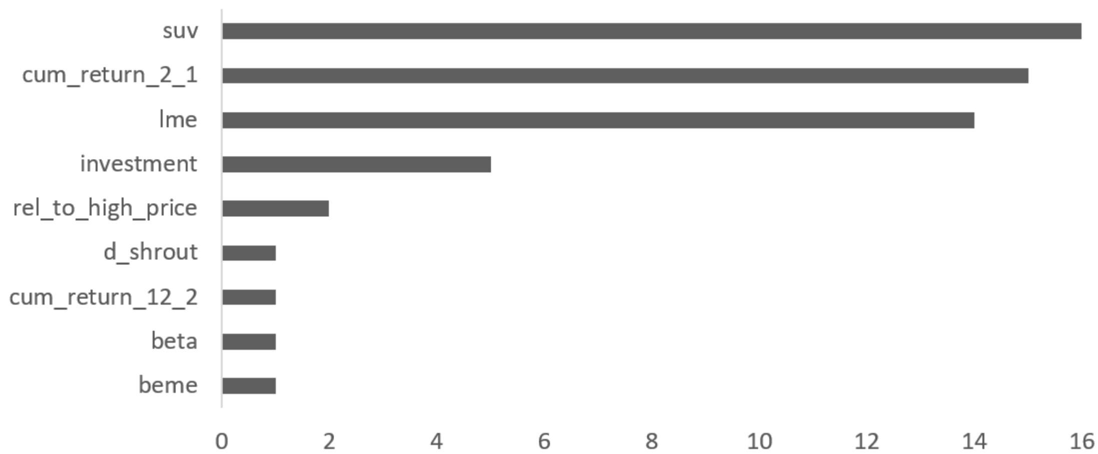


Figure 5: Characteristics chosen in the large industries

