UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

Tidying up the factor zoo

Using machine learning to find sparse factor models that predict asset returns

Oliver Klingberg Malmer & Gustav Pettersson

June 2020

A thesis presented for the degree of
Bachelor of Science in Economics

Thesis advisor: Andreas Dzemski
School of Business, Economics and Law
Department of Economics

Abstract

There exist over 300 firm characteristics that provide significant information about average asset return. John Cochrane refers to this as a "factor zoo" and challenges researchers to find the independent characteristics which can explain average return. That is, to find the unsubsumed and non-nested firm characteristics that are highly predictive of asset return. In this thesis we act on the posed challenge by using a data driven approach. We apply two machine learning methods to create sparse factor models composed by a small set of these characteristics. The two methods are one unsupervised learning method, the Principal Component Analysis, and one supervised learning method, the LASSO regression. The study is done using the S&P 500 index constituents and 54 firm characteristics over the time period 2009-07-01 to 2019-07-01. The performance of the factor models is in this study measured using out-of-sample measurements. Using established methods of post-LASSO regression and new developed techniques for variable selection based on PCA, we generate four new factor models. The latter mentioned variable selection method based on PCA is, to our knowledge, an original contribution of this thesis. The generated factor models are compared against the Fama French factors in the out-of-sample test and are shown to all outperform. The best performer is a LASSO generated factor model containing 6 factors. By analysing the results we find that momentum factors, such as price relative to 52-week-high-price, are highly predictive of return and are commonly selected factors, which confirms the results of previous responses to the same challenge.

Keywords: Asset pricing, Factor models, Machine learning, PCA, LASSO, Variable selection, Dimension reduction, Fama French Three Factor model, Fama French Five Factor model.

Acknowledgements

We would like to express our special thanks and gratitude to our advisor Andreas Dzemski, who enabled this thesis by his commitment, guidance and encouragement, and the open-source community of R, without which this thesis would not be possible.

Contents

1 Introduction

1.1 Background

The field of asset pricing is concerned with trying to understand prices and values of claims with uncertain payments. In essence, a valuation of an asset assesses two factors; *time* and *risk*. The factor of time represent the delay of the payment. This effect is not hard to calculate. However, the factor of risk plays a greater role in valuation and is much harder to assess (Cochrane 2005). In the context of valuation of equity assets, factor models is often used as a proxy for risk in order to calculate expected return. With the Capital Asset Pricing Theory (CAPM) (Sharpe 1964, Litner 1965, Mossin 1966) began the research of these models. CAPM provides investors with the expected return of an asset, based on the market risk of the asset (Bodie et al. 2018). However, it did not take long until *anomalies* were discovered. Anomalies are defined as empirical results that are inconsistent with maintained theories (Schwert 2003).

Based on these findings came the *Fama French Three Factor model* (Fama & French 1993), which included the market factor of the CAPM, a size factor and a value factor. Since then a great amount of new factors has been developed and been found to have predictive power for cross-sectional expected return (Freyberger et al. 2016). Harvey et al. (2016) identify more than 300 factors in this category. John Cochrane addresses this issue in his presidential address Cochrane (2011), and calls the current state of asset pricing research of significant factors a “factor zoo”. To clarify, there might be some small number of factor that are highly predictive on return and that the large number of other factors that has shown to affect return are simply a product of this small set of highly predictive factors. That is, it exist factors that are nested in other variables. An example of this might be cash, when asset it the predictive factor. There will exist correlations between these but the challenge of distinguishing which factors that are the highly predictive factors still remain. This is the challenge that Cochrane brings forth to the asset pricing research field in his presidential address. As factor models are used to explain market movements and market phenomenons, the task at hand is of great importance. Furthermore, as the explanation given by factor must work out-of-sample, for prediction purposes, the importance of parsimoniousity is of essence.

There have been many attempts to address this challenge. Some notable examples are Freyberger et al. (2016) with the method of adaptive group LASSO, Green et al. (2017) with the method of Fama-Macbeth with avoidance of overweighting microcap and adjusted for data-snooping bias, and Kelly et al. (2019) with the method of Instrumental Principal Component Analysis. In this thesis, we will continue in these examples' footsteps and try to find independent factors with the methods of Principal Component Analysis (PCA) and Least Absolute Shrinkage and Selection Operator (LASSO) Regression.

PCA was first developed by Karl Pearson (1901), and is a common tool in dimensionality reduction and has previously been applied in finance, see for example Feeney et al. (1964), Schneeweiss & Mathes (1995) and Zhong & Enke (2017). PCA is a unsupervised learning method (James et al. 2013) and by applying PCA to a large data set, one summarize the variability of the data with a small set of components generated by the PCA. When using a set of firm characteristics that has already shown to affect returns, it is reasonable to assume that PCA could be used to select components that explain asset returns. This comes as a handy tool for the mentioned objective.

The LASSO regression on the other hand, popularized by Robert Tibshirani (1996), is a supervised learning method that uses *shrinkage*, meaning to reduce coefficients. LASSO has the ability to force coefficients to zero (James et al. 2013), with the help of a penalty term. The non-zero coefficient variables can then be extracted and applied elsewhere, which commonly refers to as *post-LASSO* (Hastie et al. 2009). In the financial literature, LASSO has been used both for variable selection and for prediction, see for example Freyberger et al. (2016) and Feng et al. (2017).

To summaries, the field of asset pricing research is filled with a vast amount of discovered factors with significant effect on asset returns. This has become a problem for generalizing huge market movements and phenomenons (Cochrane 2011), and in prediction of out-of-sample data. By using the methods of PCA and LASSO, we will in this thesis extract the most predictive factors of average asset return and create sparse factor models.

1.2 Purpose

Factor models can help investors to evaluate if an asset is too cheap or too expensive and help academics in the pursuit of understanding market phenomenons and market movements. Although the "factor zoo" offer a great way of analyzing anomalies of the efficient market hypothesis, it lacks the ability to understand large coordinated market movements (Cochrane 2005). Sparse factor models have the advantage of being parsimonious and simple. This makes the explanation intuitive and simple. Furthermore, the ability to make out-of-sample prediction are most likely increased in sparse models.

The aim of this thesis is find predictive sparse factor models and answer Cochrane (2011)'s question of which firm characteristics can provide independent, non-nested, information about average asset return. Furthermore, the validity of Principal Component Analysis (PCA) and LASSO regression as methods for identifying independent characteristics is tested.

As mentioned before, previous researchers have used other methods such as the Fama-MacBeth regression (Green et al. 2017) and the Instrumental Principal Component Analysis (Kelly et al. 2019). With 54 firm characteristics of each company in the S&P 500 index and the usage of PCA and LASSO in our thesis, we will contribute with further research in firm characteristics to the factor literature.

1.3 Research Question

The purpose of answering Cochrane (2011)'s presidential address, leads us to the formulation of the thesis' research question. The answer to the following research question will be sought in this thesis:

Which firm characteristics should be included in a sparse factor model that predict asset return?

To answer this research question we will use a sample of listed companies on the New York Stock Exchange (NYSE) and NASDAQ which are included in the S&P 500 index. The firm characteristics are all included in Freyberger et al. (2016) and Green et al. (2017) and are calculated with help from S&P Capital IQ.

2 Literature Review

The early asset pricing research field had no direction until CAPM was developed (Cochrane 2011). Then came anomalies and the field lost its direction again. Cochrane (2011) notes that the traditional way of examining excess return, that is, portfolio sorting, see Fama & French (1993), Carhart (1997) and Fama & French (2015), works poorly when the number of included factors are excessive and researchers should find other methods. With the challenge from Cochrane (2011) began the research of alternative methods for finding and evaluating which independent characteristics explain average return. These methods will be presented now together with the corresponding results.

When Green et al. (2017) compute their Fama-Macbeth regressions with 94 characteristics during the years 1980 to 2014, they find that only 12 characteristics provided significant independent information about average return before 2003. After 2003 the predictability of returns fell and only two factors have been viable since 2003. Green et al. (2017) also test their 12 factor model with control for the benchmark models *Carhart Four Factor Model* (Carhart 1997), *Fama French Five Factor Model* (Fama & French 2015) and the *q-Factor Model* (Hou et al. 2015). First off, they find that 11 of out the 12 significant factors differ from the factors included in the benchmark models. Secondly, they find that when controlling for the benchmarks, their results does not differ much. Their factors offer new information (Green et al. 2017).

With their adaptive group LASSO Freyberger et al. (2016) get a higher explanatory power out-of-sample than linear regressions. Freyberger et al. (2016) uses 36 characteristics and finds that between 7 and 15 factors provide independent information about average returns. Factors extracted from Freyberger et al. (2016) include market capitalization, investments and various momentum factors.

Unlike Green et al. (2017) and Freyberger et al. (2016), in their effort to answer Cochrane (2011), Kelly et al. (2019) try to identify latent characteristics that provide independent information. They find that the own developed *Instrumental Principal Component Analysis (IPCA)* model outperform existing factor models such as the *Fama French Five Factor Model* (Fama & French 2015) in delivering accurate predictions, both in-sample and out-of-sample. The factors with most predictive power are 12-month momentum and size. IPCA does also deliver higher out-of-sample "mean-variance" efficiency than other methods (Kelly et al. 2019).

Another previous work in the factor literature in order to distinguish anomalies in asset pricing is Hou et al. (2015), which examines 35 firm characteristics, such as momentum (Jegadeesh & Titman 1993, Carhart 1997), return on equity & capital turnover (Haugen & Baker 1996) and market equity (Banz 1981) for naming a few. Hou et al. (2015) work results in the *q-Factor Model*, which includes a market factor, a size factor, an investment factor and a profitability factor.

