Estimating the Market Risk Exposure through a Factor Model with Random Effects

Department of Mathematics, Linköping University

Lukas Börjesson

LiTH-MAT-EX-2022/11–SE

Credits: 30 hp
Level: A
Supervisor: Martin Singull, Department of Mathematics, Linköping University
Examiner: Xiangfeng Yang, Department of Mathematics, Linköping University
Linköping: August 2021
Abstract

In this thesis, we set out to model the market risk exposure for 251 stocks in the S&P 500 index, during a ten-year period between 2011-04-30 and 2021-03-31. The study brings to light a model not often mentioned in the scientific literature focused on market risk estimation, the linear mixed model. The linear mixed model makes it possible to model a time-varying market risk, as well as adding structure to the idiosyncratic risk, which is often assumed to be a stationary process. The results show that the mixed model is able to produce more accurate estimates for the market risk, compared to the baseline, which is here defined as a CAPM model. The success of the mixed model, which we in the study will refer to as the ADAPT model (adaptive APT), most certainly lies in its ability to create a hierarchical regression model. This makes it possible to not just view the set of observations as a single population, but let us group the observations into different clusters and in such a way makes it possible to construct a time-varying exposure. In the last part of the thesis, we highlight possible improvements for future works, which could make the estimation even more accurate and also more efficient.

Keywords:
Market risk exposure, non-stationary market risk exposure, non-stationary idiosyncratic risk, factor model, linear regression, linear mixed model, random effects, mixture of gaussians.

URL for electronic version:
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-187909
Acknowledgements

I would like to thank my family and close friends, who I know are always there through good and bad times, and always comes through with good and valuable advice.

I would also like to take the opportunity to thank my supervisor, Professor Martin Singull who has been an excellent tutor during my studies at Linköping University. I am grateful to have met you and feel that I have found someone who shares the same principles and values and I hope that this is just the beginning of a prosperous relationship.
# Contents

1 Introduction ............................................. 1  
   1.1 Background ........................................... 2  
   1.2 Research Questions .................................... 2  
   1.3 Related Works ......................................... 2  
   1.4 Delimitations .......................................... 3  

2 Theory .................................................... 4  
   2.1 Random Variables ....................................... 4  
      2.1.1 Bernoulli Distribution ............................. 5  
      2.1.2 Normal (Gaussian) Distribution ...................... 5  
      2.1.3 Log-normal Distribution ............................ 6  
      2.1.4 Mixture of Gaussians ............................... 7  
   2.2 Stochastic Processes .................................... 8  
      2.2.1 Random Walk ........................................ 9  
      2.2.2 Brownian Motion .................................... 9  
      2.2.3 Geometric Brownian Motion ......................... 10  
      2.2.4 Other Processes ................................... 11  
   2.3 Regression Analysis ..................................... 11  
      2.3.1 Linear Model ....................................... 11  
      2.3.2 Linear Mixed Model ................................. 14  
   2.4 Model Selection ...................................... 18  
      2.4.1 Coefficient of Determination ....................... 18  
      2.4.2 Significance of Coefficients ....................... 18  
      2.4.3 Information Criteria .............................. 19  
   2.5 Risk Analysis ....................................... 23  
      2.5.1 Risk for a portfolio ............................... 24  
      2.5.2 Market Risk and Idiosyncratic Risk ................ 24  
      2.5.3 Risk vs. Returns .................................. 25  
      2.5.4 Capital Asset Pricing Model ....................... 25  

Börjesson, 2021.
## Method

### 3.1 Data Sampling
- 3.1.1 Assumptions ........................................ 33
- 3.1.2 Response ........................................... 34
- 3.1.3 Fixed Factors ...................................... 34
- 3.1.4 Random Factors .................................... 34
- 3.1.5 Clustering Factors .................................. 34

### 3.2 Models
- 3.2.1 CAPM Model ....................................... 35
- 3.2.2 ADAPT Model ....................................... 35

### 3.3 Training and Model Selection
- 3.3.1 CAPM Model ....................................... 38
- 3.3.2 ADAPT Model ....................................... 38

### 3.4 Backtesting
- 3.4.1 Market Neutral Portfolio ......................... 40
- 3.4.2 Portfolio Comparison .............................. 41

## Result

### 4.1 Model Fitness Comparison .......................... 44

### 4.2 Portfolio Comparison ................................ 47
- 4.2.1 Leverage ........................................... 47
- 4.2.2 Transactions ....................................... 49
- 4.2.3 Performance ........................................ 52

## Discussion and Conclusion ............................. 57

### 5.1 Discussion ........................................... 57
- 5.1.1 Method ............................................. 57
- 5.1.2 Result ............................................... 59

### 5.2 Conclusion ........................................... 60

### 5.3 Future Work .......................................... 60
Chapter 1

Introduction

Recent events, such as the recent financial crisis and the response to the Covid 19 pandemic, have shown just how increasingly dependent the global economy is on the financial markets. However, since the actors on the financial markets are also dependent on the current state of and outlook for the economy, the financial markets are dependent on the economy as well. This thesis will focus on the risk that the economy poses towards a basket of financial instruments, or more simply put; a portfolio. This risk is often denoted the market risk for a specific portfolio, and there are currently a range of different ways that financial modellers can use in order to estimate it. Most methods identifies suitable financial indicators, such as stock indices, and then try to quantify the risk that the indicator, or factor, has on the portfolio. Two popular methods are the capital asset pricing model (CAPM) and arbitrage pricing theory (APT), where regression analysis is used in order to quantify the exposure to different factors. A problem with the mentioned factor models are that they assume that the securities’ exposure to market risk is constant over the time period that the model is estimated for. Yet another problem is that the idiosyncratic risk, the risk not attributable to the market risk, is assumed to be a stochastic variable drawn from a stationary distribution.

This thesis will bring to light a method that is seldom mentioned in the scientific literature for market risk, the linear mixed-effects model. The linear mixed-effects model is a factor model, just like the CAPM and APT model, but it lets the factors vary and provides a way to add structure to the stochastic idiosyncratic risk. In order to estimate the effectiveness of the model, a suitable way would be to compare it with a standard fixed effects factor model, such as the CAPM or APT model.
1.1 Background

The work in this thesis was requested by a hedge fund located in Sweden, which has the goal of achieving risk-adjusted, but market comparable returns. In order to be risk-adjusted, the fund aims to provide returns that are uncorrelated with the market. The fund currently uses an APT model as the main source to quantify the market risk for each security in the portfolio. The fund then uses this information to optimally structure its trades. The estimation of market risk is thus an important part of the fund’s strategy. However, the fund continuously aims to improve their models, and requested a model that better accounts for the temporal structure found in financial time series.

1.2 Research Questions

As was previously alluded to, a mixed linear effects model will be compared to a linear fixed effects model when estimating the exposure a security has to the market risk. In order to accurately compare the two methods, this thesis will aim to answer the following two research questions:

Q1: Can a factor model that incorporates random effects consistently provide better goodness of fit measures, than a model with only fixed effects?

Q2: Can random effects help provide better predictive estimates for securities’ exposure to market risk?

The first question will be answered by comparing the fit between a model with only fixed effects to a model with mixed effects, using a suitable goodness of fit measure. The goodness of fit provides information on how well the model has explained the data in the past, but this gives only an indication of how well the model will explain the exposure going forward. In order to answer the second question, the estimated exposure to market risks will be used to construct beta-neutral (market risk neutral) portfolios, for the different models that were used to answer the first question. We will then also take a quick look at how well these portfolios perform in terms of return and volatility, compared to the market.

1.3 Related Works

One of the most popular factor models used in the financial industry is the CAPM proposed by Sharpe in 1964 [23]. However, the model makes critical assumptions built upon the framework of an efficient market. These assumptions
1.4 Delimitations

are openly criticized by Fama and French in [7] who modeled a three factor model based on the APT model proposed by Ross in [20]. The CAPM model and APT model have since been generalized to include random effects in order to estimate the specific effects posed by different industries and sectors [21]. The mixed model has also been used to model credit default risks [15]. However, we have not found any relevant papers that use a hierarchical structure, where different time periods are defined as clusters in the model, in order to estimate a dynamic market risk.

1.4 Delimitations

We have limited our work to only focus on the stock market, since this is the market where the fund primarily operates in. Furthermore, we have restricted our selection to only large blue-chip stocks, since again the fund primarily only takes positions in companies with a market capitalization exceeding 1 billion dollars.
Chapter 2

Theory

2.1 Random Variables

A random variable is a variable with outcomes that depends on randomness and since this randomness can take on different forms, we have many different types of stochastic variables. Furthermore, the randomness for a variable might be correlated or dependent on one or more variables, meaning that there is a relationship between the variables. This relationship might extend both ways, or one variable might have a causal effect on another variable, i.e., given an outcome from the former variable, the distribution of the outcomes for the latter will change. The resulting distribution is often called the conditional distribution and the probabilities for each outcome from this distribution is then called conditional probabilities. This concept of causality is an important one, both in this thesis and in real life, since we often try to model the possibility of future events, given prior knowledge of past events.

Furthermore, all distributions can be characterized by a set of parameters, among them the distribution’s moments, where the expected value\(^1\) (first moment) and variance\(^2\) (second central moment) probably are the two parameters that most often are mentioned. However, analysis of different distributions and scenarios calls for different characterizations and in this thesis we will take into consideration the skew\(^3\) (third standardized moment) and the kurtosis\(^4\) (fourth

---

\(^1\) A measure of the most probable outcome from the distribution.
\(^2\) A measure of the expected variation of the distribution around the expected value. The square root of the variance, the standard deviation, is often mentioned as well.
\(^3\) A measure of the asymmetry of the distribution around its expected value.
\(^4\) A measure of the "tailedness" of the distribution, i.e., it determines the probability of extreme deviations around the the expected value.

Börjesson, 2021.
standardized moment) as well. Although, the measure of kurtosis is often re-
placed by excess kurtosis, which is equal to the kurtosis minus three and gives 
a measure of the kurtosis relative to the normal distribution.

This was just a short introduction to the concept of random variables and 
the reader should be comfortable with these concepts to follow the reasoning in 
this chapter. The following four sections define distributions that will be used 
to derive the final distribution of the process, which will be used to explain the 
movement of asset prices on the market.

2.1.1 Bernoulli Distribution

The outcomes from a stochastic variable with Bernoulli distribution are binary, 
meaning that they only take on two different values, one or zero, with probability 
$p$ and $1 - p$ respectively. This gives us the following probability mass function 
for each $k$ outcome

$$p(k, p) = \begin{cases} 1 - p, & k = 0, \\ p, & k = 1, \end{cases}$$

and the expected value and variance for this distribution is $p$ and $p(1 - p)$, 
respectively. No further information for the higher moments are needed for the 
Bernoulli distribution in this thesis.

2.1.2 Normal (Gaussian) Distribution

The normal, or Gaussian, distribution is probably the most well-known distri-
bution. The reason behind this is most likely that many natural phenomena 
can be explained by its distribution and the fact that the sum of independent 
stochastic variables tends to be normally distributed. The standard normal dis-
bution is the simplest form, where the expected value and variance are equal 
to zero and one respectively. The probability density function for a standard 
normal variable is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty,$$

and Figure 2.1.a shows the curve for its density. The variance for the standard 
normal distribution is equal to one, while its expected value, skew and excess 
kurtosis are equal to zero.

It can be proved that every linear combination of standard normals are 
themselves normally distributed, which gives us the following definition

$$Y = \mu + \sigma X, \quad \text{with} \quad X \sim N(0, 1) \iff Y \sim N(\mu, \sigma^2), \quad (2.1)$$
where $X$ is a standard normal variable and the linear combination $Y$ is a normal variable and has the following density function

$$f(y) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}. \tag{2.2}$$

The mean and variance are equal to $\mu$ and $\sigma^2$, respectively, while the skew and excess kurtosis are still equal to zero.

The normal distribution can be further generalized to handle multivariate distributions

$$f(y|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)}, \tag{2.3}$$

where $y$ is a vector of random normal variables, $y_i$, and $\mu$ is a vector containing the mean value for each random variable $y_i$. The matrix $\Sigma$ is a covariance matrix matrix, that defines the covariance between each variable set $(y_i, y_j)$ in the random vector $y$.

