Scenario generation for financial market indices

Emil Ahlmann Østergaard

Kongens Lyngby 2012
IMM-B.Sc.-2012-3

Technical University of Denmark
Informatics
Building 305, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk IMM-B.Sc.-2012-3

Abstract (English)

The aim of this thesis is to generate scenarios of financial indices that can be used in the asset allocation decision process such that risk-return trade-off is optimised in accordance with the investment strategy. The statistical behaviour of the indices is described and both the correlation and dynamic structure of the weekly log returns are modelled. These models are used to generate scenarios with a high degree of credibility. Bootstrapping is also used to generate scenarios.

The result of modelling financial indices with PCA and GARCH and afterwards use this in generating scenario is an applicable method to get trustworthy scenarios that can be used in financial risk management and asset allocation. These methods give results with a higher degree of reliability if the scenario horizon is long-term compared to the bootstrapping that performs acceptable within a short time frame.

Keywords: Autoregressive conditional heteroscedasticity, Bootstrapping, Financial statistical modelling, GARCH, Principal component analysis, Scenario generation, Time series analysis.

Abstract (Danish)

Målet med denne afhandling er at generere scenarier for finansielle index, som kan bruges til at beslutte hvordan aktiver skal fordeles så risiko-afkast forholdet optimeres, samtidig med at investeringsstrategien er overholdt. De statistiske egenskaber for dataene er beskrevet og både korrelationer og den dynamiske struktur er modelleret for ugentlige logaritmiske afkast. Disse modeller bruges til at generere pålidelige scenarier med. Bootstrapping er også blevet brugt til at generere scenarier med.

Resultaterne e,r at modellering af finansielle index med PCA og GARCH modeller, og efterfølgende bruge disse til at generere scenarier er en anvendelig metoder er giver troværdige resultater der kan bruges til finansiel risikostyring som led i fordelingen af aktiver. Disse metoder giver resultater der har større pålidelighed, når tidshorrisonten for scenarierne er af længere varighed i forhold til bootstrapping, der præsterer bedre ved kortsigtede scenarier

Nøgleord: Betinget heteroskedasticitet, Bootstrapping, GARCH, Principal komponent analyse, Scenariegenerering, Statistisk modellering af finansiel data, Tidsrækkeanalyse.

Preface and acknowledgements

This thesis was prepared at the Department of Mathematics at the Technical University of Denmark in fulfilment of the requirements for acquiring a B.Sc. in Mathematics and Technology. This paper is based upon work done together with Peter Nystrup in a period from September 2011 to January 2012. The main part of the thesis is written by the author, although some parts have been made in close collaboration with Peter Nystrup.

The thesis deals with modelling financial indices using different statistical methods. This is used to generate scenarios five years ahead, which can be used by e.g. pension and hedge funds in asset allocation. The modelling is done by using R[1].

The thesis consists of this paper, containing different models used in generating scenarios, a dataset consisting of 11 stock, bond and rate indices and the scripts derived in the modelling process.

I would like to express my gratefulness to my supervisors Associate professor Lasse Engbo Christiansen (Department of Informatics and Mathematical Modelling, DTU) and Associate professor Kourosh Marjani Rasmussen (Department of Management Engineering, Operations Research, DTU). I thank them for fruitful meetings, great guidance, support and inspiration throughout the project. I would also like to thank Søren Agergaard Andersen for providing data for the project.

[1]http://www.r-project.org/ (Version 2.13.1.)

Special thanks go to my co-worker on this project, Peter Nystrup, for great cooperation and interesting discussions.

Lyngby, 20-January-2012

Emil Ahlmann Østergaard

Contents

Chapter 1

Introduction

The theory that underlies this thesis is mathematical finance, particular statistical finance. Therefore the emphasis is put on the statistical approach, and not on a speculative approach in the investment process. The general understanding of an investment is the commitment of an asset into different kinds of financial and non-financial products. It might be a bond, stamp collection, or property. The investor has no specific profile, it might be a private person, a pension or hedge fund, or a corporation, but all sharing the same motivating drift: the opportunity to gain profit or hedge. Any investment, be it financial assets or real assets has risk attached. This risk plays an important role in the investment process and any investor has to take the risk into account when looking at the potential return. There are three main steps in the decisions part of the investment process:

- Capital allocation: The investor has to decide his/her exposure to risk. How much capital should be invested in risky assets and how much in "risk-free" assets must be answered by considering the expectation for the risk-return trade-off. The investment horizon should also be considered.
- Asset allocation: Decide which asset class to invest in. There are many types of asset classes, but real assets and financial assets are the major classes. Examples of real asset are gas and oil, timber and real estate. Financial assets are fixed-income assets (bonds), equities (stocks) or cash

equivalents. Usually the asset classes with an acceptable risk level, time horizon and best risk-return trade-off are chosen.

- Security selection: Decide which specific securities to invest in. Usually the composition of securities giving the best risk-return trade-off is chosen.

This thesis deals with the asset allocation step by inspection scenarios that has been generated using statistical modelling of different financial indices tracking the three major markets in the financial asset class: fixed-income assets, equities and cash equivalents.

1.1 Scenario generation

A scenario is a description of how a sequence of actions or events might evolve. Regarding this project, a scenario is the possible future values of a financial index and not a exact prediction. The index value should also be accompanied by a description of uncertainty in order to be a satisfactory scenario. Scenarios in this thesis are modelled using time series models that take the properties of historical data into account.

What characterizes a good scenario? Besides a description of accuracy or uncertainty, the scenarios should possess correctness and consistency. Correctness mean that scenarios must obey several empirical characteristics of data. That might be non-negative index values to satisfy the no arbitrage principle etc. Correctness is also the tendency in a scenario to act like historical data but also to explain events that have not been seen before. Scenarios should be consistent, e.g. the cross correlation between indices has to be reasonably constant. The quality of the generated scenarios should be tested in order to prove the usefulness of them.

There are several different approaches when generating scenarios. In this thesis historical data is used, and there are several different methods for generating scenarios. The Monte Carlo method is widely used, and is also used in this project together with historical data. The approach often depends on the use of the scenarios, is often risk management or strategic asset allocation. Portfolio and risk managers' investment related decisions highly rely on the scenarios, and their uses have many different applications. E.g. the allocations that performs the best if the best and worst performing scenarios are identified and used as a frame of reference. Then the maximum and minimum of the risk-return trade-off arefound, assuming the investor is acting rationally. This is also known as max-min optimization and maximization. The scenarios might also be used to find the allocation that over all performs the best if the average of all the scenarios is used or just the average scenario.

The majority of this thesis is concerned wiht data analysis and modelling of data and the result from this will be used to generate scenarios.

1.2 Problem statement

This thesis is concerned with the second part of the asset allocation decision only, which is scenario generation to be exact. As outlined in the introduction, the generated scenarios are of outmost importance to the investment decision process and the risk management in for instance a pension fund. Generating sufficient scenarios is therefore a practical problem of high relevance.

The asset classes considered are limited to money market instruments, bonds, and stocks, with the goal of establishing the correlation between these asset classes. Inclusion of the other major asset classes remains a possibility for future work.

The data available is twelve years and seven months of daily values of eleven different indices covering the period from 1st of January 1999 to 12th of August 2011. Six of them are stock indices, four are bond indices, and a Danish LIBOR index, which will serve as the link to the money market. For the LIBOR-index, the data is only available from 16th of June 2003 and onwards. The indices will be explored in more detail in the following chapter.

The purpose of the project is to analyse the index data with the aim of generating scenarios that can form the basis of decisions regarding strategic asset allocation. A scenario in this sense is the future values of the indices. The time horizon of the generated scenarios will be five years, which is a reasonable horizon for a short term, strategic asset allocation decision. With ten years index data available, it would not be meaningful to look at a longer horizon than five years. There will be generated a number of scenarios, and the quality of these scenarios will be tested.