The usage of *machine learning* in the field of asset pricing has in the last years increased greatly. Rapach et al. (2013) use LASSO for predicting international equity market returns and in their comparative analysis of machine learning methods Gu et al. (2020) finds that neural networks and other machine learning based methods can help to understand empirical asset pricing. In Feng et al. (2017)'s effort of "taming the factor zoo", one once again see the usage of a LASSO based method, namely the double LASSO, as a tool for explaining return.

As presented, various researchers have tackled the challenge of identifying independent characteristics that provide information about average return. For more existing literature in the field of firm characteristics and cross-sectional returns, see the notable work of Haugen & Baker (1996), Daniel & Titman (1997), Light et al. (2017), Kozak et al. (2018) and Kozak et al. (2020).

3 Theory

3.1 Asset Pricing

This section introduces the reader to the fundamental challenges of asset pricing and to the current state of the research field. In this section we start off with an utility function expressed in terms of consumption and derive a general asset pricing function, which then is used to express price as a function of factors (firm characteristics). Note that this section is vastly based on the explanations made in Cochrane (2005).

3.1.1 Asset Pricing Function

In his book (Cochrane 2005), John Cochrane outlines that the basic objective in asset pricing is to figure out the value of an uncertain stream of cash flows. This can be seen as figuring out the value of the asset at time t with an *payoff* x_{t+1}. The payoff is different depending on the asset type, if considering a stock then the payoff equals the price in the future plus dividend, $x_{t+1} = p_{t+1} + d_{t+1}$. Although x_{t+1} is a random variable, investors can asses probabilities of outcomes. Asset pricing is more concerned with what the typical investor is willing to pay for some future payoff. To answer this, an *utility function*, expressed in terms of consumption, is used,

$$U(c_t, c_{t+1}) = u(c_t) + \beta E_t[u(c_{t+1})], \tag{1}$$

where c_t denotes consumption in time t and c_{t+1} in time $t+1$, $u(\cdot)$ denotes some increasing and concave utility function and β denotes the subjective discount which reflects the fact that the investor is impatient and values consumption today higher than consumption tomorrow. If one again consider the payoff and assume that the investor can buy and sell as much as he/she likes at price p_t, then the choice of the investor can be written as

$$\max_{\xi} u(c_t) + E_t[\beta u(c_{t+1})] \tag{2}$$

subject to

$$\begin{aligned} c_t &= e_t - p_t\xi, \\ c_{t+1} &= e_{t+1} + x_{t+1}\xi \end{aligned} \tag{3}$$

where e denotes the consumption level with no purchase of the asset, and ξ denotes the amount that the investor chooses to buy. By substituting in the constraints one gets,

$$\max_{\xi} u(e_t - p_t\xi) + E_t[\beta u(e_{t+1} + x_{t+1})]. \tag{4}$$

By then setting the derivative with respect to ξ to zero one gets,

$$u'(c_t)(-p_t) + E[\beta u'(c_{t+1})x_{t+1}] = 0, \tag{5}$$

and then finally end up with,

$$p_t = E_t \left[\beta \frac{u'(c_{t+1})}{u'(c_t)} x_{t+1} \right]. \tag{6}$$

Equation (6) is referred to as *the central asset pricing formula* and can also be written as

$$p_t = E_t \left[m x_{t+1} \right], \tag{7}$$

where

$$m \equiv \beta \frac{u'(c_{t+1})}{u'(c_t)}.$$

The interpretation goes as follows. Price at time t, p_t, is the expected value of the discount factor, m, times the payoff at time $t+1$, x_{t+1}.

3.1.2 Risk

Let's consider the risk-free rate given by

$$R^f = \frac{1}{E(m)}, \tag{8}$$

and gross-return of an asset is given by

$$R_{t+1} \equiv \frac{x_{t+1}}{p_t}. \tag{9}$$

The return of an asset can be thought of as the payoff of an asset with price 1, that is,

$$1 = E(mR). \tag{10}$$

If one were to go back to the original pricing function, $p = E(mx)$, apply the decomposition of the covariance, $cov(m,x) = E(mx) - E(m)E(x)$, and substituting in the risk-free rate equation given in Equation (8), then one gets

$$p = \frac{1}{E(x)} + cov(m,x). \tag{11}$$

One can then see that price can be expressed as the sum of a discounted present formula and a risk adjustment term. One might think that the volatility of the asset determines the risk and consequently determines part of the price. However, as one can see, if there is no correlation between the discount factor and the payoff, then the volatility has no effect on the price. Furthermore, the covariance term's affect on the price can be interpreted as that the investor does not like uncertainty. The discount factor m is large when the utility function is stable. If the covariance of m and x is positive, then this means that when x increases, m increases as well. Since $u'(c_{t+1})$ is the only term that is not fixed, it must mean that this is what increases with x. Considering that $u(\cdot)$ is increasing and concave, an increase in $u'(c_{t+1})$ comes from a decrease in $u(c_{t+1})$, i.e. decrease in c_{t+1}. The conclusion is that if $cov(m, x) > 0$ then $cov(c_{t+1}, x) > 0$. The fact that this increases the price of the asset stems from the fact that an investor prefers to own an asset which payoff increases when he/she is feeling poor over an asset which payoff increases when he/she is already feeling wealthy.

One can now move on by rewriting the formula in Equation (11) and moving into the realm of returns by applying the case in which price equals 1, as in Equation (10), and using the the risk-free rate given by Equation (8),

$$E(R^{ei}) = E(R^i) - R^f = -R^f cov(m, R^i). \tag{12}$$

One end up with an equation for calculating the expected excess return of an asset. One can see that the expected excess return will be lower for assets with return that covaries with the discount rate, i.e. have a negative covariance with consumption. If the investor is willing to take the risk, the expected excess return will be higher (Cochrane 2005).

3.1.3 Beta Pricing Models

Equation (12) can be rewritten as

$$E(R^i) = R^f + \left(\frac{cov(m, R^i)}{var(m)}\right)\left(-\frac{var(m)}{E(m)}\right), \tag{13}$$

which one then can express as a *beta pricing model*,

$$E(R^i) = R^f + \beta_{i,m}\lambda_m, \tag{14}$$

by defining $\beta_{i,m} \equiv \frac{cov(m,R^i)}{var(m)}$ and $\lambda_m \equiv -\frac{var(m)}{E(m)}$. This is useful as the $\beta_{i,m}$ is the regression coefficient of the return on the discount factor, and will come in handy later.

3.1.4 Factor Pricing Models

So far, we have concluded that consumption should be able to determine price and expected return of a portfolio. Unfortunately, the model lacks empirical support

(Cochrane 2005). This might have to do with uncertainty concerning the utility function. Motivated by the lack of empirical support, the concept of *factor pricing models* was developed as a modification of the beta pricing model by adding more variables in order to increase the predictive power of the model. Factor pricing models are models that try to explain the discount factor as a linear function composed by combinations of factors that act as proxies

$$m_{t+1} = a + b_A f_{t+1}^A + b_B f_{t+1}^B + \cdots . \tag{15}$$

3.1.5 Expected Return-Beta Representations

By using the methodology of factor-pricing models, one can construct *expected return-beta representations.* This is done by combining Equation (14) and (15), which results in the following equation,

$$E(R^i) = \gamma + \beta_{i,a}\lambda_a + \beta_{i,b}\lambda_b + \cdots , \quad i = 1, 2, \cdots , N. \tag{16}$$

The βs are defined as the coefficients in the time-series regression

$$R_t^i = a_i + \beta_{i,a} f_t^a + \beta_{i,b} f_t^b + \cdots + \epsilon_t^i, \quad t = 1, 2, \cdots , T. \tag{17}$$

The interpretation of these variables are as follows. β is the amount of exposure that an asset has to a risk factor, and λ is the price of that exposure.

There is an enormous amount of research that is focused on the average return across assets that is outlined in Equation (16). One of the most notable of such research is the *Capital Asset Pricing Model (CAPM).* CAPM was originally developed from *Modern Portfolio Theory* (Markowitz 1952), which showed that it is possible for an investor to maximize returns by diversifying the portfolio, in a sequence of papers (Sharpe 1964, Litner 1965, Mossin 1966). The CAPM model proxies return of the market portfolio for marginal utility growth,

$$E(R^i) = \gamma + \beta_i(E(R^m) - \gamma), \quad i = 1, 2, \cdots , N \tag{18}$$

where $E(R^i)$ denotes the expected return for asset i and $E(R^m)$ is the expected return of the market portfolio, and γ denotes the risk free rate and is usually proxied by the one-month US Treasury bill rate (Fama & French 2004). The sensitivity of return for asset i to the market return can be represented by β_i. The market risk premium is represented by the difference between the return of the the market portfolio $E(R^m)$ and the risk free rate γ (Fama & French 2004).

Although developed 56 years ago, CAPM is still widely used in various areas, such as the evaluation of portfolio performance (Fama & French 2004). However, no model is perfect and throughout the times there have been studies that present anomalies related to a large amount of factors and characteristics.

A notable example is Banz (1981), which finds evidence for risk premium represented by a size factor. Influenced by this, Fama & French (1992) studied the relationship between a firms book-to-market value and its returns, which resulted in the discovery of the *value premium.* Fama and French based a new model on these studies, presented in Fama & French (1993). This model is often referred to as the *Fama French Three Factor Model*, and is given by

$$E(R^{ei}) = \alpha + \beta_{1,i}(E(R^m) - R^f) + \beta_{2,i}E(R^{SMB}) + \beta_{3,i}E(R^{HML}) + \epsilon_i, \tag{19}$$

R^{SMB} denotes the difference in expected return between small and big size firm portfolios, and R^{HML} denotes the difference in expected return between high and low book-to-market ratio portfolios (Bodie et al. 2018). Close to a decade later, Fama and French found proof for new anomalies and designed the *Fama French Five Factor Model* to address these anomalies Fama & French (2015),

$$\begin{aligned} E(R^{ei}) = \alpha + \beta_{1,i}(E(R^m) - R^f) + \beta_{2,i}E(R^{SMB}) + \beta_{3,i}E(R^{HML}) \\ +\beta_{4,i}E(R^{RMW}) + \beta_{5,i}E(R^{CMA}) + \epsilon_i, \end{aligned} \tag{20}$$

where $E(R^{RMW})$ is the difference in expected return between robust operating profitability and weak operating profitability firm portfolios, and $E(R^{CMA})$ the difference in expected return between conservative investment and aggressive investment firm portfolios.