### 2.1.3 Log-normal Distribution

A log-norm distributed variable is a variable whose natural logarithm is normal distributed, i.e.,

$$Y \sim \text{lognormal}(\mu, \sigma^2) \iff \ln(Y) \sim N(\mu, \sigma^2), \tag{2.4}$$

and where $Y$ has the following probability density function

$$f(y) = \frac{1}{\sqrt{2\pi \sigma^2 x^2}} e^{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}}, \quad x > 0,$$

The expected value and variance for the log-normal distribution are equal to $e^{\mu + \frac{\sigma^2}{2}}$ and $(e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$, respectively. However, more interestingly is that the skewness and excess kurtosis are no longer equal to zero, but instead equal to $(e^{\sigma^2} + 2)\sqrt{e^{\sigma^2} - 1}$ and $e^{4\sigma^2} + 2e^{3\sigma^2} + 3e^{2\sigma^2} - 6$, respectively. This means that as long as the variance is greater than zero, the distribution has a positive skewness and excess kurtosis (i.e., fatter right tail, and also higher density for extreme values), which can be observed in Figure 2.1 (b). The fact that log-normal distribution has a positive skew and excess kurtosis makes it a more suitable distribution for asset returns, since these these conditions are often observed empirically (see Sections 2.2.2 and 2.2.3 for a continued discussion regarding normal and log-normal returns). A visualization of the positive skew and excess kurtosis can also be observed in Figure 2.1b.
2.1. Random Variables

2.1.4 Mixture of Gaussians

A Gaussian mixture is a distribution that takes two or more Gaussian distributions and then combines them into one single distribution. The probability density function for the Gaussian mixture is defined as

\[ f(y) = \sum_{k=1}^{K} \pi_k f(y|\mu_k, \Sigma_k), \]  

(2.5)

where each \( k \)th component \( f(y|\mu_k, \Sigma_k) \) is the probability density function for a multivariate normally distributed random vector as defined in (2.3). The variable \( \pi_k \) in (2.5) provides the weight for the \( k \)th component and can be viewed as the prior probability that an observation is drawn from that distribution. Since it is a probability it fulfills the condition

\[ 0 \leq \pi_k \leq 1, \quad k = 1, \ldots, K \]

and for the mixture model to make sense all \( \pi_k \) values need to sum up to one

\[ \sum_{k=1}^{K} \pi_k = 1. \]  

(2.6)
Chapter 2. Theory

Figure 2.2: A mixture of three Gaussians with equal prior probability weights (i.e., $\pi_1 = \pi_2 = \pi_3$).

Figure 2.2 displays a Gaussian mixture with three Gaussian components, and in theory one can approximate all smooth densities, with any non specific nonzero amount of error, given a Gaussian mixture model with enough components [8].

2.2 Stochastic Processes

A stochastic process is a collection of stochastic variables, where each variable is indexed according to an index set. A common notation is $\{X_i\}_{i \in I}$, where each random variable $X_i$ is indexed according to the index set $I$. The index set is often chosen to be a subset of the natural numbers, since it becomes easy to interpret the subset as a timeline and thereby to interpret the collection of variables as a sequence with respect to time, or more commonly called a *time series*. All stochastic processes in this thesis will be indexed with respect to time and the following notation will be used; $\{X_t : t \in T\}$, where $T$ is a subset of the natural numbers.
2.2.1 Random Walk

The random walk is a stochastic process and it can generally be defined by a modified Bernoulli process, \( \{X_t : t \geq 0\} \), where

\[
P(X_t = 1) = p, \\
P(X_t = -1) = 1 - p,
\]

and

\[
E(X_t) = 2p - 1, \\
Var(X_t) = 4p(1 - p).
\]

The outcome from the Bernoulli trial, at each time \( t \), determines if the walker moves to a higher state, with probability \( p \), or to a lower state, with probability \( 1 - p \). The position of the random walker, at time \( t \), can then be defined by the stochastic process \( \{Y_t : t \geq 0\} \) as

\[
Y_t = \Delta x(X_1 + X_2 + \ldots + X_n),
\]

where \( \Delta x \) is the jump distance and \( t = n\Delta t \), with \( t \) being the time period between jumps.

By using the expected value and variance of each Bernoulli trial, and the assumption that each trial are identical and independent, the expected value and variance of the position are

\[
E(Y_t) = (2p - 1)\Delta x n = (2p - 1)\frac{\Delta x}{\Delta t} t = \mu, \\
Var(Y_t) = 4p(1 - p)(\Delta x)^2 n = 4p(1 - p)\frac{\Delta x^2}{\Delta t} t = \sigma^2.
\]

By using the central limit theorem and by letting \( \Delta t \) and \( \Delta x \) go to zero, the distribution of the stochastic process \( Y_t \) can be shown to be that of a Gaussian (Normal) Process and can be seen as a continuous time random walk

\[
p(Y, t) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(Y - \mu)^2}{2\sigma^2}}, \quad (2.7)
\]

see [18] for further details.

2.2.2 Brownian Motion

Brownian motion is named after the Scottish botanist Robert Brown, who, in 1827, observed an erratic movement in pollen particles immersed in water. This
erratic movement were later studied by Albert Einstein, and in 1905, he published a paper, showing that the pollen particles collided with water molecules and thereby proving the existence of atoms \cite{6}. Apart from being a mayor breakthrough in natural sciences, the observed motion provides a statistical framework to model a path and an outcome for an object, determined by unpredictable events. These unpredictable events can be water molecules bumping into grains, determining its path in the fluid, as was observed by Robert Brown or possibly an unpredictable collection of buy/sell decisions for a specific stock, determining the change in its price.

The Brownian motion was in fact studied in financial markets 5 years before Einstein’s discovery, by the French mathematician Louis Bachelier \cite{1}, who derived the Gaussian distribution from the price fluctuations of financial contracts. However, since financial mathematics was not as highly regarded as theoretical physics, it would take decades before Bachelier’s findings got acknowledged by the scientific community.

In \cite{1}, Bachelier proposed that the asset price $S(t)$ was a Brownian motion with a drift, much like the continuous random walk previously derived in Section 2.2.1. The variable Bachelier assumed to be the Brownian motion was the price fluctuations and the differential $dS = S(t + dt) - s(t)$ was thereby considered normally distributed.

Since the Brownian motion can be derived from the random walk explained above, and has the property of a Markov process (i.e., possibilities of future states only depend on the current state). Bachelier’s findings suggests there is no way to beat the market, laying the groundwork for what today is referred to as efficient market hypothesis (EMH). However, there are a few problems with the model Bachelier suggested, where the most alarming one is that the asset price is not restricted from taking on negative values, something that is very unrealistic in a real market. Although, this problem is rectified in the geometric Brownian motion (GBM), which is explained in the next section. Another problem is that the model assumes that the variance is homoscedastic, which is often, for most assets and time periods, not the case.

2.2.3 Geometric Brownian Motion

To remedy the problem with negative values, Paul A. Samuelson suggested, in 1965, that it was the not the absolute change, $dS$, that was the variable to study, but instead the return $dS/S(t)$ \cite{22}. From this he concluded that it was the logarithm of the asset price that was a Brownian motion and that instead the change, $d\ln(S) = \ln(S(t + dt)) - \ln(S(t))$, is normally distributed. Apart from not having to deal with negative values for the asset’s price, it also makes more intuitive sense to model the return instead of the absolute change.
2.3. Regression Analysis

The fact that the return is log-normally distributed is one of the key assumptions made in the famous Black-Scholes model, which provides a formula for determining the prices of European put and call options \[3\]. Although the assumptions in the model is not all empirically valid, it provides researchers and traders, among others, with a good approximation and a general framework upon which to build more advanced models from.

2.2.4 Other Processes

As was mentioned in Section 2.1.3 logarithmic returns implies that the returns have positive skew and kurtosis, which is a desirable feature, since this is often observed on the financial markets. However, the problem with non-stationary variance remains, which makes the estimation of the distribution quite hard.

Section 2.3 will provide concepts for how to estimate the distribution of a stochastic variable, by not only looking at the observations for the variable but to model the outcomes based on outcomes from suitable predictors. Section 2.3.2 will then generalize this method, by also letting some of the predictors be random variables, which provides us with a framework to in use in order to handle the non-stationarity of the financial processes. The next section after that 2.5 will integrate these concepts with the financial market and together will provide a framework for the the models in this thesis.

2.3 Regression Analysis

Regression is a way to estimate the relationship between one or more dependent variables, often called the responses, and one or more independent variables, often called the covariates. This relationship can be linear as well as non-linear, but only the linear relationships will be taken into consideration in this thesis. Furthermore, only models with a single response variable will be considered and the following section will reflect this, i.e., no theory will be provided in order to generalize the models to handle multiple responses.

2.3.1 Linear Model

In order to quantify the relations between the responses and covariates, different approaches can be taken. One of the most used approaches is the ordinary least squares method, where the coefficients are set to minimize the sum of squared differences between the actual values for the response variable and the model’s predictions. These differences are called residuals and in regression analysis, we
further assume that these residuals are independent and identically distributed\(^5\) with a mean centered at zero. Below is a summary of the conditions, and the resulting distribution

\[
y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i = x_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n,
\]

where

\[
\epsilon_i \sim N(0, \sigma^2) \quad \implies \quad y_i \sim N(x_i^T \beta, \sigma^2),
\]

which in matrix form can be written

\[
y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} = X\beta + \epsilon, \quad (2.8)
\]

where

\[
\epsilon \sim N_n(0, \sigma^2 I) \quad \implies \quad y \sim N_n(X\beta, \sigma^2 I). \quad (2.9)
\]

The \(\beta\) coefficients is the quantified relations between the dependent variable \(y_i\) and the independent variables \(x_i\). \(\epsilon_i\) are the errors in the model which we assume are normally distributed, and they represent the residuals when the actual value of \(y_i\) is compared to the model’s prediction of \(y_i\). An important note is that we do not need to make assumptions regarding the distribution of the covariates \(X\), as they are considered known constants in the model. Nevertheless, the data for the covariates should be studied to some degree, in order to not be caught off guard by possible influential points\(^6\) which could have a great influence on the value of the coefficients.

**Parameter Estimation**

As previously mentioned, one of the most used approaches to estimate the parameters in linear regression is ordinary least squares. This estimation process sets out to find values for the parameters that minimize the sum of the squared residuals

\[
\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - x_i^T \hat{\beta})^2,
\]

- \(^5\)Unless stated otherwise, the residuals are here assumed to be normally distributed.
- \(^6\)Points that might be considered outliers in the joint distribution for the covariates and that have great impact on the value for the coefficients.
where \( r_i \) the residuals and \( \hat{y}_i = x_i^T \hat{\beta} \) is the models prediction for the true response \( y_i \).

Given that the error terms, \( y_i - \hat{y}_i \), are uncorrelated, have the same variance and with expected value equal to zero, as well as being normally distributed. The Gauss-Markov theorem states that the \( \beta_i \) coefficients estimated by the ordinary least squares method are unbiased and that they have the lowest sampling variance among all linear estimators. Which is the most likely reason to why ordinary least squares is so often used.

Under the assumption of normally distributed residuals, the log-likelihood function becomes

\[
L(\beta, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)^T(y - X\beta) \\
\propto -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)^T(y - X\beta)
\]

and by maximizing this function, with respect to \( \beta \), we come up with the following estimation for the coefficients

\[
\hat{\beta} = (X^T X)^{-1} X^T y, \tag{2.11}
\]

which means that the estimate for the coefficients that minimizes the sum of squared residuals, is a closed form function of the responses and the covariates. We here assume that the matrix \( X \) has full column rank, which implies that \( X^T X \) is invertible, i.e., non-singular. Furthermore, since the covariates are known in our model and the responses are assumed to be independent normally distributed, with variance \( \sigma^2 \), we get the following distribution for the mean estimators

\[
\hat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1}). \tag{2.12}
\]

If we instead decide to maximize the log-likelihood with respect to the variance \( \sigma^2 \), we arrive at the following estimate

\[
\hat{\sigma}^2 = \frac{(y - X\beta)^T(y - X\beta)}{n}. \tag{2.13}
\]

Since we are trying to reveal information about a population using only a sample set, the spread of the residuals will underestimate the variance of the errors, \( \text{var}(\epsilon) = \hat{\sigma}^2 \), in the regression model (i.e., the error variance is biased). In order to get an unbiased estimate for the variance, we need to account for the degrees of freedom used to estimate the coefficients, \( \beta \), which we use in when estimating
the variance. Given \( p \) independent variables, the unbiased sample variance is equal to

\[
s^2 = \frac{n}{n-p} \hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)}{n-p}.
\]

### 2.3.2 Linear Mixed Model

As was mentioned in the previous section, the coefficients that estimate the relationship between the response and covariates are, in standard regression analysis, considered to be point estimates, i.e., a deterministic best guess determined by the data provided. That is, the parameters are kept fixed for all observations in the model and the standard regression model is therefore sometimes referred to as a fixed effects model. Fixed effects models further assume that the residuals are identically distributed, but for many time series, and especially financial time series, the variance is not homoskedastic. In effect, the fixed points estimate in the standard regression model tries to catch these changes in variance as well, which might lead to less optimal estimates. An approach to handle this problem is to add structure to the variance in the model and thereby also letting it vary for different groups of observations. Linear mixed models does this by adding random effects \( \mathbf{u} \) to the model

\[
\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{\epsilon},
\]  

(2.14)

where

\[
\mathbf{\epsilon} \sim N_n(0, \sigma^2 \mathbf{R}),
\]

and

\[
\mathbf{u} \sim N_{nk}(0, \sigma^2 \mathbf{G}), \quad \sum_k n_k = n,
\]

where the design matrix \( \mathbf{Z} \) and covariance matrix \( \mathbf{G} \) adds the structure to the variance. We further assume that the random effects \( \mathbf{u} \) and the residuals \( \mathbf{\epsilon} \) are independent

\[
\text{Var} \left( \begin{pmatrix} \mathbf{u} \\ \mathbf{\epsilon} \end{pmatrix} \right) = \sigma^2 \begin{pmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{R} \end{pmatrix},
\]

and this implies that

\[
\mathbf{y} \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{V}),
\]

(2.15)

\footnote{More generally, \( p \) is equal to the rank of the matrix for the covariates; \( \text{rank}(\mathbf{X}) \).}
where

\[ V = ZGZ^T + R. \]

Since we again assume that the residuals are normally distributed and now also that the random effects are normally distributed, we consequently assume that the response are normally distributed \[19\].