The analysis will proceed according to the following steps:

1. The raw data is analysed for outliers, distribution, trends, autocorrelation, and cross-correlation.
2. A time series model is chosen and calibrated to the index series.
3. The model performance is tested on the data.
4. There will be generated scenarios using two or three different methods, and the quality of the scenarios will be assessed.

The analysis will be conducted using the statistical software R. There will be no prejudices as to what class of models that will be the better choice. The approach that will be used is therefore, through thorough data analysis to determine the necessary properties of a time series model that are able to describe the observed main features of the index data.

Part of the project work has been done in collaboration with Peter Nystrup, but the model chosen by him in connection with point two on the above list is different from the model that will be presented in this thesis. As a consequence, also the work done in connection with point three and four will differ. Apart from this subsection presenting the problem statement, the two theses have been written independently. In the concluding chapter, a comparison to the results from Peter Nystrup's work [22] will be part of the discussion.

1.3 Thesis overview

In chapter two the indices used in this thesis is presented in order to give the reader an extensive insight of the dynamics behind the indices. This is followed by an analysis of raw data in chapter three and an analysis of returns and log returns in chapter four in order to get as much knowledge about data as possible. In chapter five the data is divided into financial regimes. In chapter six the theory and models used in the later chapters are presented. In chapter seven the information about the data and log return data is used in order to find models and methods that fit data the best. After the modelling in chapter six the scenarios are generated in chapter eight with two different approaches. In chapter nine the scenarios and the method behind them are tested. At last the two different methods of generating scenarios are discussed, and ideas for further work are suggested. At the very end the appendix is found, where the R-scrip for the thesis is placed.

Chapter 2

Description of data

The purpose of this project is to generate scenarios that can be used in the decision process of allocating assets for investments. The allocation depends on risk-return trade-off estimated on the basis of the generated scenarios. This thesis is concerned with financial assets, therefore different indices from the three main assets class are used, namely fixed-income assets, equities and cash equivalents. Søren Agergaard Andersen has provided eleven different indices representing markets from all over the world, though mainly from developed countries. Data is available from the 1th of January 1999 to the 12th of August 2011, only DK00S/N starts at 16th June 2003. The data consists of daily (Monday to Friday) index values. If an index is not traded on a given day the value from the day before is used. These indices are widely used among investors and managers as benchmark etc. The compositions of the underlying securities sometimes change in order to keep the index tracking what it is meant to track. Often there is a set of rules and guidelines for the indices. These rules, guidelines and composition have been hard to find because as the company providing the indices want to held the information secret and only shares it with costumers. The indices used in this thesis will be presented below, some with more facts and information than other, but there is enough knowledge on each index to use it in the modelling and the scenario generation process [2, 10, 13, 16, 17, 18, 21].

2.1 Equity indices

KAXGI (OMX Copenhagen Stock Exchange All Share Perform Index, DKK)

This index has base date 31th December 1995 with base 100 . It consists of all the shares listed on Copenhagen Stock Exchange, and shows a general picture of the status and changes in the Danish market. It is a total return gross dividends index (GI) that shows the true performance of the index. A gross index is characterized by adjusting the index for dividends, and not including tax credits. A gross index shows a more accurate performance and measure of the total return because all the dividends are reinvested.

Morgan Stanley Capital International (MSCI) Equity Indices:

In this project four MSCI equity indices are used. They are daily total return net dividends indices in US Dollar. Net dividends indices are characterized by the reinvesting of the dividends after deduction of tax credit and withholding taxes. The tax rate used for international indices is a rate fit for use to non-resident institutional investors without profiting from double taxation treaties. The daily total return indices reinvest the dividends of the index at closing price the day the stock goes ex-dividend. All indices are free float adjusted which means that the equities listed in the index are adjusted such that the amount represented in the index is reflecting the amount available on the market. The indices are weighted by market capitalization.

The MSCI indices are often used when construction exchange-traded funds (ETF) which are securities or some financial products tracking an index.

- **NDDUE15** (MSCI Daily Total Return Net Europe, USD) measures the price equity performance of the developed European markets. NDDUE15 consists of the following 16 developed market country indices: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom. It has base day 31th December 1969.

- **NDDUJN** (MSCI Daily Total Return Net Japan, USD) is designed to measure the equity performance of the Japanese stocks listed on Tokyo Stock Exchange, Osaka Stock Exchange, JASDAQ and Nagoya Stock Exchange. It has base day 31th December 1987.

- **NDDUNA** (MSCI Daily Total Return Net North America, USD) measures the equity performance of the North American markets. On the 6th May 2010 the country weightings are 9.2% Canadian equities and 90.8% equities from USA. The total index market capitalization was $\$\,11,674,798\times 10^6$ [19] and the largest (weighted) sectors are information technology (17.74%), financials (17.42%), energy (12.66%), health care (10.96%), consumer staples (10.18%) and industrials (10.13%). Large companies trading all over the world are widely represented in the index e.g. Microsoft Corp, Coca-Cola CO, General Electric CO, Goldman Sachs Group and McDonald's Corp. The above composition is changing through time, and can only be used to give an idea of the index composition in the period that is studied. It has base day 31th December 1969.

- **NDUEEGF** (MSCI Daily Total Return Net Emerging Markets, USD) measures the equity performance of emerging markets. The following 21 emerging market country indices are used in the index: Brazil, Chile, China, Colombia, Czech Republic, Egypt, Hungary, India, Indonesia, Korea, Malaysia, Mexico, Morocco, Peru, Philippines, Poland, Russia, South Africa, Taiwan, Thailand, and Turkey. Emerging markets are fast developing countries that are in a process of industrialization. It has base day 31th December 1987.

NDDUE15, NDDUJN and NDDUNA reflects together a global performance for industrialized/developed countries, but they are also usable in studies of differences in continental (North America, Europe and Asia (Japan)) performance.

TPXDDVD (Tokyo stock Price IndeX Total Return, JPY)

TPXDDVD is a Japanese stock index representing the total return of the Tokyo stock Price IndeX (Topix) in JPY. It has base day 1th April 1968. It is directly comparable to KAXGI when taking the difference in currency into account.

2.2 Fixed income indices

CSIYHYI (J.P. Morgan High Yield Bond Index Global, USD)

This index tracks an investment fund called J.P.M. Global High Yield Bond Fund. The fund consists of a diversified portfolio. Diversification is a way to manage financial risk, where the risk is lower and the return on average is higher for the portfolio than for any single bond within the portfolio. The portfolio consists mainly of bond, unrated securities and some investment grade securities, where the issuers are corporations and banks from developed countries. The risk on bonds is often described by rating agencies e.g. Moody's and Standard & Poor's where the rating depends on credit quality of the corporation, bank etc. A rating of AAA equals a very low risk and D meaning debt in arrears. Often government bonds are considered as zero risk bonds rated above AAA. Investment grade securities are securities rated BBB or higher, and bonds below that level are known as junk bonds attractive to speculative investors. Some securities are not rated, which have different reasons. Sometimes the issuers cannot provide the information needed, or the total securities issued are relative small. A high yield bond is normally a bond rated below investment grade, where the investor speculate on a high return and calculate with higher risk. This index reflects the more volatile bonds where a higher risk is accepted to higher returns.