The Fama French Three Factor model and Fama French Five Factor model are just two examples of factor models that has been developed since CAPM. Harvey et al. (2016) has identified over 300 published factors that describe the expected excess return of assets. John Cochrane in Cochrane (2011) refers to this as a “factor zoo” and challenges researchers to, instead of researching new potential factors, find which factor independently explains return.

4 Data & Methodology

4.1 Model comparison and evaluation

We select sparse sets of firm characteristics by applying Principal Component Analysis and Least Absolute Shrinkage and Selection Operator techniques. The choice of using *PCA* and *LASSO* is not only based on the fact that these two methods has been used earlier to similar problem as a lot of methods fall into this category. The choice was also influenced by the fact that it offers a comparison analysis between *unsupervised* and *supervised* learning. Lastly, the methods were chosen based on their relative simplicity compared to other methods listed in the literature review, which offers greater interpretability of the results. To be able to measure if these techniques are successful we need a model evaluation procedure.

We apply our different techniques of variable selection and compare their predictive performance on an independent test sample. Following the examples of Freyberger et al. (2016), Green et al. (2017) and Kelly et al. (2019), the performance of out-of-sample predictions will act as a basis. Our data will be divided into two samples, one training sample and one test sample. The firm characteristics of the Fama French Three Factor model and the Fama French Five Factor model will be used as a performance benchmark. That is, we will not use the portfolio sorting approach that is used by Fama and French, instead we will include the same firm characterising that Fama and French uses to construct portfolios and model these. The factors from the Fama French Factor models are **beta**, **size**, **book-to-market**, **investments** and **profitability** (Fama & French 1993, 2015).

The model comparison procedure contains the following steps and is done for each model.

1. Perform the specific technique for variable selection on the training set data.
2. Regress log-return on the factors using the training set, with OLS regression.
3. Generate predicted values on the test set.
4. Measure deviation of predicted values from observed values in the test set.

As measurement, we will use *mean squared prediction error (MSPE)* which is calculated by

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{train})^2, \tag{21}$$

where n is the number of observations in the test sample, y_i is the log-return of observation i, where $i = 1, 2, \cdots, n$ and are all in the test set, and $\hat{y}_i^{train}$ denotes the predicted value of log-return based on model done on the training set.

To put this in the context of asset pricing theory the model comparison procedure has the following steps.

1. Select which risk factors are of interest.
2. Estimate the corresponding average exposure that risk factor has on each stock, as in Equation (17).
3. Use this exposure to predict asset returns on new data.
4. Measure the difference between predicted and real returns.

The following parts of this section will cover the necessary information of how these results will be gathered. First off, the required data will be covered, together with sample selection and data trimming. Then both PCA and LASSO regression will be covered.

4.2 Data

Our data consist of quarterly observations of 54 variables for 500 constituents of the S&P 500 index, all of which are listed companies in New York Stock Exchange (NYSE) or NASDAQ. A vast majority of the firms characteristics are derived from quarterly reports, hence; we use quarterly observations. The variables consists of 54 firm characteristics and a dependent variable. Just as our precursor Green et al. (2017), the *logreturn* variable $(\log(p_t^i) - \log(p_{t-1}^i))$ is used as dependent variable. The 54 chosen firm characteristics are listed in Freyberger et al. (2016) and Green et al. (2017) and were all the characteristics that we were able to collect using the database of Capital IQ. Each of the characteristics has proved to significantly explain return in past research. Every firm characteristic is listed in Appendix 1 together with corresponding explanation and calculation. For more information regarding theses characteristics, see Freyberger et al. (2016) and Green et al. (2017). The observations range from 2009-07-01 to 2019-07-01, and are, as mentioned, gathered quarterly.

Before starting the analysis, some data cleaning needs to be done. The data include some missing values (N/A), which, in most cases, are due to limitations in the database. These values will be set to the post-standardized mean of zero, which has been done in Green et al. (2017) as well. Some of our variables has a large number of missing values. In cases of over 100 missing values, the variable is removed as the amount of missing values will affect the results. After dropping these variables, we have 48 variables left. The dropped variables are listed in Appendix 1. This conclude our data cleaning. Next up is the data screening process. After looking at the distributions of the variables, which were highly skewed and contained a lot of outliers in almost all the cases, we decided to use a technique called *winsorizing*, just as Green et al. (2017). Winsorizing set the values that are smaller than the 5th percentile to the value of the 5th percentile and the values that are larger than the 95th percentile to the value of the 95th percentile. This is done on each factor.

Since the selected firms were taken from the list of constituents of the S&P 500 index as of today, the firms might not have been listed for the entire period that is used. For this reason we chose to drop the firms that have not been listed the entire 10 year period. 39 firms were removed based on this.

When the data is ready to use, it is divide into two parts, a training set and a test set. The training set contains data from 2009-07-01 to 2016-07-01 and the test set contains data from 2016-10-01 to 2019-07-01. The training set contains the data which all models are constructed by. The LASSO regression is conducted on the entire training set. The PCA is conducted on the firm characteristics of all firms on the last day in the training set, that is, 2016-07-01. In the model comparison stage the test data is used to calculate the MSPE-values.

4.3 Principal Component Analysis

Principal Component Analysis (PCA), developed by Pearson (1901), is a widely used tool that enables the user to transform high dimension data into a small set of factors that manages to sustain the representation of the variability of the data. In simple terms, it can be seen as a tool for removing repetitive and redundant dimensions. PCA falls into the category of *unsupervised learning* (James et al. 2013). This means that it conducts a dimension reduction without taking the effect on return predictability into account.

That a method which falls into the category of unsupervised learning is applied in a prediction context might raise suspicion. The following examples illustrates how PCA can be useful for finding relevant factors even though it does not directly target returns. Let us consider some asset characteristics that co-move with the market in some systematic way. These characteristics exhibit covariance, i.e. if x_1 goes up one can expect x_2 to also go up. But this is because they both are driven by an unobserved factor “market”. If one were to know how the market factor maps into x_1 and x_2, one can construct proxies of x_1 and x_2 using information about the unobserved factor the “market” factor. PCA plays it’s part by extracting the unobserved market factor from the covariance matrix of x_1 and x_2 and explaining it as a linear function of x_1 and x_2. The characterisation is easy and intuitive, but this is not true for most cases.

PCA provides us with independent unobserved factors, which are hard to interpret since they contain loadings of all independent variables. The objective of this thesis contain variable selection. This will not be fulfilled by just stating the loadings of the unobserved factors. Instead, our approach is to try to "proxy" the most important PCs by factors. We have chosen two different methods for the task of deciding which variables should act as proxies. The technique of choosing these factors is an original contribution by this thesis.

To use PCA to extract variables that explains a dependent variable is highly dependent on the assumption that what explains most of the variation of the independent variables also explains the most of the variation of the dependent variables. This is, of course, something that one can not know beforehand. Although, since all of our independent variables are selected based on the fact that it has been shown to significantly explain return in one or more research papers (Freyberger et al. 2016, Green et al. 2017), we believe it to be correct to argue that that assumption is correct in this case.

We now cover the technical aspects of PCA. Suppose one have a data matrix $\boldsymbol{X}$ consisting of n observations and m variables,

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}. \tag{22}$$

Assume that one have insufficient economical arguments for characterisation and grouping of the data. PCA constructs a number of independent n-sized eigenvectors called *principal components (PC)*, earlier referred to as unobserved factors (James et al. 2013). The first PC explains the most variability of $\boldsymbol{X}$ out of all PCs, the second PC explains the second most, and so on. As the investigation of the variability of $\boldsymbol{X}$ is desired, one can use $Var(\boldsymbol{X})$, and in matrix form, this is called the covariance matrix, Σ_x. If the variables are in different units of measurement, the magnitude might affect the outcome of the PCA. It is therefore a common practice to standardize the elements, which is done by taking the difference between the value of the element and the mean, divided by the standard errors. If this is done on $\boldsymbol{X}$ to get $\boldsymbol{X_s}$ the result is that the covariance matrix Σ_{X_s} is equal to a correlation matrix.

Every PC is associated with a positive numbered eigenvalue. With the help of the eigenvalue the total variance explained by that PC can be calculated. Next, one have to decide how many PCs that should be kept to still be able to efficiently explain the variability. Common rules of thumb for the selection of the number of PCs are *Kaiser's Rule* (Kaiser 1960), *cumulative variability* and *inflection point* (James et al. 2013). By using the Kaiser's rule, the number of PCs selected should be based on the number of PCs with eigenvalues over 1. The cumulative rule instead bases the selection on the cumulative variability of the selected PCs, based on a arbitrarily chosen threshold. The inflection point method uses a scree plot with eigenvalues on the y-axis and the PCs on the x-axis. The number of PCs should then be based on the the inflection point of the scree plot, that is where the plot goes from convex to concave (James et al. 2013).