### Parameter Estimation

The density function for the response \( y \) in the linear mixed model (2.14) is given as

\[
 f(y) = (2\pi\sigma^2)^{-\frac{n}{2}} |V|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) \right\}. \tag{2.16}
\]

With \( \theta \) representing the parameters in the covariance matrix \( V \) in (2.15), we can define the log-likelihood as

\[
 \mathcal{L}(\beta, \sigma^2, \theta; y) \propto -\frac{n}{2} \ln \sigma^2 - \frac{1}{2} \ln |V| - \frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta), \tag{2.17}
\]

If we assume that the parameters in the vector \( \theta \) are known, the fixed \( \beta \) effects can be estimated by maximizing the log-likelihood with respect to \( \beta \)

\[
 \hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y. \tag{2.18}
\]

However, if we aim to predict the random effects as well, we need to consider the joint likelihood for \( y \) and \( u \), where the joint density for \( y \) and \( u \) is defined as

\[
 f(y, u) = (2\pi\sigma^2)^{-\frac{n}{2} - \frac{q}{2}} \left( \det \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} \right)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \begin{pmatrix} u \\ y - X\beta - Zu \end{pmatrix}^T \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix}^{-1} \begin{pmatrix} u \\ y - X\beta - Zu \end{pmatrix} \right\}.
\]

Given Bayes’ theorem, it could also be rewritten as

\[
 f(y, u) = f(y|u)f(u),
\]

where

\[
 f(y|u) = (2\pi\sigma^2)^{-\frac{n}{2}} |R|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta - Zu)^T R^{-1} (y - X\beta - Zu) \right\},
\]

\[
 f(u) = (2\pi\sigma^2)^{-\frac{q}{2} - \frac{1}{2}} |R|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2} u^T R^{-1} u \right\}.
\]

...
and
\[
    f(u) = (2\pi\sigma^2)^{-\frac{d}{2}} |G|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2} u^T G^{-1} u \right\}.
\]

Given the log-likelihood for \( f(y|u) \)
\[
    \mathcal{L}(\beta, \theta; y|u) \propto -\frac{n}{2} \ln \sigma^2 - \frac{1}{2} \ln |R| - \frac{1}{2\sigma^2} (y - X\beta - Zu)^T R^{-1} (y - X\beta - Zu)
\]
and for \( f(u) \)
\[
    \mathcal{L}(\beta, \theta; u) \propto -\frac{n}{2} \ln \sigma^2 - \frac{1}{2} \ln |G| - \frac{1}{2\sigma^2} u^T G^{-1} u,
\]
the joint log-likelihood can be written as
\[
    \mathcal{L}(\beta, \theta; y, u) = \mathcal{L}(\beta, \theta; y|u) + \mathcal{L}(\beta, \theta; u).
\] (2.19)

By maximizing the joint log-likelihood with respect to \( \beta \) and \( u \) gives Henderson’s mixed model equations
\[
    X^T R^{-1} X\hat{\beta} + X^T R^{-1} Z\hat{u} = X^T R^{-1} y
\]
\[
    X^T R^{-1} X\hat{\beta} + (Z^T R^{-1} Z + G^{-1}) \hat{u} = Z^T R^{-1} y,
\] (2.20)
with the following best linear unbiased predictors (BLUP)\(^8\)
\[
    \hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y,
\]
\[
    \hat{u} = GZ^T V^{-1} (y - X\beta),
\] (2.21)
where we still consider the parameters \( \theta \) in \( V \), and thereby also \( G \), as known.

Provided that the matrix \( X \) is of full rank the covariance matrix for the estimation errors is
\[
    E \left\{ \begin{pmatrix} \hat{\beta} - \beta \\ \hat{u} - u \end{pmatrix} \begin{pmatrix} \hat{\beta} - \beta \\ \hat{u} - u \end{pmatrix}^T \right\} = \begin{pmatrix} X^T R^{-1} X & X^T R^{-1} z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix}^{-1},
\]
and the variance for the estimated \( \hat{\beta} \) and \( \hat{u} \)
\[
    \text{Var}(\hat{\beta}) = (X^T V^{-1} X)^{-1},
\]
\[
    \text{Var}(\hat{u}) = GZ^T \left( V^{-1} - V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} \right) ZG.
\]

---

\(^8\) Best in the sense that the mean squared error produced by these estimators is the minimum amongst all linear unbiased estimators [19].
The derivation of the BLUPs showed here is just one of many. The one here is based on Henderson’s justification, which is one of four popular derivations that Robinson pedagogically derives in [19].

As mentioned before, the previous derivation is based on the assumption that the covariance matrix $V$ is known in advance, which is often not the case. This complicates the estimation process, since we then must estimate the parameters in $\theta$ as well. This is done by an iterative process, where the effects $\beta$ and $u$ in (2.21) are updated based on the previous values of $\theta$ and where $\theta$ is then estimated from the updated values of the effects. If we assume; $y$ and $\hat{e} = y - X\hat{\beta} - Z\hat{u}$ are observed, $R = I$ and that $G$ is an arbitrary non-negative definite matrix, then $\sigma^2$ and the parameters in $\theta$ can be estimated as

$$
\hat{\sigma}^2 = \frac{\hat{e}^T \hat{e} + \text{tr} \text{Var}(\hat{e})}{n} \tag{2.22}
$$

and

$$
\hat{G} = \frac{\hat{u} \hat{u}^T + \text{Var}(\hat{u})}{m}, \tag{2.23}
$$

where $n = \sum_{i=1}^{m} n_i$ is the number of observations and $m$ the number of groups [10]. One iterative algorithm that can be used is the expectation-maximization algorithm, which is outlined, for other purposes, in Section 2.7. However, there is a range of different optimization techniques that can be used, such as penalized least squares (used in the package lme4 for R), Newton-Raphson, Broyden–Fletcher–Goldfarb–Shanno, conjugate gradient methods, and so on. Lindstrom and Bates explain both the Newton-Raphson and EM approaches in [12], which is a productive read if one has the time and interest.

The theory above takes the maximum likelihood approach, which does not take into account the loss in degrees of freedom when calculating the fixed effects $\beta$. This generally leads to biased estimates for the parameters $\theta$, which is similar to the problem of biased variance in the linear model. In order to receive an unbiased estimate for $\theta$, again similar to the unbiased variance, a restricted likelihood should be considered. In [17], Patterson and Thompson split the data into two parts, each with separate likelihoods $\mathcal{L}'$ and $\mathcal{L}''$, where

$$
\mathcal{L} = \mathcal{L}' + \mathcal{L}''.
$$

In $\mathcal{L}'$ the $\beta$ coefficients has been integrated away and Patterson and Thompson show that without the presence of $\beta$ in the likelihood $\mathcal{L}'$, unbiased estimates of

\[\text{Sometimes also denoted the residual or reduced maximum likelihood}\]
the variance parameters \((\sigma^2, \theta)\) can be derived. Meanwhile, \(\beta\) can in the next step be derived from \(\mathcal{L}''\).

Laird et al. [10] show that, using an empirical Bayes approach, the estimates \(2.22\) and \(2.23\) can be derived in a REML setting as well, using the EM algorithm.

### 2.4 Model Selection

#### 2.4.1 Coefficient of Determination

When comparing different models, such as linear models with different sets of covariates, we need a goodness of fit measure, that can indicate to us how accurate the model seems to be at predicting the responses. One approach is to compare the residual sum of squares, \(SS_E\), with the total sum of squares, \(SS_T\), for the model, in order to estimate how much of the variance for the response variables that can be explained by the model. The coefficient of determination, often denoted \(R^2\), does just that

\[
R^2 = 1 - \frac{SS_E}{SS_T} = 1 - \frac{(y - X\beta)^T(y - X\beta)}{(y - \bar{y})^T(y - \bar{y})}
\]

and its value ranges between zero and one, where a value closer to one point to a model that does a good job of explaining the variance for the response variable. However, a model with more covariates results in a higher \(R^2\) value but with it comes the risk of overfitting the model to the data, i.e., that it might not generalize to out-of-sample data points. The adjusted \(R^2\)

\[
R^2_{adj} = 1 - (1 - R^2) \frac{n - 1}{n - k - 1} \tag{2.24}
\]

puts a penalty on the number of covariates, which means that a model selection based on this adjusted value might increase the generalization of the final model.

#### 2.4.2 Significance of Coefficients

While the \(R^2\) and \(R^2_{adj}\) measures might be easy to interpret and provides a way to compare different models to each other, it does not explicitly say anything about the coefficients or more importantly the relevance of the included covariates. Given \(2.11\) we get \(2.12\) and we thereby have an uncertainty estimate for the coefficients. This makes it possible to measure the significance of the coefficients and thereby the relevance of the covariates. In order to measure the significance of the coefficients, one can set up a hypothesis; that the coefficient
is equal to zero, and then determine if this assumption is statistically valid. The \textit{p-value} provides us with a simple way to test this by measuring the probability of observing a certain value (or more extreme), for the variable under consideration, given a hypothesis

\[ p = P(T \geq t \mid H_0), \]

where \( H_0 \) is the given hypothesis, \( T \) is the random variable, and \( t \) the observed value of the statistic for the distribution of \( T \). When analyzing the coefficients, a low p-value means that there is a low probability of observing the given value for the coefficient being analyzed, given the assumption that its expected value equals zero, i.e.,

\[ \hat{\beta}_j \sim N(0, \text{Var}[\hat{\beta}_j]), \quad (2.25) \]

As an example, say our estimate for coefficient \( j \) is equal to \( b_j \) in a regression model as defined in (2.8). Then the p-value for coefficient \( j \) is equal to the probability of drawing a value, from the random variable in (2.25), with an absolute value greater than the absolute value of \( b_j \)

\[ p = P(|\hat{\beta}_j| \geq |b_j| \mid H_0) = P(|\hat{\beta}_j| \geq |b_j| \mid \hat{\beta}_j \sim N(0, \text{Var}[\hat{\beta}_j])), \]

With this in mind, one usually sets out to maximize a model selection measure, but constrained by not letting the p-value for the coefficients to be greater than a specific \textit{significance level}, which is often set to equal 0.01, 0.05 or 0.1. However, Douglas Bates\textsuperscript{10} argues that inference using p-values in mixed models is not as straightforward as in the case for linear models with only fixed effects. Bates has therefore decided to not provide calculations for the p-values in lmer and in a response to the public\textsuperscript{11} he gives he’s key reasons behind omitting the p-values.

\subsection*{2.4.3 Information Criteria}

The \textit{Akaike Information Criteria} (AIC) and \textit{Bayesian Information Criteria} (BIC) are two other measures of fitness, which differ from the \( R^2 \) measure. While \( R^2 \) focuses on the explained variance, AIC and BIC instead focus on the log-likelihood value

\[ AIC = 2k - 2 \ln \mathcal{L}(\theta), \]

\[ BIC = k \ln n - 2 \ln \mathcal{L}(\theta), \quad (2.26) \]

\textsuperscript{10}The developer of the lmer package in R, one of the most used packages for mixed models.

\textsuperscript{11}https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html
where $k$ is the number of parameters in $\theta$ and $n$ the number of observations, and both are used as penalization terms in order to restrict the bias towards complexity. By looking at the log-likelihood functions (2.10) and (2.17), we can see that such a model selection not only compares the residuals, but also the uncertainty for the parameters in the model. However, the log-likelihood methods provide no relevant information regarding the absolute quality of the model, which means that if one was to evaluate a standalone model, AIC and BIC would not be of much help and a measure such as $R^2_{adj}$ would be more useful. Although, when comparing the relative quality of two or more models, AIC and BIC might be better measures than $R^2_{adj}$, since $R^2_{adj}$ measures the quality based on the observed residuals, while the quality of the log-likelihood methods is based on the observed residuals and the uncertainty of the parameters in the model. This is especially important when using random effects in the model, since if the random effects are considered unobserved, the explained variance is going to be lower for the linear fixed model, while it will be greater if the random effects are considered observed. As a consequence, an approach that takes into account more information from the likelihood, such as the information criteria, is more suitable for the mixed models than measures that only consider the explained variance, such as $R^2$.

By comparing AIC and BIC in (2.26) we can see that BIC puts a greater penalty on the model complexity than AIC does, which means that using BIC as a model selection tool would generally select less complex models than what the AIC would. A penalty for complexity is a good thing, since it decreases the risk of overfitting. However, since our aim in this thesis is to compare a less complex linear fixed model with more complex linear mixed models, a too high focus on the penalty for complexity could make the models to similar. We deem that the consequence of having too similar model has a more negative impact on the analysis in this thesis than what a somewhat (possibly) overfitted model would have, which is why AIC is preferred in this thesis.