The composition of the portfolio changes but in 2011 the following facts are gathered. The fund received on the 30th September three starts on the Morningstar Rating, which ranks the fund in the middle group when ranking funds after risk and return adjusted for expenditures. At 31th October, 94.7% of the bond in the fund is rated BBB or less. Corporate bonds account for 95.1%, the average duration is 4.4 years and the average time to maturity is 6.3 years. Duration is the change in bond price when the interest rate changes. A large duration is equal to a large interest rate risk or large change in bond price. Duration is a bit like the maturity, but takes coupon in to account and is a weighted measure of the time the bond will pay out. The primary sectors represented in the fund are communications (18.2%), consumer cyclical (17.8%) and consumer non-cyclical (17.3%). American bonds and securities have a huge weight in the fund with 88.5% and UK bonds and securities are weighted 3.1%. Bonds and securities from non-developed countries have a small weight in the fund portfolio, Bermuda (0.3%) and Liberia (0.3%). To summarize, this fund consists mainly of high risk, low rated American corporate bonds. The funds prediction of the yield to maturity is 4.4%.

JPGCCOMP (J.P. Morgan Emerging Markets Bonds Index Global Diversified, USD)

This index tracks debt securities issued by 33 emerging markets countries rated BB+ by Standard & Poor's, e.g. Russia, Brazil and Mexico and was created in 1997. It tracks the total return of USD denominated Eurobonds and sovereign bonds with an outstanding face value of at least $\$500 \times 10^6$. Eurobonds are bonds issued in one country, but denominated in another currency, here USD.

Sovereign bonds are government bond, often from an emerging market country, issued in a foreign currency. An example of a sovereign bond is the Brady bond, which is one of the most liquid emerging markets bonds, where the issuer is a government in a developing country.

This index is a capitalization-weighted index, which means that the individual bonds are weighted according to their market capitalization. It provides some indication of expectation to a part of the emerging bond markets, but not the entire market because of rules on how countries with larger debt stock have limited weights or are excluded from the index.

NDEAGVT (Nordea Government Bond Index, DKK)

This is a government bond index denominated in DKK.

NDEAMO (Nordea Mortgage Bond Index, DKK)

NDEAMO is an index tracking mortgage bond denominated in DKK. In November 2006 the index had a composition of 63% callable mortgage bonds, 22% capped floaters and 15% non-callable mortgage bonds. Callable bonds allow the issuer of the bond to redeem the bond prior to the maturity date. A floater is a bond with a varying coupon rate determined by the short-term interest rate. If it is capped, the coupon rate has an upper limit. The modified duration is 5.8% per year for the index and it has a convexity of -1.9, which is a measure of the sensitivity of the duration, to changes in the interest rate. It is very common that mortgage (callable) bonds have negative convexity, meaning that the duration decreases when the market yields decrease.

2.3 Money market indices

DK00S.N.Index (London InterBank Offered Rate, Spot Next, DKK)

This is an interest rate index that tracks the spot/next (S/N) London InterBank Offered Rate (LIBOR). The LIBOR is the daily fixed interest rate that banks use when lending money in the London interbank market. The LIBOR is based on interbank deposit rates for larger loans with maturities from one day to one year offered by creditworthy banks. Spot/next means that the asset is handed over the day after the spot delivery date, which often is two business days after the day the transaction was made. The day count convention used is actual number of days divided by 360, and is commonly used in money markets.

It is not possible to invest in the LIBOR, but it might anyway reflect the expectations to the money market. Data for the DKK LIBOR is only available from 16 June 2003, where the fixing began.

CHAPTER 3

Analysis of index prices

This chapter deals with analysis of data from a statistical point of view. Before the process of modelling data, data needs to be examined and analysed. Through the analysis pattern in data or other important structures might be found. This knowledge will reduce the number of known usable statistical models, and ease the process of finding a model that fits the data. In this chapter the raw indices will be inspected, and in the next chapter the log returns are analysed.

In figure 3.1 the raw data is plotted. The indices have been plotted separately for the clarity. As it is seen, the indices within the same category have some of the same pattern. The stock indices seem to be more volatile in short (daily) basis than the bond indices because of the high fluctuation, but they also seems to be more sensitive to changes in the market causing a more distinct alternation. By looking at the stock indices there seem to be different types of periods with growing and falling prices distinguish by length of period, volatility and slope. Data starts in an ascending period and ends in the beginning of a decreasing period. Throughout out the whole period there seems to be three periods with growing prices and two periods with falling prices. These periods are not that distinct for the bond indices. Later on, this pattern will be compared to OECD's dates for financial peaks and crisis. As already pointed out, data for the LIBOR rate index only exists from 16th June 2003. It has a high volatility in short term and from 2009 until 2011 it takes a massive fall from DKK 675 to DKK 2. It does not seem to have much correlation with the other indices, but further analysis will clarify if that is true.

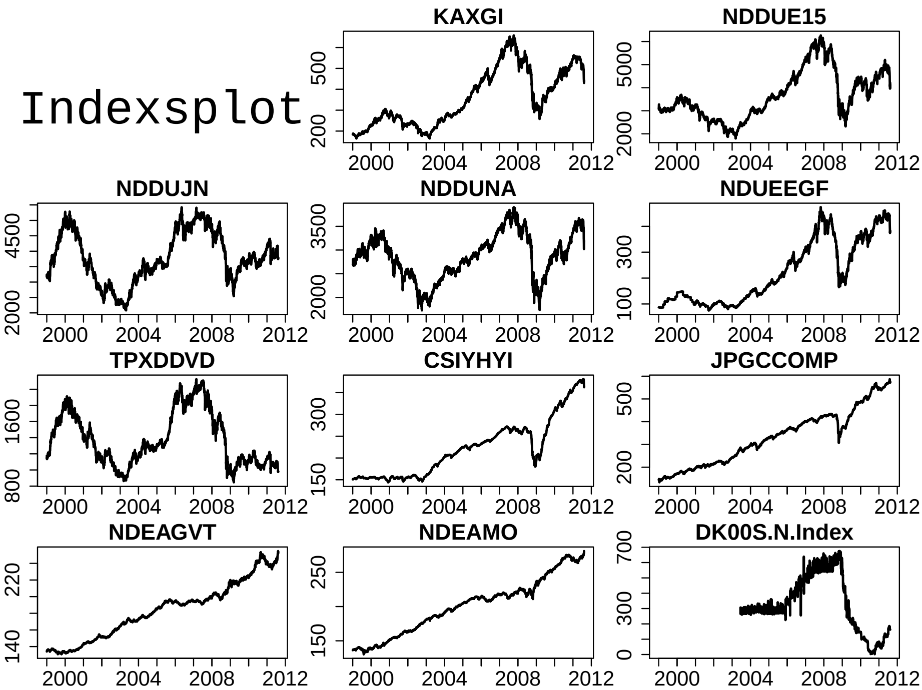


Figure 3.1: Index plot of the eleven indices, with time the on first axis and the index value on the second axis.

3.1 Generating data for NDUEEGF

Though it is not possible to see in figure 3.1, NDUEEGF only has monthly data from 1 January 1999 to 29 December 2000 where it has base date with base 100. This causes some troubles in the further analysis and modelling, so daily data is generated using the know information. This is done by using stepwise linear interpolation between the monthly data and adding normal distributed noise, with mean and standard deviation σ. The noise used is normal distributed with zero mean and standard deviation estimated as:

$$\sigma = \frac{SD(\bar{x})}{\sqrt{22}} = 1.6849,$$

where 22 is the average number of bank days in a month and $\bar{x}$ is a vector with the monthly change in index price for the two years. The generated data is plotted in figure 3.2. It is not known for sure that the volatility for new data is true but it seems reasonable compared to the known values without any suspicious outliers. There might have been a few outliers, but in the further modelling process they would have vanished because we are interested in long term asset allocation and not in single events.