We now move on to the techniques of variable selection. The first method that we use to decide which factors should act as proxies contain the procedure to choose proxies for each PC. We will regress each PC on each variable and pick the factors which model receive the highest R^2 value. We will not limit the proxy to just a single variable for each PC, instead, we will pick three variables as proxies. This choice is based on weighting the explanatory power of the proxy on the corresponding PC, which increases with more variables, and simplicity of the end model. The approach for picking proxies is closely related to stepwise selection, for more information see James et al. (2013). We are aware of the potential local minimum that it may return. The methodology is presented in the Algorithm 1 below.

Algorithm 1: Pseudo code for the selection of variables

```
For each PC, $i$, selected , select the variable $x_c$ that has the highest
  absolute correlation with $i$, $j = 1$;
while $j < K$ do
    Select variable $x_j$;
    if $x_j \neq x_c$ then
        regress $i = x_c + x_j$ if $R^2$ is higher then previously seen, save $R^2$ as
          $R^2_{sec}$ and $x_j$ as $x_{sec}$;
        set $j = j + 1$
    else
        set $j = j + 1$
set $j = 1$;
while $j \neq K$ do
    Select variable $x_j$;
    if $x_j \neq x_c$ and $x_j \neq x_{sec}$ then
        regress $i = x_c + x_{sec} + x_j$ if $R^2$ is higher then previously seen, save
          $R^2$ as $R^2_{third}$ and $x_j$ as $x_{third}$;
        set $j = j + 1$
    else
        set $j = j + 1$
```

This algorithm provides the factors which can explain the most for each PC and further the highest explanatory power for our data set. As one can see from the Algorithm 1, this procedure continues for each PC until one gets the three variables with the highest R^2 in each regression. The reason we do not use more variables in our regression is that our goal is dimension reduction; hence, we should try with as few variables possible to get explanatory power.

We will also conduct a second variable selection based on PCA, which also uses a method influenced by stepwise selection. Our first approach select variables which act as proxies for each of a number of PCs. This second approach will instead be based on the explanatory power of the variables on the entire data set. Recall that each PC has a corresponding percentage explanation of the total variation in the data. With the help of OLS regression, which is used in the first approach as well, one can receive a R^2 value of each variable on each PC. Let us use variable x_1 and principal component PC_1 as an example. If one denotes the percentage explanation of the total variation given by PC_1 by $PC_1^{\%}$ and the R^2 value from the regression $PC_1 = \alpha + \beta x_1 + \epsilon$ as R^2_{1,PC_1}. Then one can calculate the percentage explanation of the total variation given by variable x_1 through PC_1 by

$$x^{\%}_{1,PC_1} = PC_1^{\%} \cdot R^2_{1,PC_1}. \tag{23}$$

Since each of the principal components are independent of each other, the cumulative variance explained of the PCs is equal to 1. The PCs independence also means that

$$x^{\%}_{j,PC_i} + x^{\%}_{j+1,PC_i} + \cdots + x^{\%}_{K,PC_{i+1}} + \cdots = 1 \quad \text{for } j = 1, 2, \cdots, K \text{ and } i = 1, 2, \cdots, K \tag{24}$$

where K is the total number of variables, and, hence, total number of principal components. Furthermore, one can conclude that the total variance explained by variable x_i is given by

$$x^{\%}_i = \sum_{j=1}^{K} \left(PC^{\%}_j \cdot R^2_{i,PC_j} \right). \tag{25}$$

Equation (25) is what this second approach is based upon. Instead of pursuing a stepwise selection to proxy some variables for each principal component, the stepwise selection is pursued to proxy the entire set of principal component. This stepwise selection is presented in Algorithm 2 below.

Algorithm 2: Pseudo code for the selection of variables

```
while $j =< K$ do
    Select variable $x_j$;
    calculate $x^{\%}_j$;
    if $x^{\%}_j$ is higher than any previously seen then
        save $x_j$ as $x_{first}$;
    else
    set $j = j + 1$
```

The algorithm is then applied to include multivariate regression. That is, we will do the steps taken in Algorithm 2 and include x_{first} and then save the x_j which has the highest $x^{\%}_j$-value in combination with x_{first}, then save that x_j and include it, together with x_{first} to calculate a $x^{\%}_j$ again. This is repeated for some number of variables.

4.4 LASSO Regression

A relatively recent developed method, compared to PCA, is the LASSO regression. First popularized by Tibshirani (1996) because of its ability to generate interpretive results by using *shrinkage.* Shrinkage, according to Everitt & Skrondal (2002), refers to the reduction of model prediction when fitted on a new model. More particularly, it refers to a shrinkage in a variable coefficient. LASSO can also be used for dimension reduction where it performs variable selection. By using a shrinkage method LASSO receives a lower prediction error. To be able to understand the theory of LASSO, its advantages and applications, one will need to first cover the *bias-variance tradeoff.*

When creating models to be able to predict some dependent variables based on some independent variables the objective is to reduce the deviance of the fitted value $\hat{y}$ from the real value y (James et al. 2013). Consider the desired model $y = f(x) + \epsilon$. Our objective is to find the best model $\hat{y}(x) \approx f(x)$. Since the model concerns predictions, one would want to look at how well the data performs on new data. Let us consider a test set for this task. To evaluate the performance of $\hat{y}$ one can use,

$$E(y_0 - \hat{f}(x_0)) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2, \tag{26}$$

where $E(y_0 - \hat{f}(x_0))$ denotes the expected test MSE. The test MSE can be shown to be a result of three essential properties in statistical learning according to James et al. (2013). One can decompose the MSE, given data x_0, into

$$E(y_0 - \hat{f}(x_0)) = Var(\hat{f}(x_0)) + Bias(\hat{f}(x_0))^2 + Var(\epsilon). \tag{27}$$

The bias of $\hat{f}$ refers to the systematic error of that is contributed by $\hat{f}$. The variance of $\hat{f}$ is the amount that $\hat{f}$ changes when using new data. Since the variance of the error term ϵ is irreducible, the minimization of expected test MSE is only achieved by simultaneously minimizing the variance and bias of $\hat{f}(x_0)$. This is what is commonly referred to as the *bias-variance tradeoff* (Hastie et al. 2009).

As mentioned in the data section, we have a great amount of variables. In these cases, the bias is relatively low but the variance is high since the more variables included in a model, the more adaptive the model is to the training set. A great way of decreasing the variance, according to James et al. (2013) is to *shrink* the coefficients to some mean. One of the most used shrinkage methods is, again according to James et al. (2013), the LASSO regression.

The advantage of LASSO regressions compared to other shrinkage models is the fact that it can force coefficients to zero (James et al. 2013). This means that in a multivariate regression including various variables LASSO perform variable selection because it has the ability to force coefficients with low predictable power to zero. A LASSO regression yields a *sparse* model which means that only some variables from the original ones set will be used (James et al. 2013). The LASSO regression is defined as

$$\hat{\beta}^{LASSO} = \underset{\beta}{\text{argmin}} \sum_{i=t}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t \tag{28}$$

Or in Lagrangian form as

$$\hat{\beta}^{LASSO} = \underset{\beta}{\text{argmin}} \left\{ \frac{1}{2} \sum_{i=t}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{29}$$

One can notice the penalty term, $\lambda \sum_j |\beta_j|$, in the LASSO regression on the right hand side of Equation (29). LASSO is, compared to another shrinkage method, the *Ridge Regression*, easier to interpret because of its variable selection. The difference between LASSO and Ridge regression are presented in Figure 1.

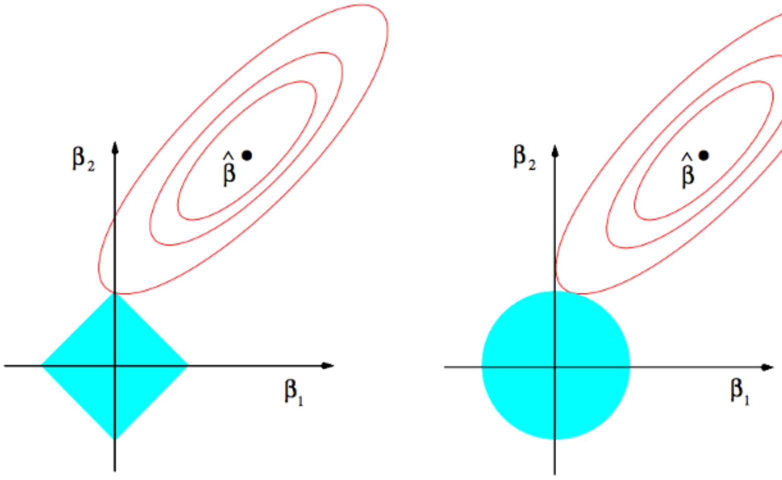
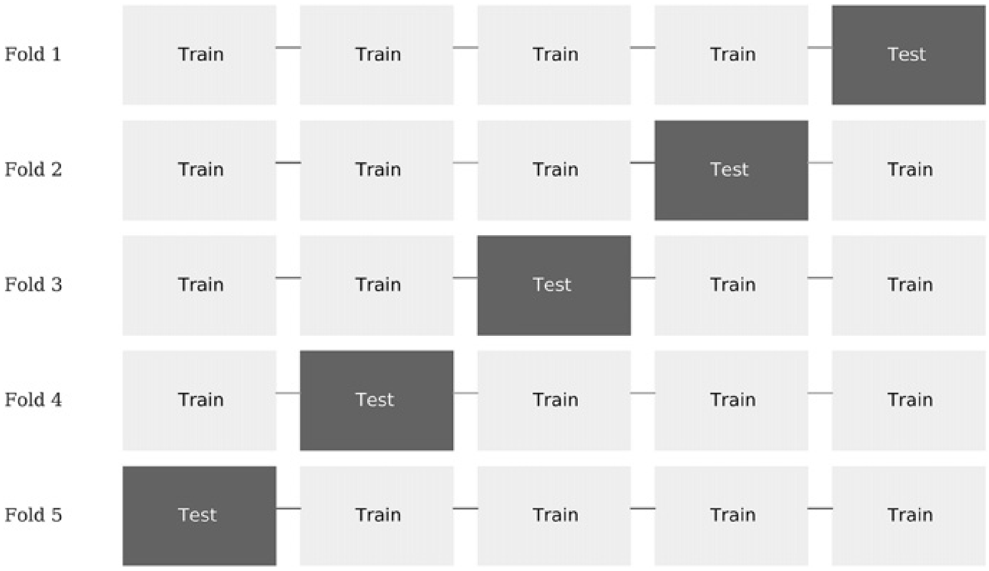


Figure 1: Figure on the right illustrates the Ridge regression penalty term. The figure on the left illustrates the LASSO regression penalty term. Source: James et al. (2013) p. 222.