The AIC, as defined in (2.26), is sufficient when analysis models with only fixed effects, but becomes insufficient in most cases when analysing mixed effects. In order for the AIC to be better suited for mixed models, the log-likelihood, $\mathcal{L}$, needs to be generalized to incorporate different types of likelihoods. In this thesis, we need to consider and define two other likelihoods in addition to the marginal likelihood. The first one is the restricted likelihood, for which we already have explained the concept of in Section 2.3.2. Derived from [17], but presented in a different fashion, the value for the restricted likelihood is

$$
\mathcal{L}_{REML}(\beta, \sigma^2, \theta; y) = \mathcal{L}_{ML}(\beta, \sigma^2, \theta; y) - \frac{1}{2} \ln |\sigma^{-2}X^TV^{-1}X|,
$$

(2.27)

where a penalty term, $\frac{1}{2} \ln |\sigma^{-2}X^TV^{-1}X|$, has been added compared to the
marginal likelihood in (2.17). Again, \( \theta \) is a placeholder for the parameters in \( V \) and \( u \) is the vector of random effects. The second one is the conditional log-likelihood

\[
\mathcal{L}(\beta, \sigma^2, \theta; y|u) \propto -\frac{n}{2} \ln \sigma^2 - \frac{1}{2} \ln |V| - \frac{1}{2\sigma^2} (y - X\beta - Zu)^T V^{-1} (y - X\beta - Zu),
\]

(2.28)

which we need in order to calculate the conditional AIC.

### Marginal AIC

The marginal AIC (mAIC) uses the marginal log-likelihood and for the linear fixed and mixed models, the criterion can be defined as

\[
mAIC = 2a_n(p + 1) - 2\mathcal{L}(\beta, \sigma^2; y),
\]

(2.29)

\[
mAIC = 2a_n(p + q + 1) - 2\mathcal{L}(\beta, \sigma^2; \theta; y),
\]

(2.30)

where \( \mathcal{L}(\beta, \sigma^2) \) and \( \mathcal{L}(\beta, \sigma^2, \theta; y) \) is the likelihoods in (2.10) and (2.17). Here, \( p \) is the number of parameters in \( \beta \), \( q \) is the number of parameters in \( V \) and the 1 accounts for the variance parameter \( \sigma^2 \). The complexity added by the variance parameter is not always accounted for and only \( p \) or \( p + q \) is used in the cases when it is not. In the infinite sample (asymptotic) case \( a_n = 1 \), while in the finite sample case \( a_n = n/(n - p - q - 1) \).

### Conditional AIC

The conditional AIC is given by

\[
cAIC = a_n(\beta, \sigma^2, \theta) - 2\mathcal{L}(\beta, \sigma^2, \theta; y|u),
\]

(2.31)

where \( \mathcal{L}(\beta, \sigma^2, \theta; y|u) \) is the conditional log-likelihood in (2.28) and \( a_n(\beta, \sigma^2, \theta) \) varies depending on the implementation. Müller et al. [16] gives a good summary of different choices for \( a_n(\beta, \sigma^2, \theta) \), where the choice depends primarily on which parameters are assumed to be known and the relative weight that each category of parameters has. One such choice for \( a_n(\beta, \sigma^2, \theta) \) is

\[
a_n(\theta) = \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i} = \text{tr} \left( \frac{\partial \hat{y}}{\partial y} \right),
\]

(2.32)

proposed by [11], where \( \hat{y} = X\hat{\beta} + Zu \). The estimate takes into account that both the fixed effects \( \beta \) and random effects \( u \), as well as the covariance matrix
for the random effects $G$ need to be estimated. However, it does not take into
account that the constant $\sigma^2$ in (2.14) also needs to be estimated, but in a
supplementary report they propose

$$a_n(\beta, \sigma^2, \theta) = \frac{\hat{\sigma}^2}{\sigma^2} \text{tr} \left( \frac{\partial \hat{y}}{\partial y} \right) + \hat{\sigma}^2 (\hat{y} - y)^T \frac{\partial \hat{\sigma}^{-2}}{\partial y} + \frac{1}{2} \hat{\sigma}^2 \text{tr} \left( \frac{\partial^2 \hat{\sigma}^{-2}}{\partial y \partial y^T} \right),$$  (2.33)

for the case when $\sigma^2$ also needs to be estimated, where $\hat{\sigma}^2$ is the true error
variance and $\hat{\sigma}^2$ the estimated error variance. As Greven and Kneib states
in [9], the calculation of the penalty term in (2.32) and (2.33) is cumbersome
and becomes prohibitive when the number of observations $n$ becomes large. As
a remedy, Greven and Kneib proposes an analytical representation of (2.32). However the representation of it is quite complex and we leave it to the reader
to read the original paper, in order to fully grasp the derivation.

An analytical representation of (2.33) could also be achieved, but Greven
and Kneib state that the calculations become heavy and their simulations show
that just adding 1 to the error term works as a good approximation. An AIC
estimate with the penalty term proposed by Greven and Kneib exists in the R
language through the cAIC4 package and its implementation is outlined in [21].
As of yet, to the best of our knowledge, there does not exist such an estimate
in any of the statistical packages in Python.

Vaida and Blanchard [24] suggest that the penalty for the estimation of the
covariance matrix for the random effects $G$ can be ignored and proposes the
following measure for the penalty term

$$a_n(\theta) = \frac{2n}{n - p - 2} \left\{ \rho + 1 - \frac{\rho - p}{n - p} \right\},$$  (2.34)

where

$$\rho = \text{tr} \{ H \} = \left\{ (X^T V^{-1} X)^{-1} X^T V^{-1} R V^{-1} X \right\} + n - \text{tr} \{ R V^{-1} \},$$  (2.35)

and where

$$H = (X \ Z) \left( \begin{array}{cc} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{array} \right)^{-1} \left( \begin{array}{c} X^T R^{-1} \\ Z^T R^{-1} \end{array} \right),$$

is the "hat" matrix that maps the observed data vector, $y$, into the fitted vector
$\hat{y} = X \beta + Zu$, i.e., $\hat{y} = y H$. Again, $p$ and $q$ are the number of fixed and random
effects and $n$ the number of observations. Which gives us the asymptotic result
$2\{\rho + 1\}$ for $a_n$ in (2.34), as $n \to \infty$.

The reasoning behind why the penalty for estimating $G$ can be ignored is
that the conditional likelihood in (2.28) does not dependent on the covariance
2.5. Risk Analysis

matrix for the random effects $G$. Furthermore, if REML is used instead of ML, Vaida and Blanchard suggest keeping the conditional likelihood in (2.31), but replacing the penalty in (2.34) with

$$a_n(\theta) = \frac{2(n - p - 1)}{n - p - 2} \left\{ \rho + 1 - \frac{p + 1}{n - p - 2} \right\},$$

(2.36)

where the asymptotic result is the same as for the ML case.

2.5 Risk Analysis

The most used method to determine the risk associated with an asset is to measure the variance for its returns over a selected period. We have the following estimate for the variance

$$\text{Var}(x) = \sigma^2 = \frac{1}{T} \sum_{t=1}^{T} (r_t - \mu)^2,$$

or when the mean is not known, the unbiased sample variance

$$s^2 = \frac{1}{T - 1} \sum_{t=1}^{T} (r_t - \bar{r})^2,$$

where $r_t$ is the assets return at time $t$ and $\mu$ and $\bar{r}$ is the true and estimated mean return. However, by using the variance as a measure of risk we introduce a critical assumption; we value positive and negative deviations from the mean equally. This assumption might seem more or less reasonable, depending on the investor’s preferences, but for the investors who do not value the deviations equally there are measures such as the negative semi-variance [14] (drawdowns)

$$\frac{1}{T^*} \sum_{t : r_t < \mu}^{T^*} (r_t - \mu)^2$$

or

$$\frac{1}{T^* - 1} \sum_{t : r_t < \bar{r}}^{T^*} (r_t - \bar{r})^2$$

which only takes into consideration the negative deviations. $T^*$ is here the number of observations that are less than the mean or the sampled mean, i.e., $T^* = \sum_{t=1}^{T} 1_{r_t < \mu}$ or $T^* = \sum_{t=1}^{T} 1_{r_t < \bar{r}}$. 


2.5.1 Risk for a portfolio

When estimating the risk for a portfolio with two or more assets, the covariance between the assets, as well as the weight for each asset need to be considered. The covariance can be estimated from the data, while the weights are considered as known parameters and the return and variance for the portfolio can therefore be estimated as linear combinations

\[
  r_p = w^T r = \sum_{i=1}^{k} w_i r_i, \tag{2.37}
\]

\[
  \sigma^2_p = w^T C w = \sum_{i=1}^{k} w_i^2 \sigma_i + \sum_{i=1}^{k} \sum_{i=1}^{k} w_i w_j \sigma_{ij}, \tag{2.38}
\]

where \( w = (w_1, \ldots, w_k)^T \) are the weights, \( r = (r_1, \ldots, r_k)^T \) the returns, \( \sigma_i \) the variance of asset \( i \), \( \sigma_{ij} \) the covariance between asset \( i \) and \( j \) and \( C \) the covariance matrix for the assets in the portfolio. The covariance can be written as an expression of the correlation and standard deviations

\[
  \sigma_{ij} = \rho_{ij} \sigma_i \sigma_j, \quad -1 \leq \rho_{ij} \leq 1,
\]

which means that as the correlation \( \rho_{ij} \) between asset \( i \) and \( j \) varies, the total risk of the portfolio varies with it.

2.5.2 Market Risk and Idiosyncratic Risk

The volatility of the returns can be seen as a measure for the total risk for an asset. However, it is often useful to break down the risk into multiple parts, such as idiosyncratic risk, market risk, credit risk, liquidity risk and many more, depending on the asset. In this thesis we are primarily interested in the exposure to market risk; how much the returns for an asset or portfolio varies in response to changes in the market, and the idiosyncratic risk; the individual risk for each asset not explained by the market risk.

The exposure to idiosyncratic risk for a single asset is often hard to predict, since the risk is unique for the time and asset. However, this risk can be handled by constructing a portfolio with suitable correlation factors and weights, which was covered in the previous section. Meanwhile, Sections 2.5.4 and 2.5.5 provides two ways to measure the exposure to the market risk and the models found there will provide a framework for the model brought to light in this thesis.
2.5.3 Risk vs. Returns

Risk can generally be viewed as the uncertainty of future outcomes. In the financial markets, investors view higher returns as desirable, but are at the same time also risk-averse. A higher expected return is typically followed by a higher risk, since otherwise there would exist arbitrage opportunities. Of course, these opportunities does exist due to market ineffectiveness, but they are generally hard to find and usually exploited fast by sophisticated investors [14].

Mean-Variance

The mean-variance portfolio, proposed by Markowitz in [14], seeks to maximize the expected portfolio returns (2.37) while including the portfolio variance (2.38) as a penalty factor

$$\max_w w^T E[r] - \gamma w^T C w,$$

(2.39)

where $\gamma \geq 0$ specifies the investor’s risk aversion. A high risk aversion leads to a higher $\gamma$ and thereby gives the variance a greater impact in the optimization. Again, the maximization should be constrained so that all weights sum to one. Although, the optimization can be constrained further by, for example, not including short selling $w > 0$ or by not letting any security be too large in the portfolio $w < w_{\text{max}}$, etc.

Efficient Frontier

By letting the risk aversion $\gamma$ vary between 0 and a large number, we get an efficient frontier [14], which is significant, since no other portfolio compositions can achieve a mean-variance relation greater than the frontier. However, by also defining a risk-free rate, $r_f$, to which investors can lend and borrow without limit, a new frontier appears. By assuming investors can borrow and lend at $r_f$, portfolios can freely be created along the line that tangents the efficient frontier and that intersects $r_f$ on the vertical axis, which then represents the new frontier, called the capital market line.

2.5.4 Capital Asset Pricing Model

If we increase the number of securities to include all investment opportunities, the tangent point on the efficient frontier represents the market portfolio. The equation for the tangent is

$$E[r_p] = r_f + \frac{\sigma_p}{\sigma_m}(E[r_m] - r_f),$$

(2.40)
where \( r_p \) and \( r_m \) is the portfolio and market return. If we assume that all investors agree on the risk free rate, as well as the mean return and standard deviation for the market, all investor can easily calculate the estimated return for a portfolio on the second efficient frontier, for any given standard deviation. By acknowledging the fact that the return for any investment, not only the ones on the frontier, should reflect the contribution the investment has towards the risk for the market portfolio, the equation (2.40) can be rewritten as

\[
E[r_p] = r_f + \rho_{pm} \frac{\sigma_p}{\sigma_m} (E[r_m] - r_f),
\]

(2.41)

where \( \rho_{pm} \) is the correlation between the portfolio and the market portfolio. (2.41) is analogous to (2.40) when the portfolio is chosen on the second efficient frontier, since the correlation between any portfolio on the tangent and the market portfolio is equal to one.