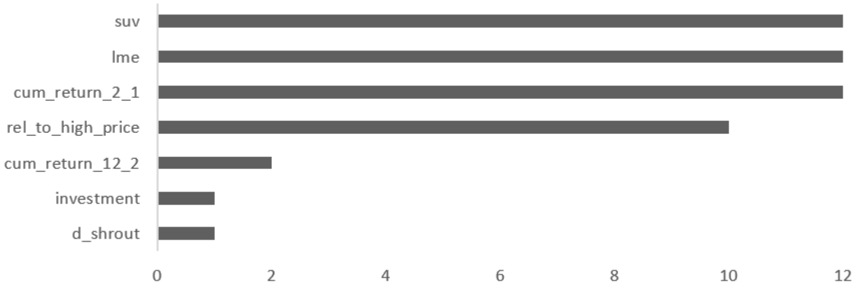


Figure 6: Characteristics chosen in CLRM

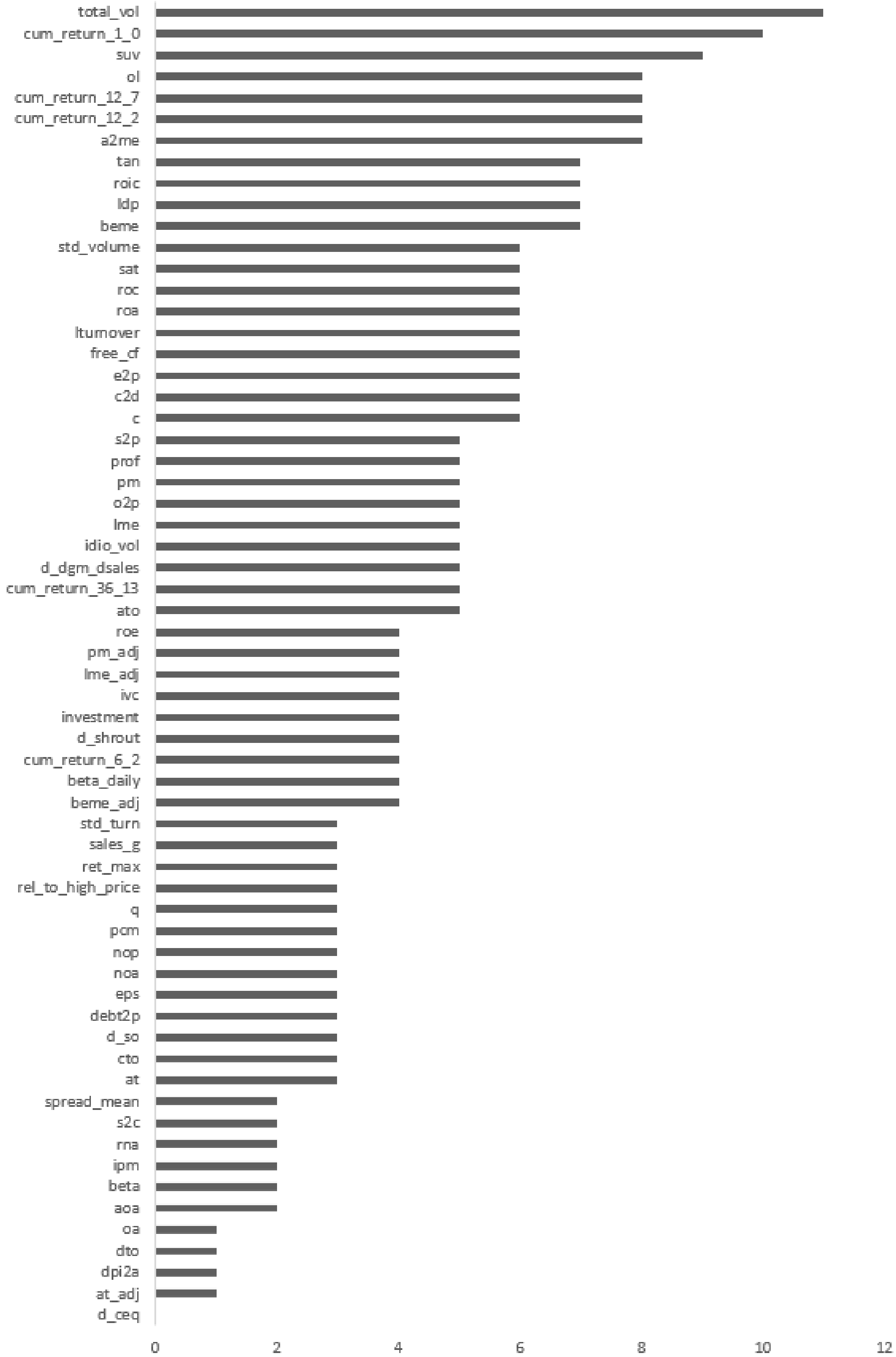