The blue circle represents the Ridge regression constraint and the blue diamond the LASSO constraint. As one can see that they differ by the way that LASSO has a *corner solution.* The ellipse represents the least squared error function where all points on the same ellipse represent the same error (James et al. 2013). Both methods will identify where the ellipse tangents the constraint. Since Ridge has a circular constraint the coefficient estimates will exclusively be non-zero as they can not intersect at the axis. In contrast, LASSOs diamond constraint has corners at both axes, the ellipse can intersect at those regions, which results in the coefficients equal zero. LASSO also tends to do better in reducing variance at a small expense of bias compared to Ridge when dealing with high variance data sets (James et al. 2013).

LASSO regression contains a penalty term which is dependent on the variable λ. λ will be selected based on the best fitted model, call this λ_{min}. In order to select a more parsimonious model, one can select the λ that is one standard error away, the λ_{1se} (James et al. 2013). We will use both and construct two models. To determine λ_{min} and λ_{1se}, we will use k-fold cross-validation, a method that uses two sets of temporary chosen in-sample and out-of-sample data, both are still part of the training set, with k folds. With k-fold cross-validation, the data is divided into k sets. Then, for each k, $k-1$ of the sets form the training set and 1 set form the test set. A model is then fitted on the training set and then predicted on the test set. Lastly, the model is evaluated by summarizing these k iterations (Hastie et al. 2009). An illustration of the fold and sets are illustrated in Figure 2 with $k = 5$. We will use a 10-fold cross-validation procedure in our LASSO regression. This will be done in R using the package *glmnet.*



Fold 1	Train	Train	Train	Train	Test
Fold 2	Train	Train	Train	Test	Train
Fold 3	Train	Train	Test	Train	Train
Fold 4	Train	Test	Train	Train	Train
Fold 5	Test	Train	Train	Train	Train

Figure 2: 5-fold cross-validation sample setup. Source: De Prado (2018) p. 104.

The extracted optimal λs is then used in LASSO regressions. The factors with non-zero coefficients in these LASSO regression are the selected factors, which can be expressed as using the post-LASSO method. Our LASSO regression will be applied to a OLS model. This means that the same assumption that applies to OLS applies to LASSO regression. Our main concern, as we use time series data, is the assumption of *stationarity*, i.e., that the distribution does not change over time (Tsay 2010).

For times series regression one needs to take into account two key assumption, stationarity and *weak time dependence*. Stationarity is the foundation in time series analysis where one assumes that the joint distribution is identical over time. In the financial literature, it is common to assume the asset return to be *weak stationary* (Tsay 2010). Weak time dependence on the other hand means that a variable at time t_0 can not predict the value for the variable at t_1. In the financial literature one assume that a stock follows a *random walk* (Cochrane 2005), which implies weak time dependence.

5 Results

5.1 Principal Component Analysis

The scree plot of the eigenvalues for the respective PC on the y-axis and the PC on the x-axis are displayed in Figure 3 and the cumulative variability plot is displayed in Figure 4.

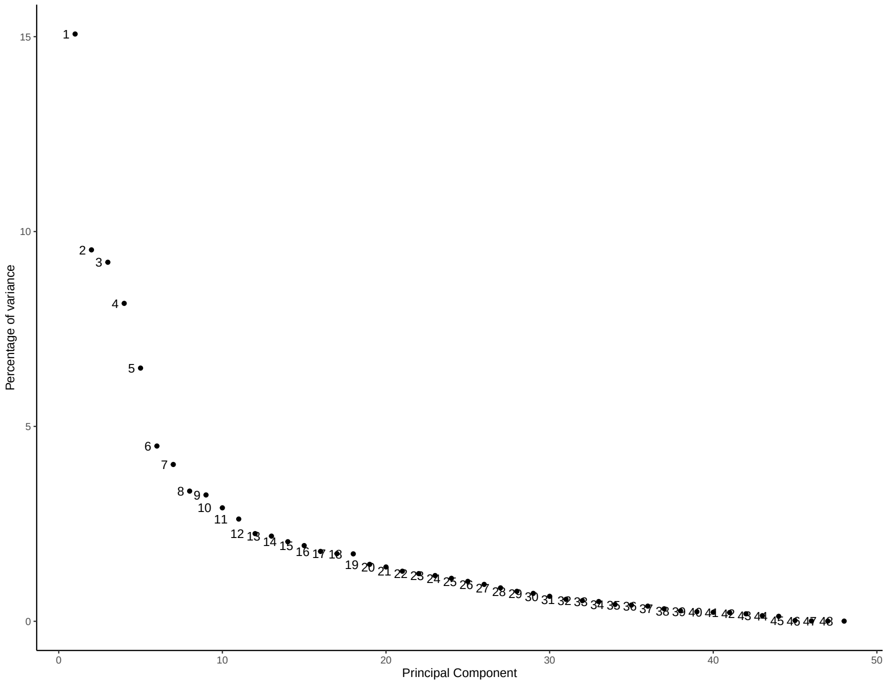


Figure 3: Scree plot of percentage of variance explained for each PC.

As one can see in Figure 3 the first PC has a relatively high eigenvalue when compared to the following PC. The same marginal drop is also seen after PC 5, which also seems to be the inflection point. If we instead look at the cumulative variability we get the sum of the explained total variance in the data for each PC and its precursors. The cumulative variability is presented in Figure 4.

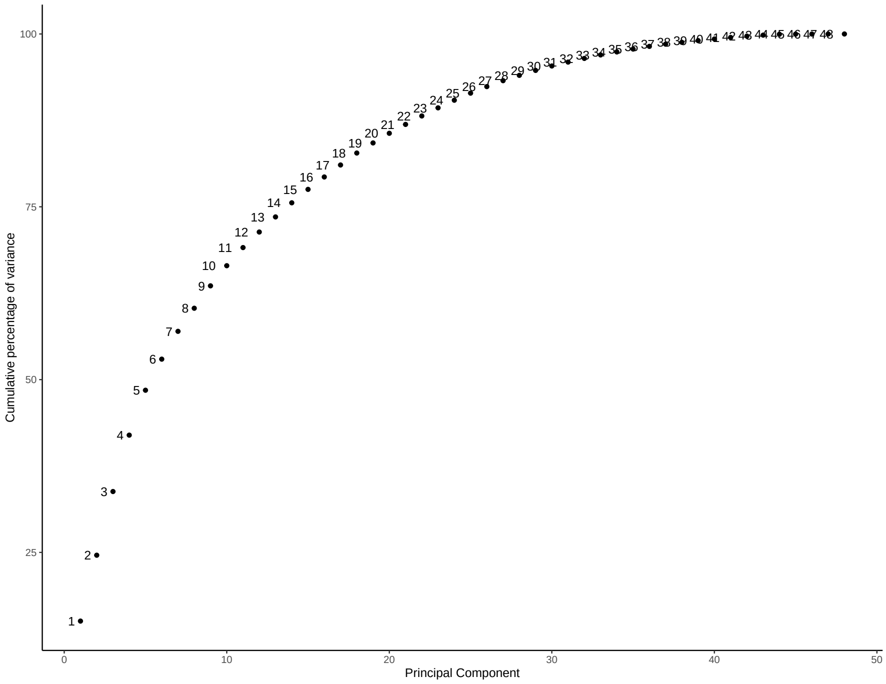


Figure 4: Cumulative percentage of variance explained by each PC.

Based on the findings presented in Figure 4 one can note that the five PCs with the highest eigenvalues explain around half the variance in the data set. The third way of choosing PCs is to select all PCs with eigenvalues over 1. In our case, the PCs with eigenvalues over 1 is in total 16. As our objective is dependent on picking a few independent factors, we decided to not use this approach. Instead we use the inflection point method, while considering adequate cumulative explained variability. This result in the selection of 5 PCs. The firm characteristics were extracted using our two methods. Algorithm 1 creates the first PCA model, later called PCA_1. With the procedure of PCA_1 we extract three factors for each chosen PC that will act as proxies for these PCs. Each chosen PC, the corresponding factors and the R^2 are presented in Table 1.

	Selected factors			
PC	1	2	3	R^2
1	roaq	a2me	sgr	0.89
2	quick	rel.to.high	salecash	0.79
3	ep	agr	mom36m	0.71
4	chcsho	gma	agr	0.75
5	betasq	mve	gma	0.63

Table 1: The PCs, the corresponding factors that result in a regression with the highest R^2 from the method of stepwise regression, and the corresponding R^2.

From Table 1 we can see that three factors for each PC seem to explain each PC by a sufficient amount by looking at the R^2. One can also see how both **agr** and **gma** has been selected for two PCs. This leaves us with 13 factors selected from this method. Notable is that **roaq**, **a2me** and **sgr** explain 89% of PC 1 which was the PC with the highest eigenvalue. These three factors seem to capture a lot of the variability. The regression on PC 5 has the lowest R^2. From Equation (25) we calculate the total variance explained by the model that contain these 13 factors and end up with 43.96%.