The term \( \rho_{pm} \frac{\sigma_p}{\sigma_m} \) is in fact the most likelihood estimator for a coefficient in a simple regression model, with \( r_p - r_f \) as the response and \( r_m - r_f \) as the covariate. This means that the following regression model could be used to model the expected return for a portfolio

\[
r_p - r_f = \beta(r_m - r_f) + \epsilon,
\]

(2.42)

where

\[
\epsilon \sim N(0, \sigma^2) \implies E[r_p] = r_f + \beta(E[r_m] - r_f),
\]

where \( \beta = \rho_{pm} \frac{\sigma_p}{\sigma_m} \). As a result, the model in (2.42), the Capital Asset Pricing Model (CAPM) \[23\], provides a simple implementation for estimating the market risk exposure for a portfolio.

The theory behind CAPM rests heavily on an efficient market, while also requiring that

(i) idiosyncratic risk can be eliminated by diversification,
(ii) investors are risk averse and diversify their portfolios, and thus eliminating the idiosyncratic risk,
(iii) investors make their investment decisions based on risk and returns, i.e., the variance and the mean returns for the portfolio,
(iv) all investors seeks to maximize the utility of terminal wealth,
(v) all investors have the similar expected values for risk and returns,
(vi) all investors have use the same time horizon.
Lastly, an intercept is sometimes added to the CAPM model \(2.42\), which gives an expected return

\[
E[r_p] = \alpha + r_f + \beta (E[r_m] - r_f),
\]

where \(\alpha\) is the intercept. However, the theory of an efficient market implies that \(\alpha = 0\), since otherwise there would exist investors that could outperform the market. Although, the concept of \(\alpha\) is important in the financial sector, since it can be used to evaluate the performance of portfolios, post hoc.

### 2.5.5 Arbitrage Pricing Theory

The model for Arbitrage Pricing Theory (APT) given as

\[
r_p - r_f = \beta_1 f_1 + \cdots + \beta_k f_k + \epsilon = f^T \beta + \epsilon,
\]

where

\[
E[\epsilon] = 0 \implies E[r_p] = r_f + f^T \beta,
\]

generalizes the expression in CAPM by dividing up a portfolio’s exposure to risk among multiple factors, not just the market portfolio \([20]\). The assumptions are also less restrictive

(i) an asset’s returns can be explain by a set of factors, as in \(2.44\),

(ii) idiosyncratic risk can be eliminated by diversification,

(iii) the diversified portfolio in (ii) provides no arbitrage, opportunities,

and does not limit the the asset’s returns to be of any specific distribution, as the theory behind CAPM does. However, the random idiosyncratic risk \(\epsilon\) must have a zero mean, a stationary variance \(\sigma^2\) and be sufficiently independent, i.e., \(E[\epsilon \epsilon^T] \approx \sigma^2 I\).

### 2.6 Cluster Analysis

With cluster analysis the aim is to partition a set \(S\) of \(n\) observations \(S = \{x_1, \ldots, x_n\}\) into \(k\) clusters, where each observation \(x_i\) is a \(d\)-dimensional vector. A range of different methods for clusters analysis exist, such as more simple approaches that groups the data depending on their euclidean distance to each other or more advanced approaches that implements neural networks. In this thesis we will use a Gaussian mixture model, which sets out to cluster each data
point to one of \( k \) Gaussian distributions, depending on its \( d \)-dimensional value. The model is easier to grasp when one has an understanding of the \( k \)-means algorithm, since it is a natural generalization of this approach, which is why \( k \)-means is first introduced here.

### 2.6.1 K-means

The \( k \)-means algorithm sets out to minimize the within cluster variation simultaneously for all \( K \) clusters

\[
W = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathbf{(x}_n - \mu_n)^T (\mathbf{x}_n - \mu_n).
\]

The variable \( W \) above is the sum of within cluster variation, where the binary variable \( r_{nk} \) specifies if observation \( n \) belongs to cluster \( k \) or not. The value of \( r_{nk} \)

\[
r_{nk} = \begin{cases} 
1, & \text{if } k = \arg \min_j (\mathbf{x}_n - \mu_j)^T (\mathbf{x}_n - \mu_j), \\
0, & \text{otherwise},
\end{cases} \tag{2.45}
\]

is determined so that observation \( n \) is included in the cluster with a mean closest\(^{12}\) to the observation. The mean \( \mu_k \) for each cluster \( k \) is initialized to an appropriate value, which gives us our first estimate for the \( r_{nk} \), but by minimizing the sum of cluster variations \( W \) with regards to the means \( \mu_k \)

\[
\frac{\partial W}{\partial \mu_k} = 2 \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \mu_n) = 0,
\]

we end up with a new estimate for the mean vector

\[
\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n, \tag{2.46}
\]

where

\[
N_k = \sum_{n=1}^{N} r_{nk}.
\]

Since the mean vectors might have changed, the values for \( r_{nk} \) might have changed in response to this as well, which in turn might give new values for the mean vectors, and so on. The algorithm iterates like this until the binary values for \( r_{nk} \) have converged, see [13] for more details.

\(^{12}\)The measure of closeness is here the euclidean distance.
2.6.2 Gaussian Mixture

We define the vector $z$ to be any of the unit vectors in a $K$-dimensional space. If $z$ is the $k$th unit vector, then by definition, $z_i = 1$ if $i = k$ while $z_i = 0$ if $i \neq k$. The probability of $z$ to be the $k$th unit vector can then be written as

$$p(z_k = 1) = \pi_k, \quad k = 1, 2, \ldots, K. \quad (2.47)$$

If we assume independence between all $z_k$, we can define the distribution function for $z$ as

$$p(z) = \pi_1^{z_1} \pi_2^{z_2} \cdots \pi_K^{z_K} = \prod_{k=1}^{K} \pi_k^{z_k}. \quad (2.48)$$

If we assume that $x$ is an observation from a Mixture of Gaussians as defined in (2.5) and let each $z_k$ in $z$ represent a latent variable that determines if the observation is drawn from component $k$ or not. Following the definition in (2.5), the conditional distribution of $x$, given $z_k = 1$ (i.e., drawn from the $k$th component), is a Gaussian

$$f(x|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k),$$

which in turn can, given (2.48), be generalized into

$$f(x|z) = \prod_{k=1}^{K} f(x|z_k)^{z_k} = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}. \quad (2.49)$$

Combining (2.48) and (2.49) and using the law of total probability, gives us the marginal distribution for the observation $x$

$$f(x) = \sum_z p(z)f(x|z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k), \quad (2.50)$$

which is the definition of a Gaussian mixture, see (2.5). In other words, we now have defined the Gaussian mixture model using a explicit latent variable $z$, much like in the case for the $k$-means approach, where the latent variable is $r_{ij}$.

Since we now have the distribution for $z$ (2.48), the marginal distribution of $x$ (2.50) and the conditional distribution of $x$ given $z$ (2.49) and by utilizing Bayes theorem we get the conditional distribution for $z$ given $x$

$$p(z|x) = \frac{p(z)p(x|z)}{p(x)} = \frac{p(z)p(x|z)}{\sum_z p(z)p(x|z)}. $$
The probability that the observation \( x \) comes from the \( k \)th component is then
\[
p(z_k = 1|\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x} | \mu_j, \Sigma_j)}.
\]
(2.51)

Parameter Estimation

The log-likelihood for the Gaussian Mixture is defined as
\[
\mathcal{L}(X | \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left( \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k) \right),
\]
where \( X = (x_1^T, x_2^T, \ldots, x_n^T)^T \) is a \( N \times D \) dimensional matrix and where each observation \( x_i \) is drawn independently from the \( k \)-dimensional mixture. As for the \( k \)-means algorithm, the optimal mean vectors \( \mu_k \) and now also the optimal covariance matrices \( \Sigma_k \) need to be estimated. However, instead of minimizing the sum of within cluster variances, we now set out to maximize the log-likelihood \( (2.52) \).

So by maximizing the log-likelihood with respect to the mean vectors \( \mu_k \), we end up with a similar estimate as for the \( k \)-means algorithm \( (2.46) \)
\[
\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} p(z_k = 1|x_n) x_n,
\]
(2.53)
where
\[
N_k = \sum_{n=1}^{N} p(z_k = 1|x_n),
\]
(2.54)
and where \( r_{nk} \) \( (2.45) \) has been replaced by the more general \( p(z_k = 1|x_n) \) in \( (2.51) \). As stated previously, we now also need to consider the covariance matrix \( \Sigma_k \) and by maximizing the likelihood with respect to the symmetric\(^{13}\) covariance matrix \( \Sigma_k \), we end up with the following estimate
\[
\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} p(z_k = 1|x_n) (x_n - \mu_k)(x_n - \mu_k)^T.
\]
(2.55)

Lastly, the mixing coefficients \( \pi_k \) also needs to be estimated, but in order for the condition that the coefficients must sum up to one, a Lagrange multiplier is

\(^{13}\)By definition, the covariance matrix is always symmetric.
2.7 Expectation-Maximization

added to the expression

$$\frac{\partial}{\partial \pi_k} \left[ \mathcal{L}(X | \pi, \mu, \Sigma) + \lambda \left( \sum_{k=1}^{K} \pi_k - 1 \right) \right] = 0. \quad (2.56)$$

By solving for $\pi_k$ and by using the fact that $\lambda$ is here equal to $-N$, we get the following estimate for the coefficients

$$\pi_k = \frac{N_k}{N}, \quad (2.57)$$

and the weight for each Gaussian is thereby the number of observations that each distribution covers.

2.7 Expectation-Maximization

The expectation-maximization (EM) algorithm is an iterative method that consists of two steps, one estimation step and one maximization step [5]. The algorithm makes it possible to find maximum likelihood estimates for models that depend on latent variables. In the following section, we show how the EM algorithm is used in order to estimate the parameters in a Gaussian Mixture.

2.7.1 Gaussian Mixture

E-step

The latent variables in the Gaussian mixture model are the $z_k$ variables, which specifies the distribution that the observations $x_n$ most likely originates from

$$p(z_k = 1|x_n)^{(t)} = \frac{\pi_k^t \mathcal{N}(x_n | \mu_k^t, \Sigma_k^t)}{\sum_{j=1}^K \pi_j^t \mathcal{N}(x_n | \mu_j^t, \Sigma_j^t)}. \quad (2.58)$$

The initial values for the unknown mean parameters $\mu_k$ are often chosen as the best estimate from the $k$-means algorithm, which can in turn give appropriate initial estimates for the covariance matrix $\Sigma_k$ as well as the mixing coefficients $\pi_k$.

---

14Variables that are not observed in the data, but rather inferred from the data.
M-step

In the M-step, the unknown variables $\mu_k$, $\Sigma_k$ and $\pi_k$ are estimated using the values for the latent variables estimated in the E-step

$$\mu_k^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_{n=1}^{N} p(z_k = 1|x_n)^{(t)} x_n,$$

$$\Sigma_k^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_{n=1}^{N} p(z_k = 1|x_n)^{(t)} (x_n - \mu_k^{(t)}) (x_n - \mu_k^{(t)})^T,$$

and

$$\pi_k^{t+1} = \frac{N_k^{(t)}}{N},$$

where

$$N_k^{(t)} = \sum_{n=1}^{N} p(z_k = 1|x_n)^{(t)}.$$

The EM algorithm continues to iterate until the log-likelihood

$$\mathcal{L}(X|\pi, \mu, \Sigma)^{(t+1)} = \sum_{n=1}^{N} \ln \left( \sum_{k=1}^{K} \pi_k^{(t+1)} N(x_n|\mu_k^{(t+1)}, \Sigma_k^{(t+1)}) \right)$$

has converged.
Chapter 3

Method

3.1 Data Sampling

The data collected for this thesis will be dated between 2006-04-30 to 2021-04-30. All data concerning stock market indices, stock returns, stock volume, and other company-specific information will be gathered from FactSet\(^1\). However, the market data could also be gathered free of charge from a source such as Yahoo Finance, while the company-specific data can be collected by sourcing the data from each individual company. Other macroeconomic indicators such as inflation, bond yields, economic surveys, and more will be gathered from the Federal Reserve Bank of St. Louis (FRED\(^2\)) which provides an open-source API that developers can source data from.

3.1.1 Assumptions

Financial Assets

Backed by the extensive theory on distributions outlined in Section\(^2.2\) we will assume that the price of financial assets follows a geometric Brownian motion process, i.e., they have independent log-normally distributed returns. Furthermore, we also assume that the risk-free rate is analogous to the 10-year US treasury yield.

\(^{1}\)https://www.factset.com/
\(^{2}\)https://fred.stlouisfed.org/

Börjesson, 2021.
3.1.2 Response

Since we assume that the stocks under consideration have independent log-normally distributed returns, both the linear and mixed linear model, as defined in the theory section, can be used in order to model a specific asset’s returns. The models will be tested on a subset of the stocks currently in the S&P 500 index, which consists of 505 stocks issued by 500 large cap companies divided into 11 industrial segments. Of these 505 stocks, 251 of them were in the S&P 500 as of 2006-04-30 and this is the subset that will be used.