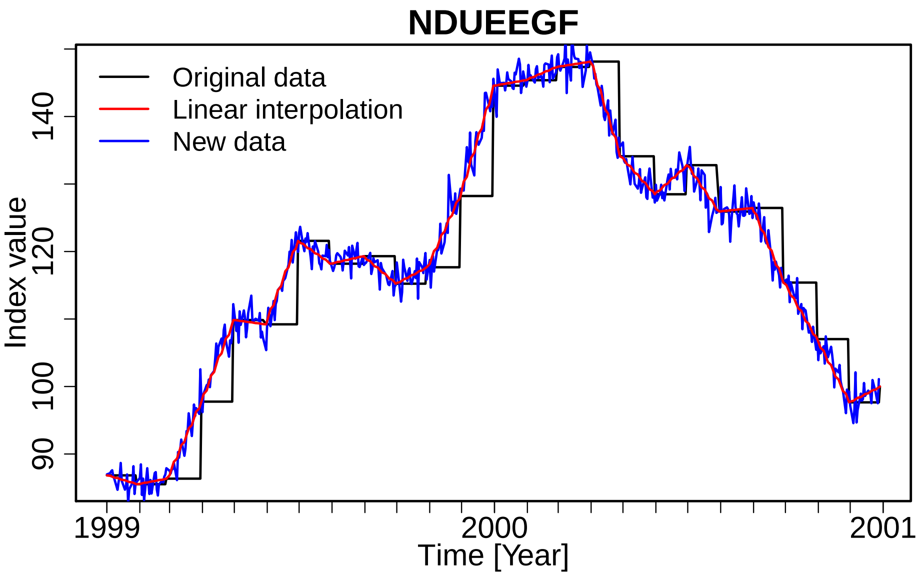


Figure 3.2: The original monthly data for NDUEEGF together with the linear interpolation and the new data generated by the linear interpolation and normal distributed noise.

3.2 Correlation

Figure 3.1 gives an indication on correlation between the indices, especially it is easy to see a similarity in the indices' behaviour within an index type. The reactions on one financial market can easily spread cross-border because almost all assets are traded online, and the digitisation has removed these limits. Therefore the financial markets are expected to have high correlations coefficients. The correlation between two indices is calculated as [14]:

	1	2	3	4	5	6	7	8	9	10
1 KAXGI										
2 NDDUE15	0.95									
3 NDDUJN	0.83	0.87								
4 NDDUNA	0.93	0.95	0.85							
5 NDUEEGF	0.88	0.83	0.56	0.79						
6 TPXDDVD	0.56	0.63	0.89	0.62	0.17					
7 CSIYHYI	0.66	0.54	0.28	0.60	0.85	-0.14				
8JPGCCOMP	0.65	0.52	0.24	0.51	0.88	-0.20	0.96			
9 NDEAGVT	0.43	0.26	-0.01	0.23	0.71	-0.42	0.85	0.94		
10 NDEAMO	0.44	0.29	0.01	0.26	0.73	-0.40	0.88	0.95	0.99	
11 DK00S.N	0.23	0.38	0.46	0.23	-0.04	0.62	-0.49	-0.42	-0.51	-0.54

Table 3.1: Correlations between the indices.

$$\rho_{X,Y} = \frac{E\left[(X-\overline{X})(Y-\overline{Y})\right]}{\sigma_X \sigma_Y}$$

where $\overline{X}$ and $\overline{Y}$ are the means and σ_X and σ_Y are the standard deviation on the indices. The correlations between all the indices are found in table 3.1. In general, the indices are positively correlated, especially the stock indices are highly correlated with each other. Only DK00S.N.Index seems to behave a little independently. Also the correlation between the Danish bond indices and the Japanese stock index is almost zero.

3.3 Autocorrelation in indices

It would be reasonable to think that the index value today has some dependency on yesterday's value, because the today's trading price starts at yesterday's closing price. This dependency is also known as autocorrelation. Generally speaking autocorrelation is the correlation of a time series with its own past and future. The autocorrelation of an index is described by the autocorrelation function (ACF). The coefficient in ACF, is given by [14]:

$$\begin{aligned}\rho_{XX}(t_1,t_2) &= \frac{\gamma_{XX}(t_1,t_2)}{\sqrt{\sigma^2(t_1)\sigma^2(t_2)}}\\ &= \frac{\mathrm{Cov}\left[X(t_1),X(t_2)\right]}{\sqrt{\sigma^2(t_1)\sigma^2(t_2)}}\\ &= \frac{\mathrm{E}\left[\left(X(t_1)-\mu(t_1)\right)\left(X(t_2)-\mu(t_2)\right)\right]}{\sqrt{\sigma^2(t_1)\sigma^2(t_2)}},\end{aligned}$$

where $\gamma_{XX}(t_1,t_2)$ is the autocovariance function, $\mu(t_j)=\frac{\sum_{i=1}^{j}X_1}{j}$ is the mean of the index until time t_j and $\sigma^2(t_j)$ is the variance of the process until time t_j. If the process is stationary the coefficients at lag $\tau=t_1-t_2$ simplifies to:

$$\rho_{XX}(t_1,t_2) = \frac{\gamma_{XX}(\tau)}{\sigma_X^2},$$

with $\gamma_{XX}(\tau)=\mathrm{Cov}\left(X(t),X\left(t+\tau\right)\right)$ and σ_X^2 being the variance of the process.

The Partial autocorrelation function (PACF) is a measure of the conditional correlation in time of a time series. At lag $\tau=1$ the coefficient in PACF is equal to the coefficient ACF [14].

For $\tau=2$ the coefficient is given by:

$$\rho_{XX}(\tau)=\frac{\mathrm{Cov}\left[X_t,X_{t-2}|X_{t-1}\right]}{\sqrt{\sigma^2\left(X_t|X_{t-1}\right)\sigma^2\left(X_{t-2}|X_{t-1}\right)}},$$

For $\tau=3$ the coefficient is given by:

$$\rho_{XX}(\tau)=\frac{\mathrm{Cov}\left[X_t,X_{t-3}|X_{t-1},X_{t-2}\right]}{\sqrt{\sigma^2\left(X_t|X_{t-1},X_{t-2}\right)\sigma^2\left(X_{t-3}|X_{t-1},X_{t-2}\right)}}$$

and for $\tau>3$ the procedure is the same as above just with more conditions.

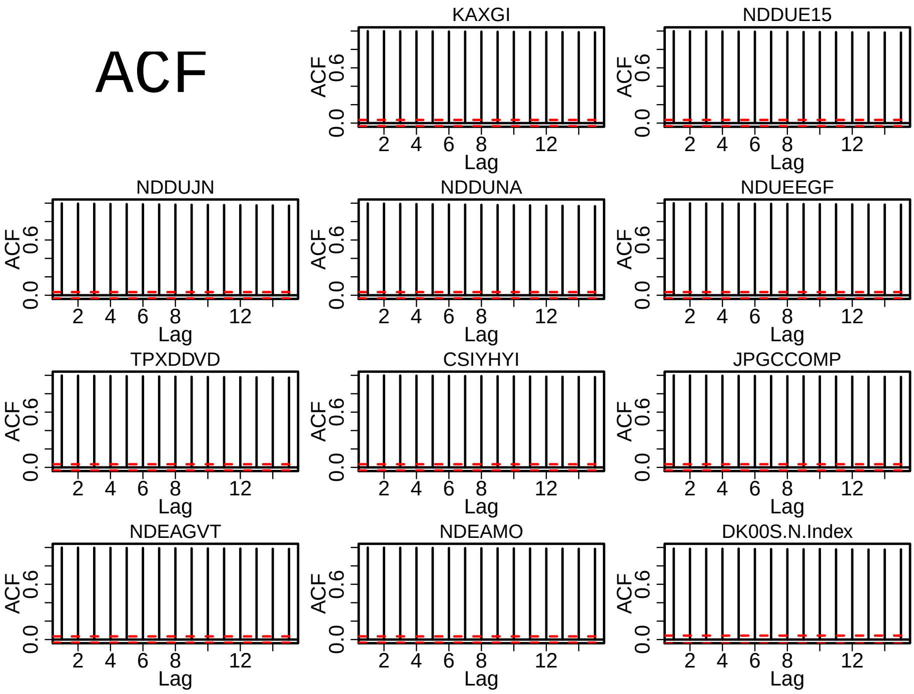


Figure 3.3: ACF in data with 95 % confidence interval (red).