The second model, later called PCA_2, uses the total explained variability stepwise method, explained in Algorithm 2. In Figure 5 the total explained variability is presented.

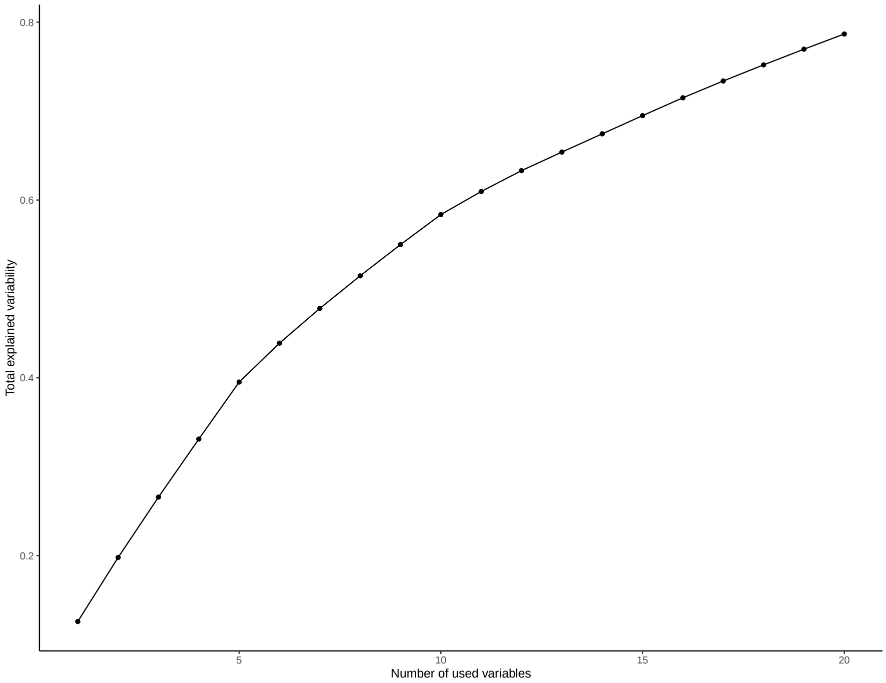


Figure 5: Total explained variability by a number of factors.

We can in Figure 5 see how a small set of factors manage to generate explanatory power close to 50% of the total variability. Figure 5 compared to Figure 4 shows real factors instead of generated PCs. From this, we choose to extract 8 factors as the marginal explanation drops after 8 factors. Again, the objective of the thesis requires us to weight explanation and simplicity. As these 8 factors explain 49.75% of the variability, which is sufficient, and the marginal explanation drops, we make the decision of choosing 8 factors. The factors extracted are: **a2me** (assets-to-market capitalization), **agr** (asset growth), **cto** (capital turnover), **quick** (quick ratio), **roaq** (return on assets), **rd_sale** (R&D to sale), **rel.to.high** (price to 52 week high price) and **tb** (tax income to book income).

5.2 LASSO Regression

First off, the λs are chosen using 10-fold cross-validation and the result are presented in Figure 6 below.

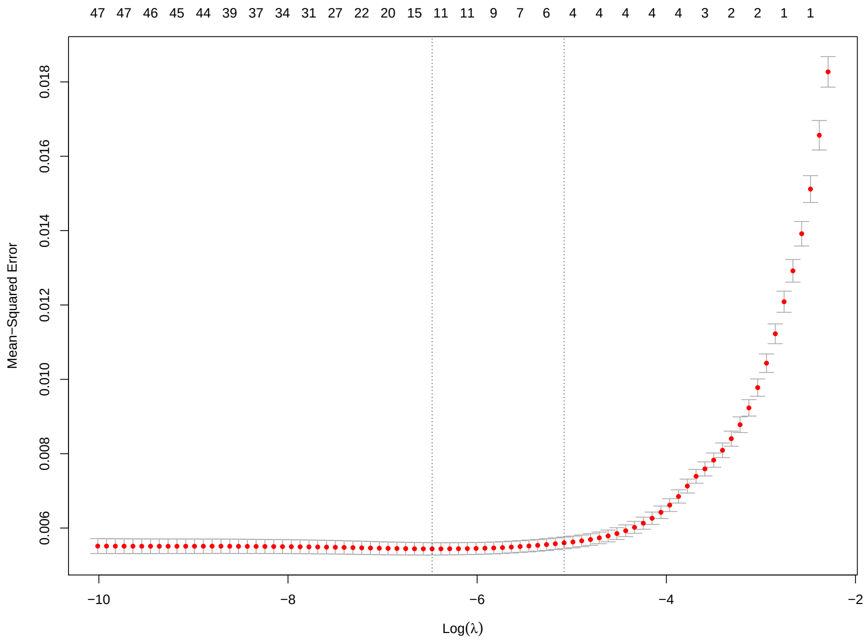


Figure 6: Cross validated λ with corresponding MSE. The bottom x-axis shows the $\log(\lambda)$ value and the top x-axis shows the corresponding number of non-zero coefficients. The two dotted lines are the λ_{min} and λ_{1se} values.

Figure 6 shows the selected tuning parameter (λ) from the cross-validation. One can see how the MSE increases when the variable selection generate fewer characteristics. The red dots represent the MSE for corresponding log λ and the grey bars represent the confidence interval. The λ_{min} generated a model containing 12 factors and the λ_{1se} generates a model containing only 6 factors.

The characteristics generated from LASSO λ_{min} is **agr** (asset growth), **beta** (beta), **currat** (current ratio), **dy** (dividend to price), **lev** (leverage), **mom12m** (12-month momentum), **mom1m** (1-month momentum), **mom36m** (36-month momentum), **mom6m** (6-month momentum), **quick** (quick ratio), **rel.to.high** (price to 52 week high price) and **tang** (debt capacity/firm tangibility).

The characteristics generated from LASSO λ_{1se} is **dy** (dividend to price), **mom12m** (12-month momentum), **mom1m** (1-month momentum), **mom36m** (36-month momentum), **mom6m** (6-month momentum) and **rel.to.high** (price to 52 week high price).

5.3 Model comparison

The result from out-of-sample prediction by each of the four created models and two benchmark models are presented in Table 2 below.

Method	MSPE	Number of factors
PCA_1	0.009964	13
PCA_2	0.009938	8
$LASSO_{\lambda_{min}}$	0.005006	12
$LASSO_{\lambda_{1se}}$	0.005003	6
FF3	0.017297	3
FF5	0.017323	5

Table 2: Mean square prediction error values. Calculated with out-of-sample set of quarterly log-returns from 461 firms in 2016-10 to 2019-07.

With the measure of MSPE, $LASSO_{\lambda_{1se}}$ is the best performing model. All constructed model also outperforms the two benchmark models. Overall, the two LASSO models outperform both the benchmark and the PCA models.

6 Discussion

Both LASSO and PCA are successful in the context of minimizing the mean squared prediction error as the four models based on these techniques outperform the benchmark. However, as a part of the objective is to build sparse models, the value of parsimoniousity should not be ignored. Each of our models are larger than the two benchmark models. On the other hand, these results might indicate that the two established models are too parsimonious and more factor should be included. Still, a key takeaway from the in-group results, meaning the difference of results in the technique models, is that in each technique, the more sparse model outperforms the less sparse model; PCA_2 outperforms PCA_1 and $LASSO_{\lambda_{1se}}$ outperforms $LASSO_{\lambda_{min}}$. This is most likely a consequence of reduction in variance of the models.

Our main concern with the results are the lack of similarity of selected factors between the methods. The only variable that is included in all four models is price relative to 52-week-high-price (rel.to.high). The 36-month momentum is included in three out of four models. In total, the four models have extracted 23 unique firm characterises. These results indicate that there are no distinguished patterns found and suggests that the factors selected in the best model does not have to be superior to any other model. That is, the selection is not stable. However, it could indicate that the unsupervised learning method of PCA is not adequate for the task at hand and should not be included in such an analysis.

When comparing the extracted characteristics from our PCA and LASSO methods to existing factor models such as *Fama French Three Factor model*, *Carhart Four Factor model*, *Fama French Five Factor model* and the *q-Factor model*, one can see that out of the 7 factors included in these existing models, 4 are included in at least one of our four models. These are, market beta, momentum, size and profitability. Missing are the factors book-to-market (Fama & French 1993, Carhart 1997, Fama & French 2015, Hou et al. 2015), return on equity (Hou et al. 2015) and investments (Hou et al. 2015, Fama & French 2015).

Freyberger et al. (2016) receives significant power for market capitalization and investments as mention earlier. The only model which extracts market capitalization is PCA_1, none of our other models extract these variables. As Green et al. (2017), we extract the size factor. The size factor from Green et al. (2017) is together with 12-month momentum the most powerful predictors in their paper but we only extract it from one of our four models, the PCA_1.

Looking at the apparent superior technique of LASSO regression one can see that various variables which falls in the category of momentum are represented. This is in line with Jegadeesh & Titman (1993)'s argument for a momentum factor in 1993 and confirms the validity of inclusion of the momentum factor in the factor model of Carhart (1997). Our results also reaffirm the results of Freyberger et al. (2016), Green et al. (2017) and Kelly et al. (2019) which all find the momentum factor as a highly predictable characteristic of average asset return. To put this into the context of asset pricing, momentum seems, from the results, to proxy marginal utility growth well.