3.1.3 Fixed Factors

Only the excess returns for the S&P 500 will be used as a fixed factor in the models in this thesis, but the models can by all means be further generalized to incorporate multiple fixed factors.

3.1.4 Random Factors

The set of random factors consists of; (i) the excess returns for the S&P 500, (ii) the number of shares traded for each stock divided by outstanding shares (averaged across each month), (iii) the price-to earnings-ratio, and (iv) the debt to equity ratio.

The first random effect (i) makes it possible to model the shifts in market risk exposure throughout the time period, while the other three make it possible to add structure to the individual idiosyncratic risk for each stock.

3.1.5 Clustering Factors

In order to use the mixed model, as defined in Section 2.3.2, the observations in the response $y$ in (2.14) need to be labeled into different groups, or clusters. These clusters will be found by implementing an unsupervised classification approach, where the observations in each cluster will be assumed to come from different Gaussian distributions. Each stock in the S&P 500 is categorized into 11 different sectors, and as the optimal cluster factors most likely differ between the sectors, different sets have been chosen for each sector.

As will be explained further on, a limit for the number of observations has been set for each time period and model, due to the temporal structure in the time series. This means that we are also limited to the number of factors that can be used as cluster factors, since we would otherwise be struck by the curse of dimensionality. In order to catch as much information as possible, from a small number of factors, principal component analysis (PCA) will be utilized.
on each set of cluster factors $C_s$. PCA provides us with orthogonal factors that together explain the variance for the set of factors.

There is no need for any assumption regarding the distribution of the clustering factors, but in order to not let the variance of a few factors dominate, each factor must be standardized prior to the PCA decomposition.

### 3.2 Models

As has been stated in Sections 2.5.4 and 2.5.5, both the CAPM and APT are linear regression models. These models usually only includes fixed effects, but this thesis will aim to generalize it by also including random effects.

#### 3.2.1 CAPM Model

As a baseline model, we will use a univariate (fixed) factor model. We will refer to the baseline as the CAPM model, with its definition already given in (2.42).

#### 3.2.2 ADAPT Model

The mixed model proposed in this thesis is univariate with respect to the fixed effects, but the model has more incommon with the ATP model than the CAPM model, since we have allowed random effects to be included in the model. Furthermore, there is nothing that stands in the way of including additional fixed effects in the model. The main differences between the mixed model presented here, compared to the APT model, is that we have allowed the market risk exposure to vary during the time period and the idiosyncratic risk is no longer a stationary process. While the mixed model is built on the frame work of the APT model, it is in a way more adaptable, which is why we have chosen to name it Adaptive Arbitrage Pricing Theory (ADAPT).

As stated in Section 2.5.5, the APT model does not restrict the distribution of the response to be normally distributed. However, since the log-normal distribution is an adequate approximation of the market returns and since the linear mixed models are more manageable with normally distributed effects and residuals, we will here restrict the ADAPT model to have normally distributed errors.

Using similar terminology for the ADAPT model as in the theory for linear mixed models (but with the response $y_k$, in group $k$, replaced by each stock’s excess return $r_k - r_f$) we get

$$ r_k - r_f = x_k \beta + Z_k u_k + \epsilon $$

(3.1)
where

\[ u_k \sim N(0, G_k) \]

and

\[ \epsilon \sim N(0, \sigma^2 I). \]

This implies that

\[ E[r_k] = r_f + x_k \beta \]

and

\[ E[r_k | u_k] = r_f + x_k \beta + Z_k u_k, \]

where \( r_k, x_k \) and \( Z_k \) denotes the log returns and the fixed and random factors for a specific group \( k \). \( u_k \) is the random effects for that same group and \( \beta \) is the fixed effect, which is the same for all groups.

The generalization by including random effects in the ADAPT model could potentially make it more favourable compared to the APT model, primarily for the following two reasons; (i) The random effects will make it possible to view the residuals as non-stationary. This means that we as a result can view the response (excess log returns for each stock) in the regression model as a non-stationary process, which is often the case in the financial markets, see Section 2.2.4. (ii) Depending on the covariates used in the matrix \( Z_k \), the market risk as well as the idiosyncratic risk (discussed in Section 2.5.2) can be viewed as non-stationary as well. In Section 2.5.4 we explained that "the return for any investment, not only the ones on the frontier, should reflect the contribution the investment has towards the risk for the market portfolio". Considering this statement, it should not be hard to convince anyone that, in return, the market risk for any investment is prone to vary throughout the estimation window, which usually spans at least a few years of data. If this not was the case, then the risk an investment contributes towards the market portfolio should be stationary throughout the estimation window, which is quite a strong assumption to make.

In addition to allowing the stock returns be non-stationary in (i), the random effects might also be able to explain random shocks in the market, which otherwise would be regarded as outliers.
3.3 Training and Model Selection

All models, CAPM and ADAPT (with different sets of hyperparameters\(^3\)), will be trained on a set of \(T\) observations, where the \(T\) observations are chosen from a moving window that moves through the time series. In this thesis, the value of \(T\) is fixed at 60 observations (60 months of data), but it could be seen as a hyperparameter in future works. Figure 3.1 shows how the window will move through the time series, where at time \(t\) all models will be trained on the observations between time \(t\) and \(t - T\) and then give an estimate for the market risk exposure between time \(t\) and \(t + 1\). The total number of observations for training the models is 180 (15 years of monthly data), which together with a moving window of size 60, means that we have 120 number of time-periods to estimate the market risk exposure for. With the total number of stocks under consideration equaling 251, we get \(120 \times 251 = 30\,120\) market risk estimates to use, for each model, during the backtesting phase.

\(^3\)Includes the dimension of the Mixture of Gaussians (2- or 3D), the components used in the mixture of Gaussians, etc. Will be explained in more detail in Section 3.3.2.
3.3.1 CAPM Model

Training

By definition, only the excess log returns for the market portfolio will be considered as the fixed effect in the CAPM (baseline) model. Since there really is no such thing as a market portfolio per se, we will use the S&P 500 as a proxy for that portfolio.

During the training of the CAPM model, the option to include an intercept in the regression model will be seen as a hyperparameter. This means that two models for all 251 stocks across 120 time periods will be trained, i.e., a total of $120 \times 251 \times 2 = 60240$ models.

Model Selection

The regression model for stock $i$ at time $t$ with the log-likelihood in (2.10) that produces the lowest AIC score (2.29) will be selected. The $\beta_4^4$ generated by that model will then be used during the backtesting.

3.3.2 ADAPT Model

Training

The subset of random factors $Z \in Z$ will be seen as a hyperparameter, together with the subset of cluster factors $C \in C$ (i.e., the dimension of the Gaussian distributions representing the clusters). In this study, the number of clusters (i.e., the number of groups in the mixed model) will be set to ten, but could be seen as a hyperparameter as well for future works. In order to limit of models to evaluate, we will limit the subset of random effects to only include one to four factors and the cluster factors to only include two or three factors. The number of available random factors is equal to five (the four random effects outlined in Section 3.1.4 together with an intercept term) and the number of cluster factors is equal to seven, so we get

$$n_r = \sum_{i=1}^{4} \binom{5}{i} = \binom{5}{1} + \binom{5}{2} + \binom{5}{3} + \binom{5}{4} = 5 + 10 + 10 + 5 = 30$$

A scalar value and not a vector, since we are using a univariate model. A non-bold $\beta$ should in this thesis, unless stated otherwise, be interpreted as the market risk, rather than a typical coefficient in the definition of linear and mixed regression models.

The first seven PCA factors explains between 89.8% and 96.1% depending on the assigned sector for the stock, and in order to limit the number of models to evaluate, we will not include components beyond the first seven.
and

\[ n_c = \sum_{i=2}^{3} \binom{7}{i} = \binom{7}{2} + \binom{7}{3} = 21 + 35 = 56 \]

number of possible subsets. However, we will, for reasons that will become clear during the backtesting section, further constrain each subset of random effects to have to include the excess log returns for the S&P 500 as an element. This means that we instead have

\[ n_r = \sum_{i=1}^{2} \binom{4}{i} = \binom{4}{1} + \binom{4}{2} + \binom{4}{3} + \binom{4}{4} = 4 + 6 + 4 + 1 = 15 \]

number of subsets for the random factors. By also viewing the option to fit the model with ML or REML as a hyperparameter, we get a total of \( 2 \times 15 \times 56 = 1680 \) of models to evaluate for each stock and time period. Which ultimately result in \( 1680 \times 120 \times 251 = 50,601,600 \) models to train and evaluate.

**Model Selection**

Only mixed models with the same fixed effects can be compared with respect to the AIC score if we use the REML approach, but since we only consider S&P 500 as the fixed effect for all models, we never have this problem during the model evaluation.

Müller et al. in [16] states that mAIC is asymptotically equivalent to leave one cluster out cross validation, while cAIC is asymptotically equivalent to leave on observation out cross validation. This means that the mAIC is best to use in order to predict future values for the entire population (the population is here the time series for the response with time window \( t - T - t \)), while cAIC is best to use in order to predict future values for a specific cluster within the population.

During the backtesting phase, we will test models that both make estimates for the market risk exposure for the entire population as well for a specific cluster. We will therefor use the model with lowest mAIC (2.30) for when predicting the market risk exposure for the entire population (i.e., unconditional on any cluster) while the model with lowest cAIC (2.31) will be used for when predicting the risk exposure for the population from a specific cluster.

The penalty term for the mAIC is given in (2.30) and since only 60 observations is used, the finite sample approach is used rather than the asymptotic approach. As for the cAIC, the penalty is chosen to be the one proposed by Vaida and Blanchard in (2.34) and (2.36) for the ML and REML respectively, where again the finite sample case is preferred.
When comparing the AIC score for the CAPM model with the mAIC and cAIC for the ADAPT models estimated with the REML approach, we believed that we could alter the likelihood in the AIC score to account for the bias for the variance parameter. We did this by replacing \( n \) in (2.10) with \( n - p \), where \( p \) is the number of parameters used in the linear fixed model. However, this correction most likely failed, since there was a huge difference between the fixed and mixed models. We have therefore not included the REML approach when comparing the goodness of fit, which was our first research question. The REML approach is, however, included when we set out to construct market neutral portfolios.

3.4 Backtesting

In the previous sections, unique market risk exposures for each stock has been estimated for all models. The baseline model is again the CAPM model, while the mixed models will be referred to as the mADAPT and cADAPT. The prefix m and c specifies if we set out to estimate the exposure for the entire population or a specific cluster from that population. In order to not clutter the expressions going forward, \( \beta = (\beta_1, \beta_2, \ldots, \beta_{251}) \) is generalized to represent the estimate for any model \( m \) at any time \( t \), i.e., \( \beta = \beta^{(m,t)} = (\beta_1^{(m,t)}, \beta_2^{(m,t)}, \ldots, \beta_{251}^{(m,t)}) \).

3.4.1 Market Neutral Portfolio

In order to backtest the effectiveness in estimating the market risk, each model’s estimate at time \( t \) will be used to construct a market neutral portfolio. The market neutral portfolio is here defined equivalently to a zero-beta portfolio. However, each portfolio also needs an objective function to be maximized or minimized, restricted by the zero-beta condition. One such function is given in (2.39), which maximizes the return penalized by the variance of the returns. Another approach is to only maximize returns or to only minimize the variance. However, our aim is not to maximize the portfolios’ returns or to only minimize its variance, but to minimize the correlation with the market, while also keeping the variance low. This means that none of these objectives are straightforward candidates to use in our optimization.

A better candidate comes together if we replace one of the vectors of weights \( w \) in the variance expression, with \( \tilde{w} \). This expression is in fact the covariance between portfolio \( w \) and \( \tilde{w} \) and by letting \( \tilde{w} = \beta \), we get the covariance between a portfolio and the market risk for the same portfolio. Since our aim is to minimize the correlation between the portfolio and the market, while decreasing the variance, this expression is indeed the most suitable of the candidates. The
optimization then becomes

\[
\min_w \left| w^T \Sigma \beta \right|
\]

subject to \( w^T \beta = 0, \quad w^T 1 = 1, \) \hspace{1cm} (3.4)

where \( w^T \beta = 0 \) is the constraint for the zero-beta portfolio, and \( w^T 1 = 1 \) constrains the funds in the portfolio to be fully invested.

By looking at the definition of portfolio variance in (2.38) we can see that the objective expression in the optimization problem in (3.4) also benefits from a well diversified portfolio. However, in order to not let a few positions to still potentially dominate the portfolio, a limit \( \omega \) can be placed on each asset. Furthermore, since only a minority of all stocks in S&P 500 have a negative exposure to market risk (i.e., \( \beta < 0 \)), together with the constraint \( w^T 1 = 1 \), the portfolio must most likely be leveraged in order to achieve a zero-beta portfolio and a low value for the objective. Although the objective function should punish portfolios with excess leverage, so it is prudent to also include a limit for the leverage \( \ell \). The optimization then becomes

\[
\min_w \left| w^T \Sigma \beta \right|
\]

subject to \( w^T \beta = 0, \quad w^T 1 = 1, \quad \| w \|_1 \leq \ell, \quad \| w \|_\infty \leq \omega, \) \hspace{1cm} (3.5)

where \( \| w \|_1 \leq \ell = 3 \) (i.e., a maximum of 200% leverage) is the limit for the leverage and \( \| w \|_\infty \leq \omega = 0.1 \) is the limit for each position.