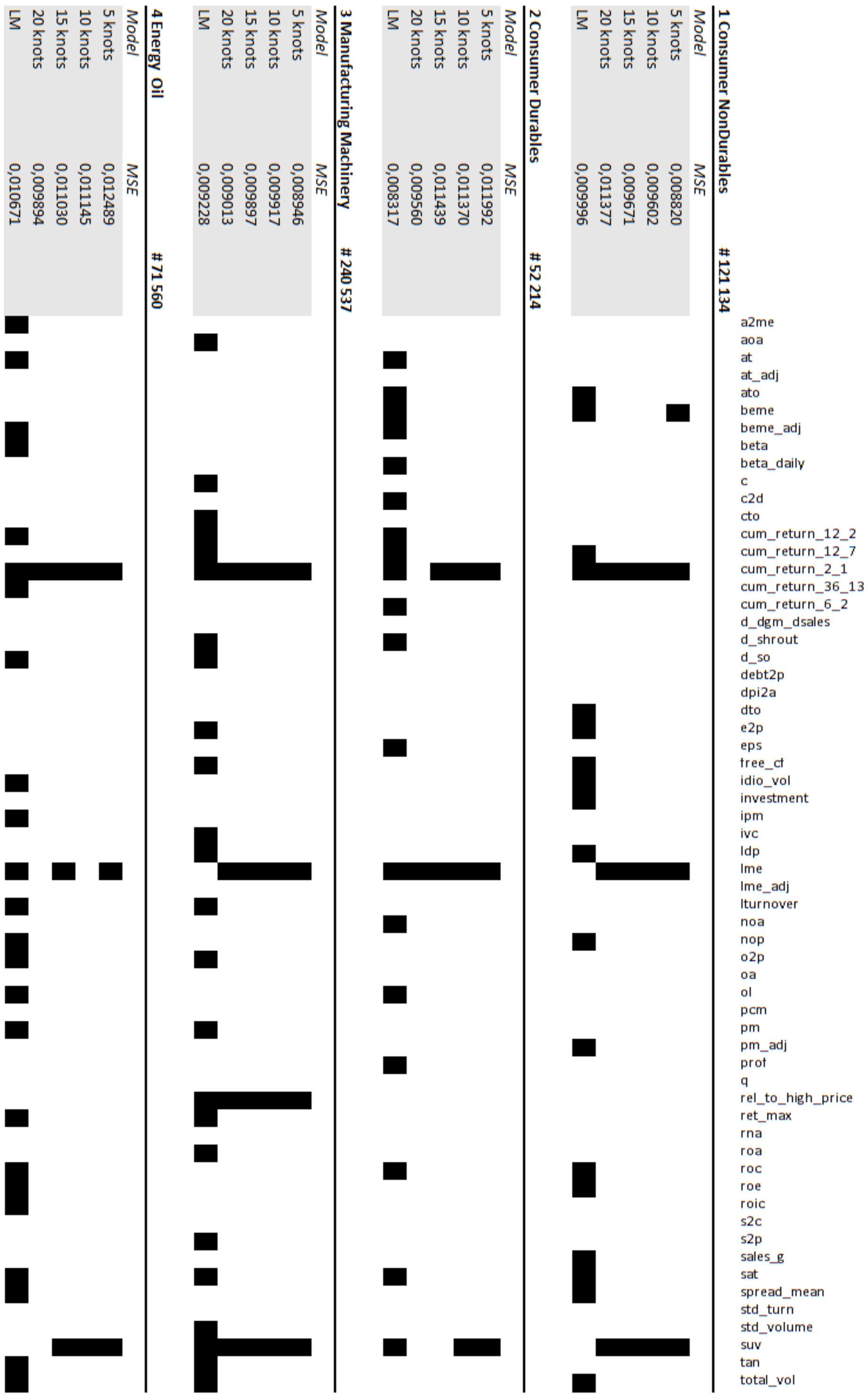
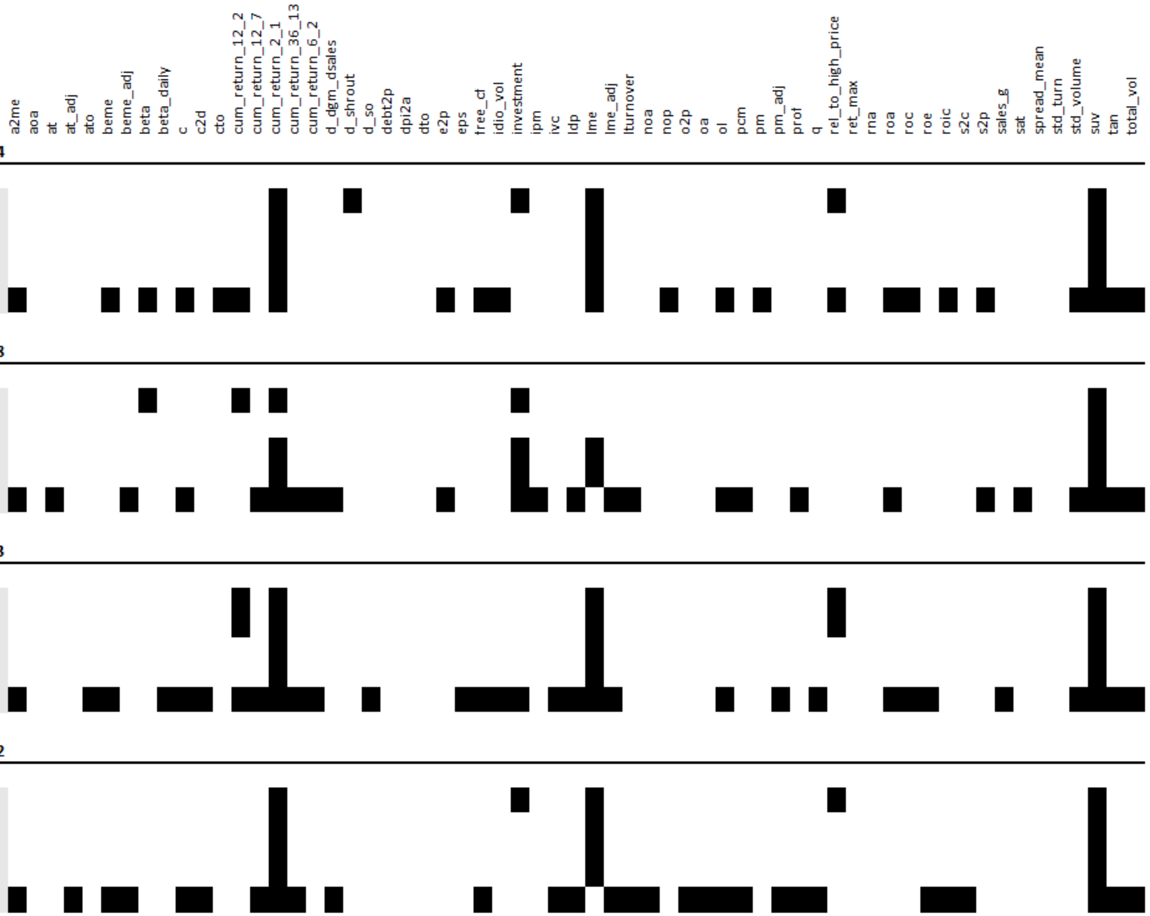