When considering the research question of this thesis, an obvious answer to it could be the constituents of $LASSO_{\lambda_{1se}}$: **dy** (dividend to price), **mom12m** (12-month momentum), **mom1m** (1-month momentum), **mom36m** (36-month momentum), **mom6m** (6-month momentum) and **rel.to.high** (price to 52 week high price). However, this would not be correct as, as mentioned earlier, the results lack consistency in variable selection. Moreover, as with all empirical work, one has to be careful before drawing any conclusion. The findings are always the result of the used data, the method and the design of the tests. We have used an out-of-sample test which estimates the βs (the risk exposure) of the factors and then keep them constant to be able to predict asset prices over a period of over two years. Hence, the result of the predictive power of momentum factors might benefit from the fact that the risk exposure towards momentum perhaps are more stable over time compared to other factors.

6.1 Robustness

Hou et al. (2015) notes the alleviation of the impact of microcaps. Even though the number of microcap stocks are a majority of the stocks in the index, the total value of these stocks are insignificant. As transaction costs and the probability of illiquidity are much greater in these stocks, anomalies are less likely to be exploited in practice. Even though our set of firms does not include any microcaps, the concern might still be the same in our study. As a test for not overweighing smaller firms in our research, we conduct a robustness test for verifying the variable selection but only use the 100 largest companies in the S&P 500 index. The technique of PCA_1 resulted in 12 selected factors. 8 out of these factors were included in the original model as well. The originally chosen factors which were now excluded are the following: **quick**, **rel.to.high**, **ep**, **mve**. The technique of PCA_2 resulted in 7 selected factors. Only 3 out of these were included in the original model. These are **a2me**, **agr** and **cto**.

The discovery of momentum being a highly predictive factor category is shown to be robust. In both LASSO regressions various momentum factors were selected. When conducting the model comparison the LASSO regression once again outperformed, and once again $LASSO_{\lambda 1se}$ was the best model. For more details of the results of the robustness test, see Appendix 2.

6.2 Limitations

There are many areas in which we had to limit our research, both in aspects of limited data but also to limit the magnitude of the thesis. The fact that we only use 54 variables instead of a larger sample for analysis, as Harvey et al. (2016), which was a result of limited data, may influence our results. The set of not selected variables might contain variables with stronger predictive power. It might also affect the size of our created models. For example, the required number for selected PCs might be larger as the variability would most certainly increase. By increasing the time period one may also receive different results of independent characteristics.

Furthermore, our time series runs over 10 years, which, compared to other papers for example Freyberger et al. (2016) whose data consist of the time period 1964 to 2014, is a small amount of time. This may affect the results since market fluctuations affect our analysis more because of the shorter time.

In the context of comparing supervised and unsupervised learning methods, one has to consider the limitation of just one model of each are used. To be able to make these results more robust more methods would have to be added. This is also true for the overall results. To draw more certain conclusions about the selected variables, one can add more methods for factor selection. This is always true as one can always do more.

6.3 Future research

In the up to date literature of asset pricing theory, Cochrane (2011)'s question still remains; "which firm characteristics provide independent information about average return?" Our approach with PCA and LASSO resulted in characteristics that still needs a lot of verification before drawing conclusions. By including more variables and conduct the analysis on a larger random sample and over different markets for comparison one can yield more interesting results.

There are also improvements in our used techniques that could be employed. An obvious example is to improve the method used in PCA_1 and PCA_2. Instead of looking at the independent variable's variance explained by each PC, one may instead first run a Principal Component Regression on log return and base the selection of PCs on that and then proxy them.

7 Conclusion

We have proposed two different machine learning methods to be applied to the challenge posed by Cochrane (2011). One unsupervised method (PCA) and one supervised method (LASSO) for variable selection. Our results are twofold. First, the model comparison shows how LASSO tend to do better then PCA in both selecting fewer variables and perform better out-of-sample. Second, the predictable power of momentum factors has been further supported. Furthermore, even though the usage of the unsupervised learning method of PCA applied on a supervised learning problem is not obvious, we have managed to show that the method is viable as it outperformed the benchmark.

The scope of this thesis has been to use two different methods to generate sparse factor models in order to predict asset returns. What we have found is how LASSO tend to choose various momentum factors while PCA choose more diversified characteristics. In order to minimize MSPE the predictable power of momentum has shown to work best. Our contribution to the literature of asset pricing is a further reaffirmation of the predictable power of momentum factors, which is in line with previous studies (Freyberger et al. 2016, Green et al. 2017, Kelly et al. 2019).

In the method of PCA, we have created an original contribution for proxying factors for PCs. Although the technique of unsupervised learning seems to be inferior to the supervised learning at hand, our method can be adjusted to a supervised learning method, as proposed in the previous section.

8 References

Banz, R. W. (1981), ‘The relationship between return and market value of common stocks’, *Journal of financial economics* **9**(1), 3–18.

Bodie, Z., Kane, A. & Marcus, A. J. (2018), *Investments*, 11 edn, McGraw-Hill Education, 2 Penn Plaza, New York, NY 10121.

Carhart, M. (1997), ‘On persistence in mutual fund performance’, *Journal of Finance* **52**(1), 57–82.

Cochrane, J. (2005), *Asset Pricing*, Princeton University Press, Princeton, New Jersey.

Cochrane, J. (2011), ‘Presidential address: Discount rates’, *Journal of Finance* **66**(4), 1047–1108.

Daniel, K. & Titman, S. (1997), ‘Evidence on the characteristics of cross sectional variation in stock returns’, *The Journal of Finance* **52**(1), 1–33.

De Prado, M. L. (2018), *Advances in financial machine learning*, John Wiley & Sons.

Everitt, B. & Skrondal, A. (2002), *The Cambridge dictionary of statistics*, Vol. 106, Cambridge University Press Cambridge.

Fama, E. F. & French, K. R. (1992), ‘The cross-section of expected stock returns’, *The Journal of Finance* **47**(2), 427–465.

Fama, E. F. & French, K. R. (2004), ‘The capital asset pricing model: Theory and evidence’, *Journal of economic perspectives* **18**(3), 25–46.

Fama, E. & French, K. (1993), ‘Common risk factors in the returns on stocks and bonds’, *Journal of Financial Economics* **33**(1), 3–56.

Fama, E. & French, K. (2015), ‘A five-factor asset pricing model’, *Journal of Financial Economics* **116**(1), 1–22.

Feeney, G. J., Hester, D. D. et al. (1964), Stock market indices: A principal components analysis, Technical report, Cowles Foundation for Research in Economics, Yale University.

Feng, G., Giglio, S. & Xiu, D. (2017), ‘Taming the factor zoo’, *Chicago Booth research paper* (17-04).

Freyberger, J., Neuhierl, A. & Weber, M. (2016), 'Dissecting characteristics nonparametrically', *The Review of Financial Studies* **33**(5), 2326–2377.

Green, J., Hand, J. & Zhang, F. (2017), 'The characteristics that provide independent information about average u.s. monthly stock returns', *The Review of Financial Studies* **30**(12), 4389–4436.

Gu, S., Kelly, B. & Xiu, D. (2020), 'Empirical asset pricing via machine learning', *The Review of Financial Studies* **33**(5), 2223–2273.

Harvey, R., Liu, Y. & Zhu, H. (2016), '… and the cross-section of expected returns', *The Review of Financial Studies,* **29**(1), 5–68.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

Haugen, R. & Baker, N. (1996), 'Commonality in the determinants of expected stock returns', *Journal of Financial Economics* **41**(3), 401–439.

Hou, K., Xue, C. & Zhang, L. (2015), 'Digesting anomalies: An investment approach', *The Review of Financial Studies* **28**(3), 650–705.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013), *An Introduction to Statistical Learning*, Springer.

Jegadeesh, N. & Titman, S. (1993), 'Returns to buying and selling losers: Implications for stock market efficiency', *Journal of Finance* **48**(1), 65–91.

Kaiser, H. F. (1960), 'The application of electronic computers to factor analysis', *Educational and psychological measurement* **20**(1), 141–151.

Kelly, T, B., Priutt, S. & Su, Y. (2019), 'Characteristics are covariances: A unified model of risk and return', *Journal of Financial Economics* **134**(3), 501–524.

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.

Light, N., Maslov, D. & Rytchkov, O. (2017), 'Aggregation of information about the cross section of stock returns: A latent variable approach', *The Review of Financial Studies* **30**(4), 1339–1381.

Litner, J. (1965), 'Security prices, risk, and maximal gains from diversification', *Journal of Finance* **20**(4), 587–615.

Markowitz, H. (1952), 'Portfolio selection', *The Journal of Finance* **7**(1), 77–91.

Mossin, J. (1966), 'Equilibrium in a capital asset market', *Econometrica: The Econometric Society* **34**(4), 768–783.

Pearson, K. (1901), 'Liii. on lines and planes of closest fit to systems of points in space', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**(11), 559–572.

Rapach, D. E., Strauss, J. K. & Zhou, G. (2013), 'International stock return predictability: what is the role of the united states?', *The Journal of Finance* **68**(4), 1633–1662.

Schneeweiss, H. & Mathes, H. (1995), 'Factor analysis and principal components', *Journal of multivariate analysis* **55**(1), 105–124.

Schwert, G. W. (2003), 'Anomalies and market efficiency', *Handbook of the Economics of Finance* **1**, 939–974.

Sharpe, W. F. (1964), 'Capital asset prices: A theory of market equilibrium under conditions of risk', *Journal of Financial Economics* **9**(3), 424–442.

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

Tsay, R. S. (2010), *Analysis of Financial Times Series*, 3 edn, John Wiley Sons, Inc., Hoboken, New Jersey.

Zhong, X. & Enke, D. (2017), 'Forecasting daily stock market return using dimensionality reduction', *Expert Systems with Applications* **67**, 126–139.

9 Appendix

The following tables and figures shows the variables that's been used in this thesis and robustness test.