In figure 3.3 and 3.4 the autocorrelation (ACF) and partial autocorrelation (PACF) are plotted for each index series together with a 95% confidence interval. The interval is calculated by $\pm 1.96/\sqrt{n} = \pm 0.0341$, where n is the sample size and ± 1.96 corresponds to the 2.5% and 97.5% quantile in the standard normal distribution. Using this confidence interval assumes data following a multivariate normal distribution. It has not been shown that the series are normal distributed, but the confidence interval is still used with this observation in mind. All the plotted lags in the ACF-plot and lag=1 in PACF-plot are highly significant. As expected the index value of today depends highly on yesterday's value. By looking at figure 3.4 some of the stock indices are also significant at lag 2 and 3. The reason for this can be explained by the volatile behaviour. Stock markets often have longer periods with smaller volatility, followed by shorter periods with high volatility, also known as volatility clumping. The reason why lags larger than 1 for the bond indices are absolutely not significant, is the more stable behaviour, where volatility clumping is more unusual. The rate index shows a bit strange tendency in the PACF-plot having significant lags at lag 2, 3, 4, 5, 10 and 15. This might be caused by some special mechanism or trading behaviour in the market, but there is no reason to deal with that now, the different behaviour might vanish in the modelling process.

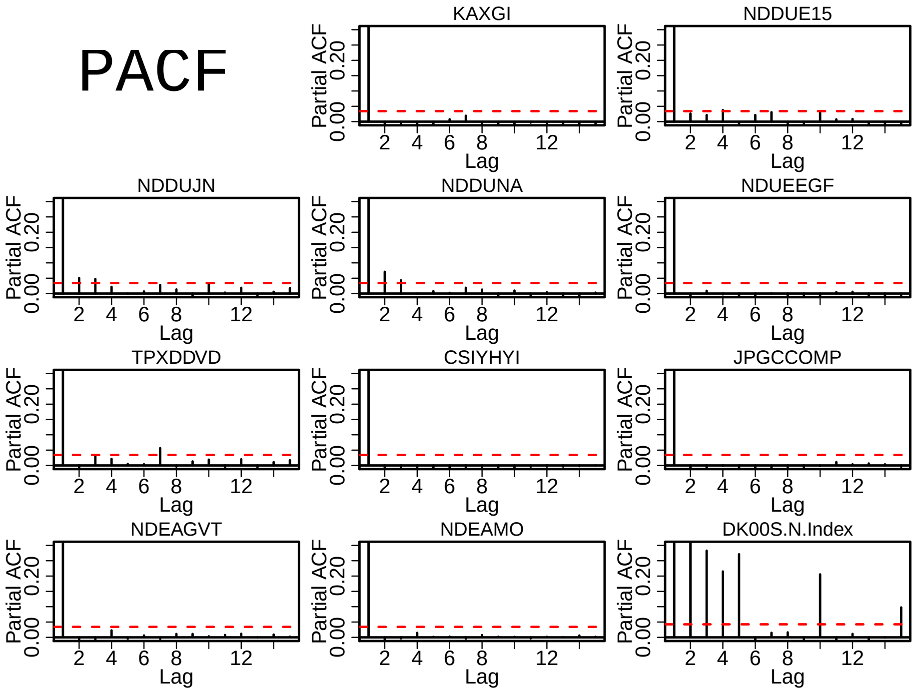


Figure 3.4: PACF in data with 95 % confidence interval (red).

3.4 Normality and stationarity

Modelling data can be done easily if we know the true distribution, mean and the variance. Looking at the index series on figure 3.1, it seems hard to use a direct estimate of the mean and the variance for model that would be acceptable, because the series do not look stationary.

It would be comfortable if the data follows a normal distribution, because many statistical test and assumptions are based on data being normal. The Shapiro–Wilk test [20] tests the null hypothesis that the index values comes from a normal distribution. The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} \alpha_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - \mu)},$$

where $X_{(i)}$ is the order statistics of the index values, X. $\mu = \frac{\sum_{i=1}^{n} X_n}{n}$ is the index mean and n is the number of values. α_i are constants generated from means, variances and covariances of order statistics of n independent and identical distrubuted (i.i.d.) random variables sampled from a normal distribution. Of course all the series are tested separately using the Shapiro–Wilk test. All the results give p-value $< 2.2 \times 10^{-16}$, and thereby rejecting the null hypothesis of normality as expected.

Knowing that data is highly autocorrelated, we might expect that data is not stationary. All the series are tested for stationarity using the Kwiatkowski-Phillips-Schmidt-Shin test (KPSS-test) [34]. All the tests give p-value < 0.01, meaning that the null hypothesis of level-stationarity can be rejected and mean and variance cannot be estimated easily.

Knowing that the indices might be non-stationary, a test for data following a random walk is relevant. A random walk is a unit root non-stationary process and is defined as:

$$X_i = X_{i-1} + a_t$$

where a_t is a white noise process. The Augmented Dickey–Fuller-test tests if data has a unit root [33]. The null hypothesis is that data has a unit root, and testing all the series give large p-values and the null hypothesis cannot be rejected in any cases. The indices might therefore follow a random walk, so further analysis is necessary.

Index prices have shown properties that makes the modelling difficult. Therefore it is appropriate to transform data, which is the topic for the next chapter.

Chapter 4

Analysis of returns

In the previous chapter the index prices were analysed, but from an investors perspective the price of an asset or index is not as relevant as the return. The aim is to gain profit or hedge when investing, and the index price is not a directly measure of how well that is done. Instead the return is a scale-free measure of the investment. Furthermore the index prices have shown statistical properties that make the modelling difficult. It is desirable that the data is stationary without any autocorrelation and if possible normal distributed. For this reason the returns of the indices are analysed trying to meet these qualities.

4.1 Calculating returns

There are different kinds of returns [29], and only returns based on the same period length and calculation method, returns can be compared.

4.1.1 Simple return

First let us consider one-period returns where the period is equal to one day, but it might as well be an hour, a week etc. The simple net return, R_t from yesterday, $T = t - 1$, to today $T = t$, is given by :

$$R_t = \frac{P_t}{P_{t-1}} - 1 = \frac{P_t - P_{t-1}}{P_{t-1}}. \tag{4.1}$$

Where P_{t-1} and P_t are the (closing) price of yesterday and today. $R_t + 1$ is also known as simple gross return. Now consider a multiple-period return, where for instance one period still is one day, and we want to know the net return of the last three days equal to $k = 3$ periods, then one period gross returns are simply multiplied :

$$\begin{aligned}
R_t[k] &= \frac{P_t}{P_{t-k}} - 1 && (4.2)\\
&= \left(\frac{P_t}{P_{t-1}} \cdot \frac{P_t}{P_{t-2}} \quad \cdots \quad \cdot \frac{P_t}{P_{t-k}}\right) - 1 && (4.3)\\
&= (R_t + 1)\,(R_{t-1} + 1) \cdots (R_{t-k+1} + 1) - 1 && (4.4)\\
&= \prod_{i=0}^{k-1} (R_{t-i} + 1) - 1. && (4.5)
\end{aligned}$$