Figure 7: Selected models and out-of-sample MSE



		a2me	aoa	at	at_adj	ato	beme	beme_adj	beta	beta_daily	c	c2d	cto	cum_return_12_2	cum_return_12_7	cum_return_2_1	cum_return_36_13	cum_return_6_2	d_dgm_dsales	d_shrout	d_so	debt2p	dpi2a	dto	e2p	eps	free_cf	idio_vol	investment	ipm	ivc	ldp	lme	lme_adj	lturnover	noa	nop	o2p	oa	ol	pcm	pm	pm_adj	prof	q	rel_to_high_price	ret_max	rna	roa	roc	roe	roic	s2c	s2p	sales_g	sat	spread_mean	std_turn	std_volume	suv	tan	total_vol
5 Chemicals and Allied Products	**# 49 468**																																																													
Model	*MSE*																																																													
5 knots	0,008631															■																																														
10 knots	0,008670															■																																														
15 knots	0,009203															■																														■																
20 knots	0,009319															■																																														
LM	0,008074	■	■			■				■		■	■		■		■		■			■			■						■		■		■			■						■	■		■	■				■						■				■
6 Business Equipment	**# 257 930**																																																													
Model	*MSE*																																																													
5 knots	0,011982															■													■				■													■														■		
10 knots	0,010954															■				■													■													■														■		
15 knots	0,011509															■																	■													■														■		
20 knots	0,011215															■																	■													■														■		
LM	0,011346					■	■			■	■			■		■			■	■		■	■			■	■					■			■					■		■	■			■		■	■			■			■					■	■	■
7 Telephone and TV	**# 32 891**																																																													
Model	*MSE*																																																													
5 knots	0,014558															■																																														
10 knots	0,010424															■																																														
15 knots	0,010214															■																																														
20 knots	0,009375															■																																														
LM	0,013707	■					■					■		■																								■			■			■						■		■	■	■				■	■	■		■
8 Utilities	**# 67 537**																																																													
Model	*MSE*																																																													
5 knots	0,009159						■																										■																											■		
10 knots	0,009615															■																																														
15 knots	0,009373															■																																														
20 knots	0,012188															■																																														
LM	0,009486	■										■		■	■	■		■	■	■		■			■			■	■			■		■		■				■		■							■			■		■		■		■	■	■		■



9 Wholesale and Retail **# 178 114**

Model	*MSE*
5 knots	0,011794
10 knots	0,012098
15 knots	0,010085
20 knots	0,008826
LM	0,011971

10 Healthcare **# 130 898**

Model	*MSE*
5 knots	0,010305
10 knots	0,010381
15 knots	0,010719
20 knots	0,011377
LM	0,010148

11 Money Finance **# 225 793**

Model	*MSE*
5 knots	0,011552
10 knots	0,012043
15 knots	0,009671
20 knots	0,009560
LM	0,013280

12 Other **# 180 352**

Model	*MSE*
5 knots	0,009643
10 knots	0,009493
15 knots	0,011439
20 knots	0,009013
LM	0,009202