Appendix 1

Acronym	Definition of characteristic
a2me	Assets-to-market cap
agr	Asset growth
ato	Net sales over lagged net operating assets
beta	Beta
betasq	Beta squared
bm	Book-to-market
c*	Ratio of cash and short-term
cto	Capital turnover
cash	Cash holdings
cashdebt	Cash flow to debt
cashpr	Cash productivity
chscho	Change in shares outstanding
chinv	Change in inventory
chtx	Change in tax expense
cinvest	Corporate investment
currat	Current ratio
d2a	Capital intensity
depr	Depreciation / PP&E
dy	Dividend to price
egr	Growth in common shareholder equity
ep	Earnings to price
hire*	Employee growth rate
gma	Gross profitability
grCAPX	Growth in capital expenditures
invest	Investments
lev	Leverage
lgr	Growth in long-term debt
mom12m	12-month momentum
mom1m	1-month momentum
mom36m	36-month momentum
mom6m	6-month momentum
mve	Natural log of market capitalization at end of month t-1

(continued)

operprof	Operating profitability
pchcurrat	% change in current ratio
pchdepr	% change in depreciation
pchquick	% change in quick ratio
pchsale_pchrect	% change in sales - % change in A/R
pchsale_pchxsga	% change in sales - % change in SG&A
pchsale_pchinvt*	% change in sales - % change in inventory
pchsaleinv*	% change sales-to-inventory
quick	Quick ratio
rd_mve	R&D to market capitalization
rd_sale	R&D to sales
realestate*	Change in real estate
rel.to.high	Closeness to 52-week high is the ratio of stock price
roaq	Return on assets
roeq	Return on equity
salecash	Sales to cash
saleinv*	Sales to inventory
salerec	Sales to receivables
sgr	Sales growth
sp	Sales to price
tang	Debt capacity/firm tangibility
tb	Tax income to book income

* = removed

Appendix 2

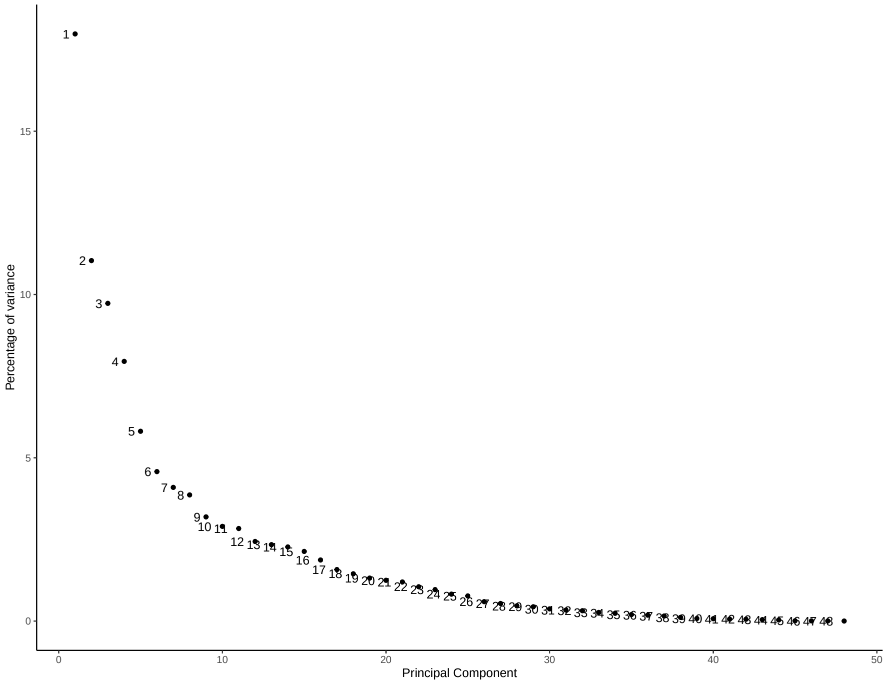


Figure 7: Scree plot of percentage of variance explained for each PC with 100 firms.

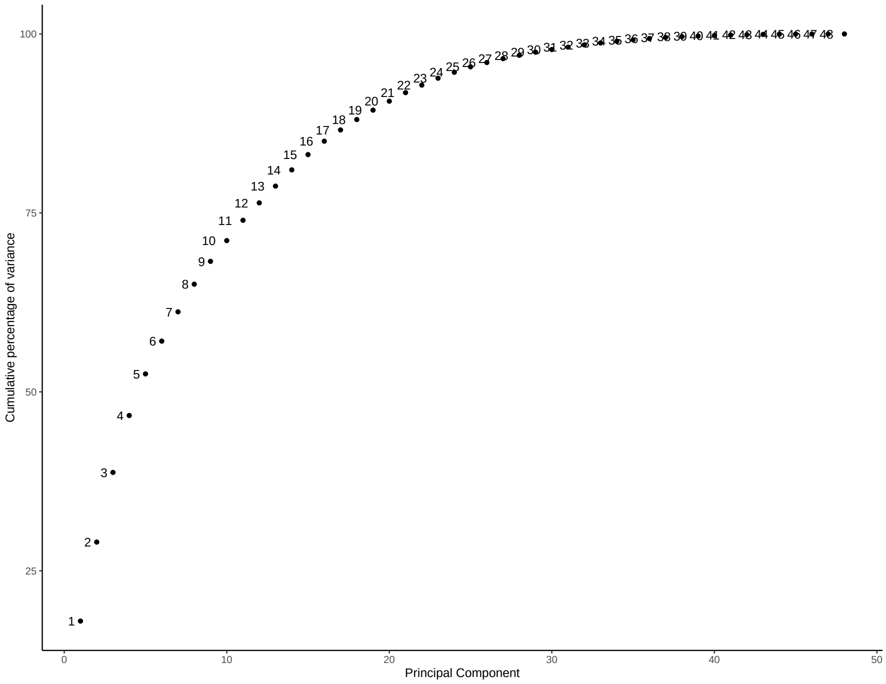


Figure 8: Cumulative percentage of variance explained by each PC with 100 firms.

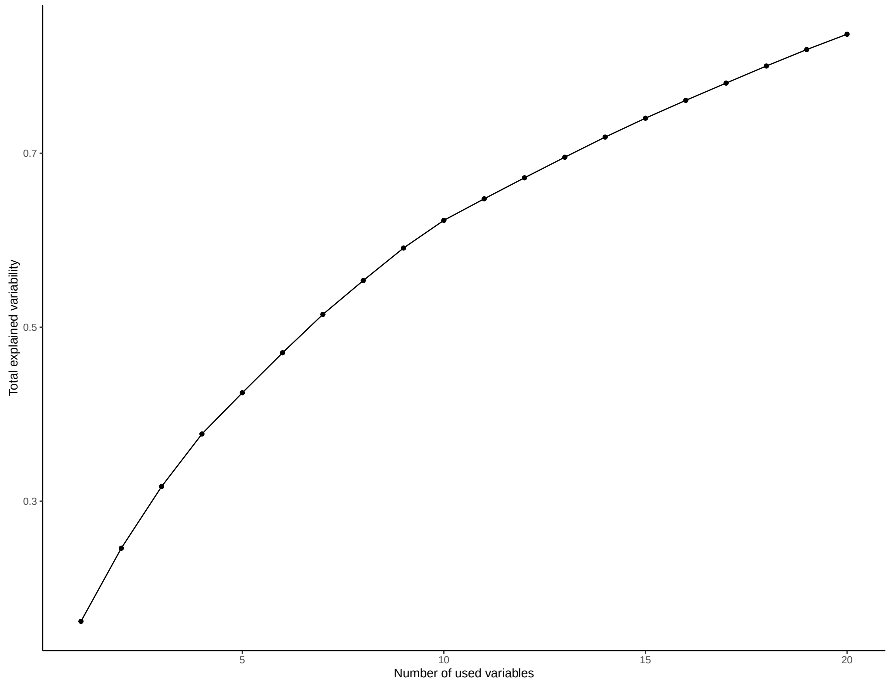


Figure 9: Total explained variability by a number of factors with 100 firms.

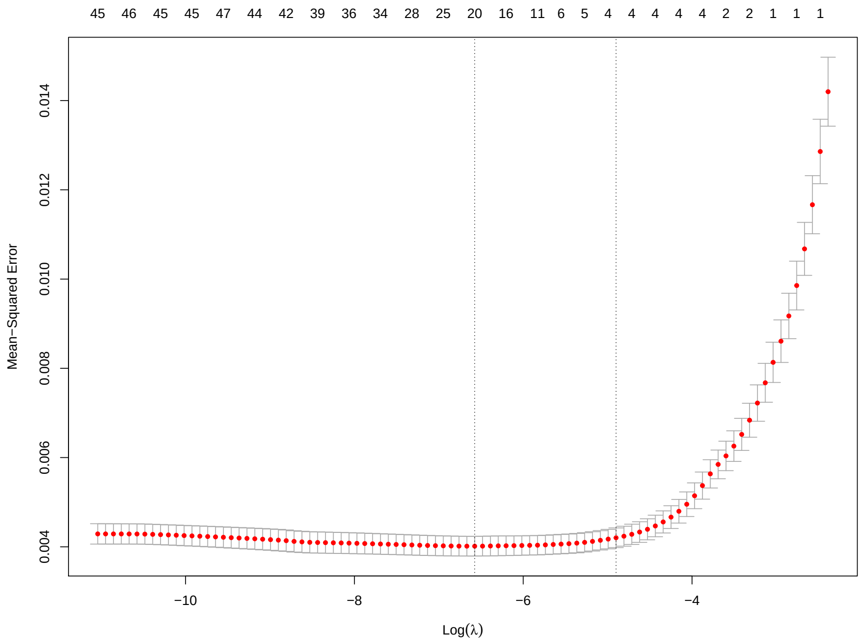


Figure 10: Cross validated λ with corresponding MSE with 100 firms.