4.1.2 Log return

Log return is actually the natural logarithm of the simple gross return:

$$LogR_t = \ln(R_t + 1) = \ln\left(\frac{P_t}{P_{t-1}}\right). \tag{4.6}$$

The log return is also called continuously compounded return. Log transformation of returns has different advantages. Extreme values in a set of returns will be reduced, and finding a model that fits the returns is now easier. A multi-period log return is simply the sum of all the one-period log return:

$$\begin{aligned}
LogR_t[k] &= \ln(R_t[k]+1) && (4.7)\\
&= \ln\left[(R_t+1)(R_{t-1}+1)\cdots(R_{t-k+1}+1)\right] && (4.8)\\
&= \ln(R_t+1)+\ln(R_{t-1}+1)+\cdots+\ln(R_{t-k+1}+1) && (4.9)\\
&= \sum_{i=0}^{k-1} LogR_{t-i}. && (4.10)
\end{aligned}$$

Equation 4.6 is used to transform the data into a log return space. Figure 4.1 is plot of log returns. Log return data seems to be more stationary, but it was also expected because of the transformation which also is the same as using a backward difference operator on ln-data:

$$LogR_t = \ln\left(\frac{P_t}{P_{t-1}}\right) = \ln\left(P_t\right) - \ln\left(P_{t-1}\right) = \nabla \ln(P_t).$$

Calculating the difference removes the autocorrelation at lag $= 1$, that we already have seen was highly significant.

It is even clearer that the volatility is not constant, because of the high fluctuation. Again the stock indices have a more fluctuating behaviour than the bond indices, again indicating higher sensitivity to variation in the market. The volatility clumping has also been more distinct, especially in the beginning of crisis starting in 2008 is easy to see. The rate index has a moderate volatility, but has some enormous outliers ultimo 2010. This is not caused by unrealistic changes in the index value, but the huge fluctuation is caused by relatively large daily changes compared to the low level of interest rate, which also can be seen of figure 3.1. There are other conspicuous log returns for the other series, and some of them can be explained. CSIYHYI has an outlier in 2001, and taking in to account that it mainly consists of American corporate bonds, it might have a relation to the terror attack 11 September The Japanese NDDUJN and TPXDDVD indices haves outliers around 11 March 2011 where an earthquake and tsunami hit Japan causing a tense and nervous market. The rest of the outliers will also be kept in the data set, because they are unacceptable extreme, it is not possible to reject that they are not true values and they might vanish when modelling. Taking a closer look at NDUEEGF index, the generated data seem to behave close to the rest of the series, and therefore the generated values are still accepted.

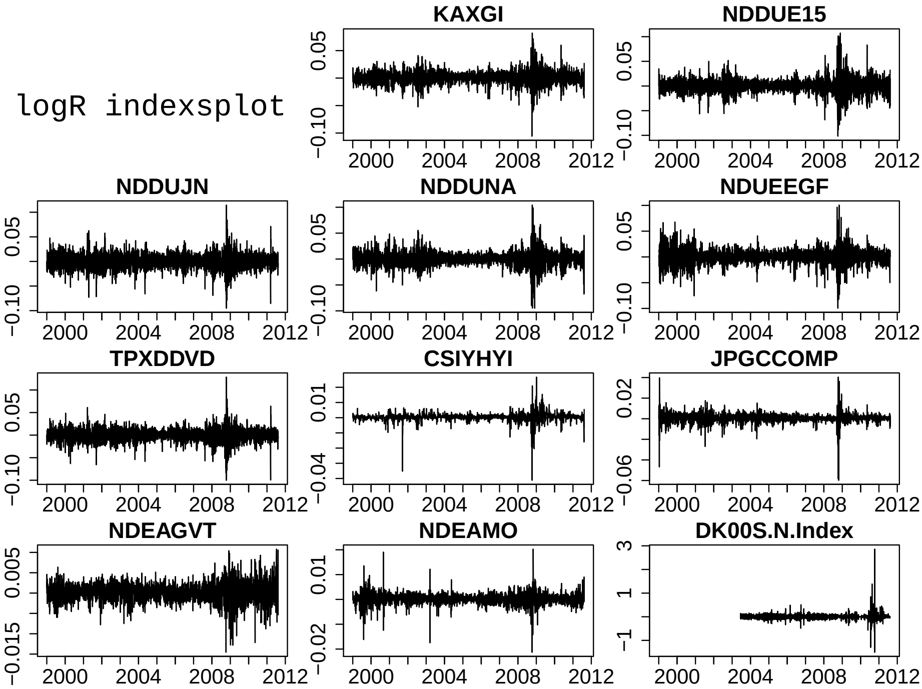
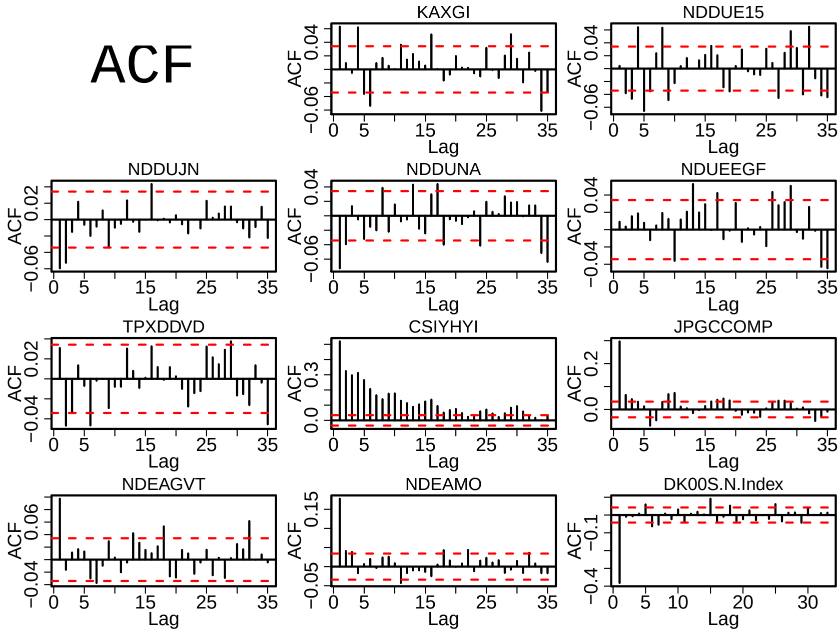


Figure 4.1: Plot of log return, where the time is on first axis and the log return on the second axis.

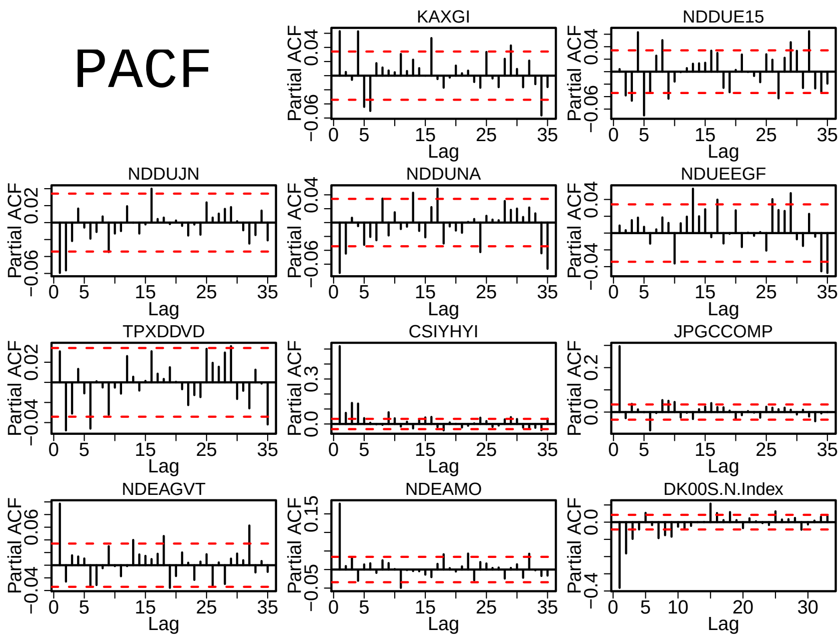
4.2 Autocorrelation in log return indices

After data has been transformed it would be interesting to see if there is any autocorrelation left. If it is possible to remove some time dependency in data the modelling process gets simpler.