Since an investment that has a negative expected return is seldom attractive to investors, an expression that constrains the expected returns to be greater than zero could be added to the optimization. However, we found that the portfolios performed worse when including this condition, and thus decided to not include it.

3.4.2 Portfolio Comparison

The CAPM model gives us a single expected value for the log returns \( r \)

\[
E[r] = r_f + x \beta,
\]
while the ADAPT model gives us two expected values $E[r] = r_f + x_\beta$ and $E[r|u_k] = r_f + x_k\beta + Z_ku_k$ depending on if we condition the returns on the random factors $u_k$, i.e. the mADAPT and cADAPT models. As mentioned previously; for each time window, $t - T$ to $t$, the ADAPT model groups the observations into different clusters, where the random effects $u_k$ are unique for each cluster $k$. In the unconditional case, we make a prediction for the return $r_{t+1}$ based solely on the fixed effects for the entire time window. In the conditional case, however, we make a prediction for $r_{t+1}$ based on the assumption that the market conditions between $t$ and $t + 1$ reflects the conditions for a specific cluster $k$.

If we have no informed prior estimate to what cluster the observation at $t + 1$ will be labeled, then the unconditional approach is more suitable than the conditional case, while the reverse might be true if we have an informed prior estimate. In an efficient market; the best estimate for future conditions at time $t + 1$ in the financial market, are the observed conditions at time $t$. Although, the clusters are estimated based on the output from the actual economy or company-specific information, rather than financial assets, and these factors are generally not covered by the efficient market hypothesis.

The mADAPT and cADAPT will be compared against the CAPM and against each other. The market risk estimates for the CAPM and mADAPT will be just $\beta$, although with differing values. Meanwhile, the market risk estimate for the cADAPT will be $\beta + u_k^{(S&P 500)}$ where two different approaches will be tried out. The first will follow the reasoning in the efficient market hypothesis and the prediction for the label of the observation for the cluster factors at time $t + 1$ will be the label at time $t$. We will refer to this model as the backward-looking cADAPT model (cADAPTb). The second approach will assume that we can make an accurate prediction of the label at time $t + 1$, which we will refer to as the forward-looking cADAPT (cADAPTf). An investor could most likely predict future conditions for the economy better than what the efficient market hypothesis (if we extend the theory to cover economical development as well) states, but to assume that the same investor could make an accurate prediction for all 251 stocks across all 120 time periods are far from reasonable. So in practice, the result of the model should fall somewhere between the cADAPTb and cADAPTf models.

Taking into consideration the moving time window with size $T = 60$, the portfolios will be constructed at time $t_{60} = 2006-03-31 + 60 \times \text{months} = 2011-03-31$. At time $t_{60}$, we have our first estimate of the equity market risk exposure for the CAPM, mADAPT, cADAPTb and cADAPTf models and we can then make the first portfolio optimizations given by (3.4). Then, for each $t_i$, $i = 37, 38, \ldots, 120$, we get new estimates for the equity market risk exposures, which will be used to
continuously optimize the weights in each portfolio. At time $t_{120} = 2021-03-31$, we have updated the portfolios for the last time and at time $t_{121} = 2021-04-30$ we can compare the returns for each portfolio to the market portfolio.

The market portfolio is here chosen to be the S&P500, where an optimal estimate for the equity market risk exposure would yield a correlation equal to zero. This means, that the model with the correlation closest to zero is implied to be the model with the best market risk estimate, providing a straight forward approach to answer Q2.
Chapter 4

Result

4.1 Model Fitness Comparison

Figure 4.1 shows the AIC score for each model, where each datapoint is the average score for all 251 stocks. As can be seen, the ML approach achieves a higher AIC than the REML. This is inline with the theory presented in Section 2.3.2 and thereby not surprising. As the number of random effects increase, the cADAPT models, evaluated with cAIC, gets a higher and higher score, while the mADAPT, model evaluated with mAIC, actually sees a decrease in model fitness. This probably means that; as the complexity of the model increases, the clusters can be estimated better, and so the predictive capabilities for each cluster will become better. However, as we make the clusters more separable, the (unconditional) predictive performance for the entire population gets worse.

Looking at the scores for when REML is used, the baseline CAPM model performs better for all hyperparemeters. Either the mixed models fitted by the REML approach are not able to compete with the CAPM model, or the option to replace $n$ in (2.10) with $n - p$ (where, again, $p$ is the number of parameters used in the linear fixed model) made less sense than we originally thought. As we will see in the next sections, its probably the latter. We have therefor opted not to include the REML approach in 4.1 that outlines the the average for all stocks and timepoins.

In 4.1 we can clearly see that when the complexity reaches three and four random effects, the cADAPT models actually surpasses the CAPM model for large parts of the time period. This can be further visualized in Table 4.1, where the mean clearly shows that the model fitness for the more complex cADAPT models can be compered or even considered better than the CAPM model. The
first of the two research questions has thus been answered.

Table 4.1: The table outlines the comparison between the linear fixed and mixed models. As mentioned before, AICc measures the performance of the CAPM model, while mAIC and cAIC measures the performance of the unconditional and conditional ADAPT models. See Section (3.4) for more details regarding the different AIC scores. The conditional mixed models clearly outperforms the fixed model, as the complexity increases. $c$ is the dimension of the clusters and $e$ the number of random effects used.

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Mean AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AICc</td>
</tr>
<tr>
<td>$c$ 1 $e$</td>
<td>-175.3</td>
</tr>
<tr>
<td>$c$ 2 $e$</td>
<td>-175.3</td>
</tr>
<tr>
<td>$c$ 3 $e$</td>
<td>-175.3</td>
</tr>
<tr>
<td>$c$ 4 $e$</td>
<td>-175.3</td>
</tr>
<tr>
<td>$c$ 1 $f$</td>
<td>-175.3</td>
</tr>
<tr>
<td>$c$ 2 $f$</td>
<td>-175.3</td>
</tr>
<tr>
<td>$c$ 3 $f$</td>
<td>-175.3</td>
</tr>
<tr>
<td>$c$ 3 $f$</td>
<td>-175.3</td>
</tr>
<tr>
<td>$c$ 3 $f$</td>
<td>-175.3</td>
</tr>
<tr>
<td>$c$ 3 $f$</td>
<td>-175.3</td>
</tr>
</tbody>
</table>
Figure 4.1: AIC scores for the CAPM, mADAPT and cADAPT models. The plots on the left show the AIC scores when using a 2-dimensional Gaussian Mixture when determining the clusters, while the plots on the right show the scores when using a 3-dimensional Gaussian Mixture. The first row of plots shows the AIC scores when only using one random effect (the S&P 500), while the second, third and fourth row shows the scores when using the S&P 500 plus one, two, or three company-specific factors. The solid graphs represent the models that use ML when determining the market risk, while the dashed graphs represent the models that use REML.
### 4.2 Portfolio Comparison

#### 4.2.1 Leverage

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Mean Leverage (% added to portfolio)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CAPM</td>
</tr>
<tr>
<td>$c$ $e$ $r$</td>
<td></td>
</tr>
<tr>
<td>2 1 0</td>
<td>101.1</td>
</tr>
<tr>
<td>2 1 1</td>
<td>101.1</td>
</tr>
<tr>
<td>2 2 0</td>
<td>101.1</td>
</tr>
<tr>
<td>2 2 1</td>
<td>101.1</td>
</tr>
<tr>
<td>2 3 0</td>
<td>101.1</td>
</tr>
<tr>
<td>2 3 1</td>
<td>101.1</td>
</tr>
<tr>
<td>2 4 0</td>
<td>101.1</td>
</tr>
<tr>
<td>2 4 1</td>
<td>101.1</td>
</tr>
<tr>
<td>2 f 0</td>
<td>101.1</td>
</tr>
<tr>
<td>2 f 1</td>
<td>101.1</td>
</tr>
<tr>
<td>3 1 0</td>
<td>101.1</td>
</tr>
<tr>
<td>3 1 1</td>
<td>101.1</td>
</tr>
<tr>
<td>3 2 0</td>
<td>101.1</td>
</tr>
<tr>
<td>3 2 1</td>
<td>101.1</td>
</tr>
<tr>
<td>3 3 0</td>
<td>101.1</td>
</tr>
<tr>
<td>3 3 1</td>
<td>101.1</td>
</tr>
<tr>
<td>3 4 0</td>
<td>101.1</td>
</tr>
<tr>
<td>3 4 1</td>
<td>101.1</td>
</tr>
<tr>
<td>3 f 0</td>
<td>101.1</td>
</tr>
<tr>
<td>3 f 1</td>
<td>101.1</td>
</tr>
</tbody>
</table>

Table 4.2: The utilized leverage for the portfolios with different hyperparameters. Here, $c$ is the dimension of the clusters, $e$ the number of random effects and $r$ if REML is used or not.

Table 4.2 shows the mean leverage utilized in the portfolio. The CAPM portfolio clearly uses the greatest leverage, while the cADAPT models nearly use only half of what the CAPM model does. Furthermore, we can see in Figure 4.2 that the cADAPT models not only use lower leverage on average, but consistently across the time period as well. We can also see that the leverage for the portfolios is quite correlated.
Figure 4.2: Time series for the utilized leverage for the portfolios, i.e., the percentage added to the portfolios. The solid graphs represent the models that use ML when determining the market risk, while the dashed graphs represent the models that use REML.
4.2.2 Transactions

Table 4.3 shows the mean monthly transaction as a percentage of the total portfolio. The mADAPT model demands by far the most transactions, for all hyperparameters, while the CAPM model demands the least. However, an important note is that the transactions is displayed as a percentage of the leveraged portfolio and since the CAPM model takes on far more leverage, the actual transaction cost would most likely be closer to the cADAPT models, than what the table at first might suggests.

We have opted to not highlight the model that performs the best in Table 4.3, since it is not necessarily the model that achieves the lowest transaction cost that are the best one. However, all things equal, a low number of transactions are generally preferred since high transaction costs will lead to diminishing returns. This means that mADAPT model would need to achieve a low covariance with the market, in order to justify its high number of transactions.

The transactions needed at each time $t$ for the two cADAPT models are highly correlated. Furthermore, the cADAPT models’ need for transactions are both mildly correlated with the CAPM model’s need for transactions, which can be observed in Figure 4.3.
Table 4.3: Mean monthly transactions, as a percentage of the portfolios with different hyperparameters. Here, $c$ is the dimension of the clusters, $e$ the number of random effects and $r$ if REML is used or not.
Figure 4.3: Time series for the transactions as percentage of the leveraged portfolios. The solid graphs represent the models that use ML when determining the market risk, while the dashed graphs represents the models that use REML.
4.2.3 Performance

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Covariance (with S&amp;P 500)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CAPM</td>
</tr>
<tr>
<td><strong>c</strong></td>
<td><em>e</em></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td><em>f</em></td>
</tr>
<tr>
<td>2</td>
<td><em>f</em></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td><em>f</em></td>
</tr>
<tr>
<td>3</td>
<td><em>f</em></td>
</tr>
</tbody>
</table>

Table 4.4: Covariance between the S&P 500 and the portfolios with different hyperparameters. Here, *c* is the dimension of the clusters, *e* the number of random effects and *r* if REML is used or not.

Our goal was to construct portfolios which has a low correlation with the market, as well as a low volatility. Tables 4.4 and 4.5 shows the covariance and correlation between the portfolios and the market (S&P 500). The covariance and correlation is clearly lower for the portfolios based on the cADAPT models than for the portfolios based on mADAPT and CAPM. We can also see that the mADAPT model performs better than the CAPM model. This is in stark contrast to what the analysis for the model fitness suggest. Where the AIC scores suggested that the CAPM model should outperform the mADAPT model,
Table 4.5: Correlation between the S&P 500 and the portfolios with different hyperparameters. Here, \( c \) is the dimension of the clusters, \( e \) the number of random effects and \( r \) if REML is used or not.

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Correlation (with S&amp;P 500)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c )</td>
<td>( e )</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>f</td>
</tr>
<tr>
<td>2</td>
<td>f</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>f</td>
</tr>
<tr>
<td>3</td>
<td>f</td>
</tr>
</tbody>
</table>

as well as the cADAPT models for many of the hyperparameter sets.

We can also see that the backward looking cADAPT model outperforms the forward looking cADAPT model for almost all of the hyperparameters.

Table 4.6 and 4.7 shows the volatility and drawdown of the portfolios.

According to the theory of CAPM, a portfolio that has a beta equal to zero should have the same return as the risk-free rate. Since we only cared about minimizing the covariance with the market constrained by the portfolio beta to be equal to zero, one could make the assumption that the return should mimic that of the risk-free rate. However, since we probably never can estimate a perfect forward-looking beta, we can never completely remove it from our
Table 4.6: Volatility for the portfolios with different hyperparameters. Here, \( c \) is the dimension of the clusters, \( e \) the number of random effects and \( r \) if REML is used or not.