Figure 4.2a is a plot of the autocorrelation in daily log return data. Comparing this with figure 3.3 it is easy to see that many significant lags has been removed through the transformation. Some has even switched to being negative. Comparing figure 4.2b with partial autocorrelation in daily log return data to figure 3.4, it is easy to see that the transformation has removed a lot of significance at lag $= 1$. But there is still a lot of autocorrelation left in data after transformation that cannot be ignored, especially CSIYHYI and DK00S.N:Index have many significant lags of lower order that certainly not can be assumed to be white noise.



(a) ACF in daily log return data with 95 % confidence interval (red).



(b) PACF in daily log return data with 95 % confidence interval (red).

Figure 4.2

A way to deal with autocorrelation in data is to use weekly data instead. Using weekly data, we might lose some extreme events, but using e.g. data from Friday every week the variance is kept realistic. If the mean value for the week is used instead the true variance is reduced resulting in a weak model. The few extreme events that are not in weekly data would anyway have vanished on the long run when modelling and generating scenarios. Therefore the use of weekly (Friday) data is acceptable, and is a technique already widely used in statistical finance exactly to get independent data. Using weekly data, the estimate of weekly volatility is more accurate.

If the Shapiro-Wilk test is applied on the weekly log return indices the result is that all p-value>0.1, and thereby the null hypothesis of level-stationarity cannot be rejected. Another way to check if weekly log return indices are stationary is to estimate their mean recursively. The recursive estimation has been done using a forgetting factor $\lambda = 0.9$ such that the recursive estimate at time t, becomes a weighting of the previous $t-1$ observations. The weighting of the i'th observation is given by :

$$W(i) = \lambda^{-(i-t)},$$

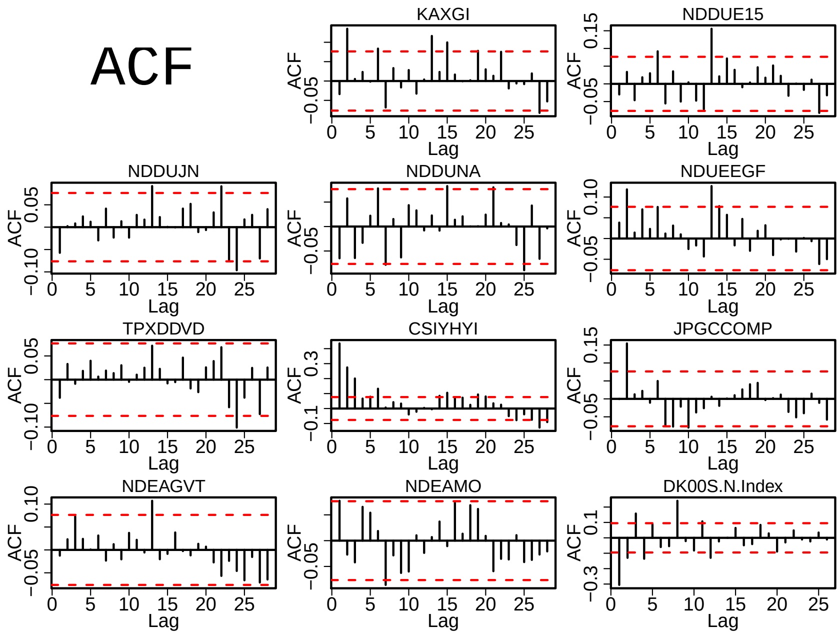
where $i \in [1;t]$. Afterwards, the weighting is scaled such that $\sum_{i=1}^{t} W(i) = 1$. In practice the effective number of previous values used in the estimation is given by:

$$n_{eff} = \frac{1}{1-\lambda} = \frac{1}{1-0.90} = 10.$$

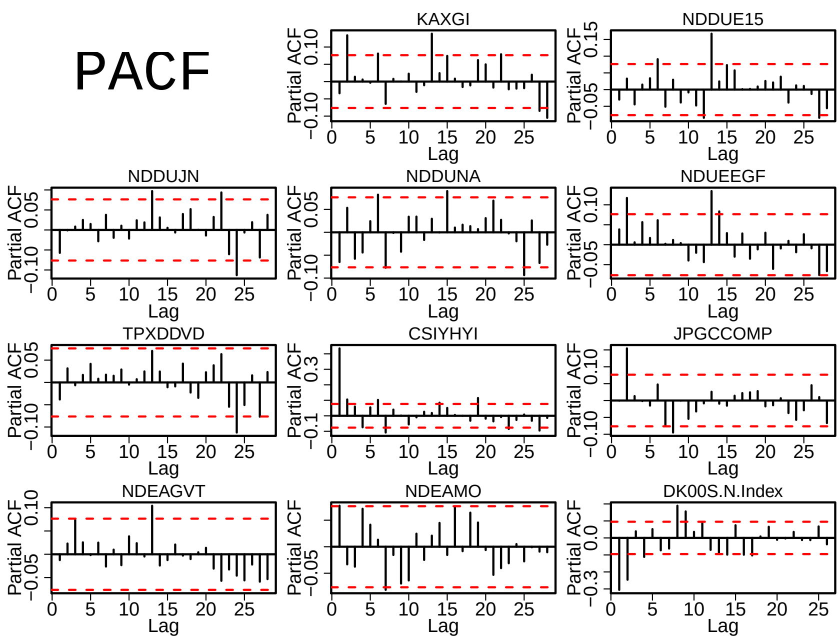
In figure 4.4 the recursive estimate of the mean for each weekly log return series is plotted. It is clearly seen that the mean has small fluctuations around zero (except DK00S.N.Index), therefore the weekly log return indices might be stationary.

This was already expected cf. earlier results and thereby the plots of ACF and PACF show a more exact picture of what is going on and not disturbed by time dependency. Stationarity is a nice property when we want to model the data, because a lot of different models require that the input must be stationary. The ACF and PACF for weekly log returns are plotted in figure 4.3a and 4.3b.

As expected even more significant autocorrelation have been removed, now to an acceptable level. The bond indices except CSIYHYI have only one or two lags just outside the 95 % confidence bands in the ACF, which acceptable. The



(a) ACF in weekly log return data with 95 % confidence interval (red).



(b) PACF in weekly log return data with 95 % confidence interval (red).

Figure 4.3

stock indices have a few more lags just outside the confidence bands but this is still acceptable. CSIYHYI still has some pattern in autocorrelation with lag = 1, 2 and 3 being very significant and lag 1 significant in the partial autocorrelation. The other bond and stock indices also have a few significant lags in PACF, but it is acceptable on a 95 % significance level even though it is a bit suspiciously that almost all the stock indices have significance lag around lag =13. There is no trading or market related explanation for this structure and as long as there only is a few lags of higher order just outside the confidence bands then data is accepted as being independent. The ACF and PACF in DK00S.N.Index now behave more like the other indices but there still seems to be too much time dependency left.