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Volatility (annualized from monthly returns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c )</td>
<td>( e )</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>f</td>
</tr>
<tr>
<td>2</td>
<td>f</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>f</td>
</tr>
<tr>
<td>3</td>
<td>f</td>
</tr>
</tbody>
</table>

The correlation for the 251 stocks individually with the market, is on average 0.573 during the same time period, so a correlation around 0.4 is of course a great achievement, but we should clearly follow the trend of the market to some extent. Together with the fact that the portfolios are leveraged, we expected the portfolios’ returns to fall somewhere between the risk-free rate and the market portfolio.

Figure 4.4 shows the performance based on the return on investment, and all portfolios actually outperform the market up to the point before the pandemic. During the pandemic, the CAPM and mADAPT models have greater drawdowns than the market, possibly due to their high leverage. On the other
### Hyperparameters

<table>
<thead>
<tr>
<th>c</th>
<th>e</th>
<th>r</th>
<th>S&amp;P 500</th>
<th>CAPM</th>
<th>mADAPT</th>
<th>cADAPTb</th>
<th>cADAPTf</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0.101</td>
<td>0.103</td>
<td>0.103</td>
<td>0.077</td>
<td>0.078</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0.101</td>
<td>0.103</td>
<td>0.101</td>
<td>0.078</td>
<td>0.078</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0.101</td>
<td>0.103</td>
<td>0.095</td>
<td>0.078</td>
<td>0.079</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0.101</td>
<td>0.103</td>
<td>0.093</td>
<td>0.079</td>
<td>0.079</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0.101</td>
<td>0.103</td>
<td>0.098</td>
<td>0.081</td>
<td>0.081</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0.101</td>
<td>0.103</td>
<td>0.098</td>
<td>0.080</td>
<td>0.080</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0.101</td>
<td>0.103</td>
<td>0.098</td>
<td>0.079</td>
<td>0.080</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0.101</td>
<td>0.103</td>
<td>0.098</td>
<td>0.079</td>
<td>0.079</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0.101</td>
<td>0.103</td>
<td>0.102</td>
<td>0.079</td>
<td>0.080</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0.101</td>
<td>0.103</td>
<td>0.100</td>
<td>0.079</td>
<td>0.079</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0.101</td>
<td>0.103</td>
<td>0.106</td>
<td>0.080</td>
<td>0.080</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0.101</td>
<td>0.103</td>
<td>0.104</td>
<td>0.079</td>
<td>0.080</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0.101</td>
<td>0.103</td>
<td>0.098</td>
<td>0.079</td>
<td>0.080</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0.101</td>
<td>0.103</td>
<td>0.100</td>
<td>0.079</td>
<td>0.080</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0.101</td>
<td>0.103</td>
<td>0.101</td>
<td>0.080</td>
<td>0.082</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0.101</td>
<td>0.103</td>
<td>0.101</td>
<td>0.080</td>
<td>0.081</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0</td>
<td>0.101</td>
<td>0.103</td>
<td>0.103</td>
<td>0.080</td>
<td>0.080</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0.101</td>
<td>0.103</td>
<td>0.102</td>
<td>0.079</td>
<td>0.080</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0.101</td>
<td>0.103</td>
<td>0.106</td>
<td>0.080</td>
<td>0.080</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0.101</td>
<td>0.103</td>
<td>0.104</td>
<td>0.079</td>
<td>0.080</td>
</tr>
</tbody>
</table>

Table 4.7: Drawdowns for the portfolios with different hyperparameters. Here, $c$ is the dimension of the clusters, $e$ the number of random effects and $r$ if REML is used or not.

In the other hand, the drawdowns for the cADAPT models are comparable (and in some cases less) to the market drawdown and they are also able to achieve relatively good returns during the recovery.

If we also consider the volatility of the models and calculate the Sharpe ratio for the best performing model and compare it with the market, we get that the ratio is significantly better for the best cADAPT model. In fact, even the worst performing cADAPT model generates a greater Sharpe ratio.
Figure 4.4: Return on investment for the optimized portfolios, compared to the S&P 500 index. The solid graphs represent the models that use ML when determining the market risk, while the dashed graphs represents the models that use REML.
Chapter 5

Discussion and Conclusion

5.1 Discussion

5.1.1 Method

There are primarily four things in the method that we think are worth discussing. (i) The first one is how we define the clusters, (ii) the second is why we decided not to include other factors than the market portfolio and the company-specific factors, while the third and fourth regard (iii) the stocks we selected to be included in our portfolios and (iv) the AIC scores that we used to evaluate the goodness of fit for each model.

(i) Our method to cluster the time period into different sub-periods, in order to catch the temporal structure of the market risk, is quite basic. Although, we utilized a dimension reduction method (PCA) in order to not have to check all subsets of the cluster factors, and also to find the components that contributed the most to the variability in the market. The method quickly experience the curse of dimensionality, which is why we only used 2 and 3 as hyperparameters for the dimension of the clusters. Since we only used 60 observation (5 years of monthly data) and looked for 10 different clusters, the space would be to sparse if we increased the dimensions further. A solution we thought of was to also see the number of clusters as a hyperparameter, but as we decreased the number of clusters well below 10 (say 5 or 6) we got that the covariance matrix $V$ in (2.14) became singular during the estimation, for many of the different clusters and random effects combinations that we tested. This could mean that a low number of clusters were not able to effectively catch the temporal structure in the time series and could possibly impact the performance of our portfolios, since we have fewer estimates for the market risk in the cADAPT models. Another solution
would then be to move on from the Gaussian mixture clustering method, to a method that better handles high-dimensional inputs. A method that also more effectively determines the number of clusters could possibly increase the performance of the portfolios, rather than seeing the number as a fixed parameter. However, this could possibly lead to a large increase in iterations, where many of them would render bad or useless results.

(ii) Next we will discuss the reason behind why we decided to not include other factors in the models, such as the interest rate or other financial indicators. We first thought of including some of the factors that were later used in the clustering algorithm, but it would be quite hard to find indicators that fits all stocks\(^1\) in each sector during the entire time period. Of course we could have let the baseline CAPM model become an APT model and see the set of indicators as a hyperparameter, and then have used the same factors as fixed effects in the mixed model. Although, we would then have to look at the APT model for each stock and then try to find a set of indicators that, on average, works well for all stocks. However, this could make the market risk estimate for some stocks quite poor, since we only care about the average, which would be very bad since we only really cared about the market risk in the first place.

(iii) The reason behind why only selected 251 stocks from the S&P 500 was that we otherwise would have to switch stocks in the portfolio, every time the S&P changed the constituents in the index. This is of course not such a hard task to solve, we only need to specify during which months the market risk for each stock would need to be estimated and which stocks to include in the portfolio optimization at time \(t\). However, we also considered the cost of estimating the market risk for all stocks in the S&P 500, which would effectively double the computation time. To estimate the market risk for the 251 stocks, for a total of 120 periods and for all sets of hyperparameters took us 7 days with a total of 80 CPUs, which means that a doubling of the number of shares would make the training take 14 days instead of 7 days. Of course the cost of letting the 80 CPUs run for the double amount time is not astronomical in any way, but we believed that the upside of having 500 stocks in the portfolio instead of 251 was not big enough. A portfolio of 251 stocks across all sectors is already very diverse and to include even more stocks would probably have small impact on the diversification. However, it might have an impact on the drawdowns and performance compared to the SP 500, since a majority of the stocks that have fallen out of the index, have done so due to bad performance.

(iv) Lastly, the AIC scores proved to be a good measure of fitness for the ML models, but not so good for the REML models. When we compared the mixed

\(^1\) We need all stocks to share the same fixed indicators, otherwise the market beta would become harder to compare between each stock.
models optimized with the REML method to the fixed model (with unbiased likelihood), we got an indication that the mixed models would produce less accurate estimates for the market risk. However, as we can see in Tables 4.4, 4.5, 4.6 and in Figure 4.4, the results for the cADAPT models are very similar for the ML and REML methods. In line with the theory, we should be able to compare models fitted with the REML method, as long as we use the same fixed effects. We had hoped that if we used a likelihood that took into account the bias for the variance term, we could be able to compare it with a mixed model that was optimized with REML. However, as we can see in Table 4.4 the account for bias has a larger effect on the mixed model than it has on the fixed effects model. This means that we probably can not compare a fixed model, with a mixed model when using the REML estimator.

5.1.2 Result

Given the success of the mixed models in estimating the market risk, we are surprised that the model is rarely mentioned in literature for this application. By looking at the definition of the mixed models in (2.14), we can see that we have actually implicitly created a mixture of Gaussian, as defined in Section 2.1.4, where the different clusters represent the Gaussian distributions. Since we used 10 different clusters, we effectively created a mixture of 10 Gaussian distributions. The random effects for the S&P 500 moves each distribution either to the left or right (negative or positive contribution to the fixed market risk), while the random effects for the company-specific factors contribute to the distribution’s variance.

As was further mentioned in Section 2.1.4, the mixture of Gaussians has the ability to approximate all continuous distributions. This means that we could, theoretically, approximate all continuous distributions. However, as was mentioned in section 2.2.4 the financial stochastic processes are not necessarily completely continuous, and it is not as easy to mix continuous and discrete random effects. Furthermore, the relationships between the effects and levels are not necessarily linear in nature. Although, there are non-linear mixed effects models that could set out to model these non-linear relationships as well.

We hoped that we could decrease the volatility of the portfolios, by setting out to covariance with the risk that each stock contributed to the market portfolio. However, the fact that a majority of the mixed linear models (with different hyperparameters) actually achieved higher returns than the market was a result that we did not expect. But as was observed in Table 4.7, we effectively decreased the drawdowns for the cADAPT portfolios, and together with the fact that the portfolios had a quite low correlation with the market, they could mitigate heavy losses to the portfolio. This can be observed in Figure 4.4 where
the cADAPT models were able to handle the first two months of the pandemic (from the perspective of the western countries) better than the market portfolio. Since the market portfolio is heavily influenced by the current state and current projections for the economy, we might thus have created portfolios that are less dependent on the economy. Although, we would have to test the model during a time period where the economy and market performed poorly during a longer period (i.e., when the fiscal and monetary policies where not as responsive and or extreme as they have been during the financial crisis and even more so during the pandemic) to know this for sure.

5.2 Conclusion

The results in this thesis suggests that linear mixed models provide better estimates for the market risk for stocks large (and stable) enough to take place in the S&P 500. The conditional ADAPT model is able to take into account the temporal structure in the financial time series. It does this by estimating a unique exposure to the market for each sub-window and by letting the idiosyncratic risk be drawn from distributions unique for each cluster. Furthermore, the fact that the backward-looking model shows results comparable to the forward-looking model, tells us that the trade-off in trying to predict the next month’s market conditions is quite small, since we are probably not able to make accurate predictions every single month anyway.

5.3 Future Work

More studies need to be made in order to state definitively if the mixed model is superior to the fixed model. This study focuses on large and established companies under a limited time period of 10 years and so we suggest that future work should also incorporate different asset classes and different time periods as well.

Furthermore, it could be worthwhile to further improve the clustering algorithm in order to come up with more appropriate clusters for the market risk exposure to vary between. Lastly, by adding even more levels to the model, such as sectors specific effects, might improve the model further, since the stocks in each sector could theoretically "help" each other, by providing more observations. Where each stock then adds its own fixed effect in order to account for its own unique exposure to the market, apart from the estimated sector effects.
Bibliography


Copyright

The publishers will keep this document online on the Internet – or its possible replacement – from the date of publication barring exceptional circumstances.

The online availability of the document implies permanent permission for anyone to read, to download, or to print out single copies for his/her own use and to use it unchanged for non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional upon the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility.

According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement.

For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its www home page: [http://www.ep.liu.se/](http://www.ep.liu.se/)

Upphovsrätt

Detta dokument hålls tillgängligt på Internet – eller dess framtida ersättare – från publiceringsdatum under förutsättning att inga extraordinära omständigheter uppstår.

Tillgång till dokumentet innebär tillstånd för var och en att läsa, ladda ner, skriva ut enstaka kopior för enskilt bruk och att använda det oförändrat för ickekommersiell forskning och för undervisning. Överföring av upphovsrätten vid en senare tidpunkt kan inte upphäva detta tillstånd. All annan användning av dokumentet kräver upphovsmannens medgivande. För att garantera äktheten, säkerheten och tillgängligheten finns lösningar av teknisk och administrativ art.

Upphovsmannens ideella rätt innefattar rätt att bli nämnt som upphovsmann i den omfattning som god sed kräver vid användning av dokumentet på ovan beskrivna sätt samt skydd mot att dokumentet ändras eller presenteras i sådan form eller i sådant sammanhang som är kränkande för upphovsmannens litterära eller konstnärliga anseende eller egenart.

För ytterligare information om Linköping University Electronic Press se förlagets hemsida [http://www.ep.liu.se/](http://www.ep.liu.se/)

© 2021, Lukas Börjesson