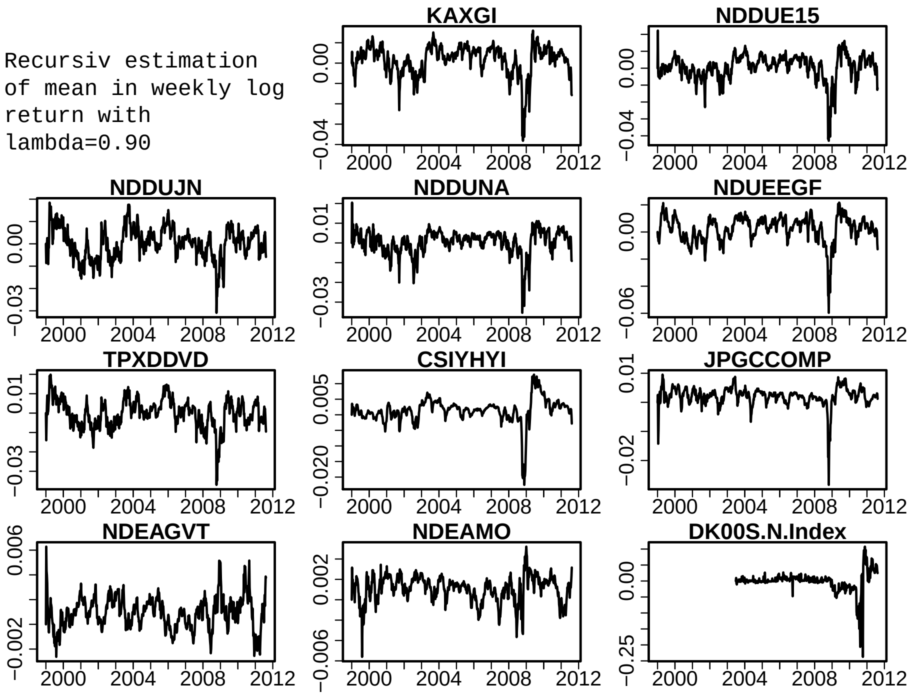


Figure 4.4: Recursive estimate of mean of each weekly log return index using forgetting factor $\lambda = 0.90$. Time is on the first axis, and mean on the secondary axis.

The reason for the strange behaviour of CSIYHYI might be that the log transformation is too "effective". Therefore a square root of simple gross return might be a usable transformation for exactly this index. The ACF and PACF for the square root simple gross CSIYHYI index is plotted in Appendix A. There is only a slightly difference compared to the log return data, and therefore the log return transformation will be used for now.

Now it can be assumed, a little roughly, that weekly log returns are independent, with the exception of CSIYHYI and DK00S.N.Index. This is an important feature that is very useful when the indices are modelled in chapter 7.

4.3 Volatility

The plots of ACF and PACF in daily and weekly log return data has already shown indications of non-constant variance, also called heteroscedasticity. This can also be tested if a recursive estimate of the standard deviation is made. Again a forgetting factor $\lambda = 0.9$ is used and the estimates are plotted in 4.5. As it is clearly seen the standard deviations, for the indices are not constant despite the smoothing. The conclusion is that weekly log return indices have heteroscedastic behaviour.

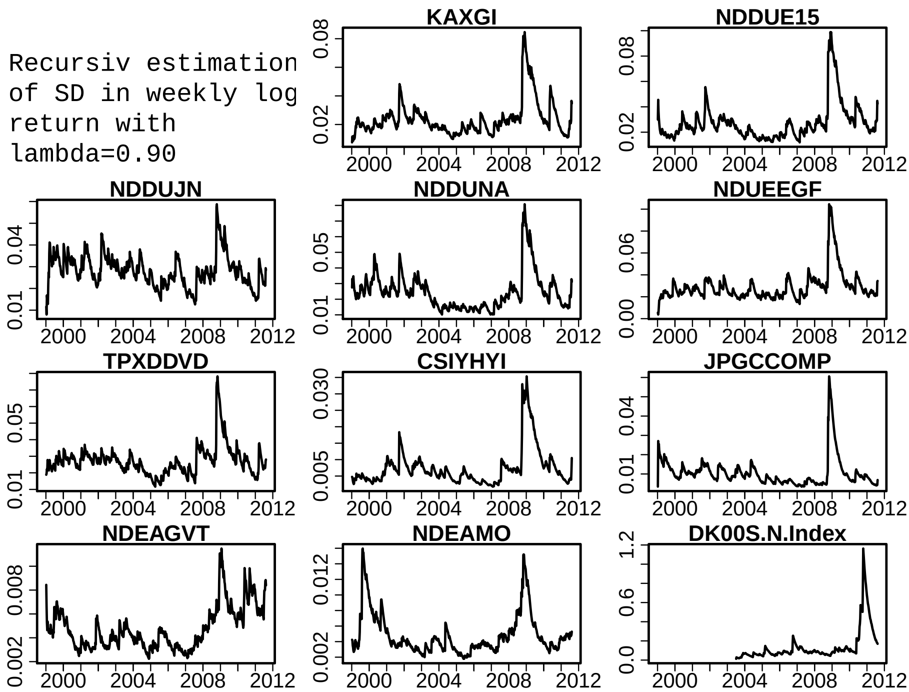


Figure 4.5: Recursive estimate of the standard deviation of each weekly log return indices using forgetting factor $\lambda = 0.90$. Time is on the first axis and the standard deviation on the secondary axis.

The log return plots of the indices on figure 4.1 show that the variance evolve continuously with a few and rare jumps often gathered in a cluster, known as volatility clumping or clustering. On figure 4.5 the volatility clumping of the stock indices tend to be more intense that the bond indices. This volatility behaviour indicates that it is conditional. E.g. if the market was volatile yesterday, the market tends to be volatile again today. The variance does not diverge to infinity, but are within a limited set, meaning that the variance is stationary. Using figure 3.1 to identify periods with positive or negative price changes and then comparing with figure 4.1, it is easy to see that the variance is higher in periods with falling prices than periods with growing prices.

4.4 Cross correlation in log return indices

From the index plot in figure 3.1 it seems plausible that all indices are somehow cross correlated, which also is known from theory and from other analyses. A boom or a crisis in the world economy will of course affect the financial markets somehow. Years ago bonds and stocks were negatively correlated, but in the recent years the inflation rate and the interest rate have drop so much that they are in general positively correlated. Cross correlation is an expression of how much two stochastic processes are correlated, and can be described by the cross correlation function (CCF). The CCF can be used to identify lags of one index that has some determining property on another index at lag 0. The coefficients in the CCF of two processes X and Y is defined as:

$$
\begin{aligned}
\rho_{XY}(t_1,t_2) &= \frac{\gamma_{XY}(t_1,t_2)}{\sqrt{\sigma_X^2(t_1)\sigma_Y^2(t_2)}} \\
&= \frac{Cov\left[X(t_1),Y(t_2)\right]}{\sqrt{\sigma_X^2(t_1)\sigma_Y^2(t_2)}} \\
&= \frac{E\left[\left(X(t_1)-\overline{X}(t_1)\right)\left(Y(t_2)-\overline{Y}(t_2)\right)\right]}{\sqrt{\sigma_X^2(t_1)\sigma_Y^2(t_2)}}
\end{aligned}
$$

where $\gamma_{XY}(t_1,t_2)$ is called the cross covariance, $\overline{X}(t)$ and $\overline{Y}(t)$ are the means and $\sigma_X^2(t)$ and $\sigma_Y^2(t)$ are the variances, all until time t. If the bivariate process $(X(t),Y(t))^{\mathrm{T}}$ is stationary the coefficient at lag τ is defined as:

$$
\begin{aligned}
\rho_{XY}(\tau) &= \frac{\gamma_{XY}(\tau)}{\sqrt{\gamma_{XX}(0)\gamma_{XY}(0)}} \\
&= \frac{\gamma_{XY}(\tau)}{\sigma_X\sigma_Y}
\end{aligned}
$$

where the cross covariance function is:

$$\begin{aligned}\gamma_{XY}(\tau) &= Cov\left[X(t), Y(t+\tau)\right] \\ &= E\left[\left(X(t)-\overline{X}(t)\right)\left(Y(t+\tau)-\overline{Y}(t+\tau)\right)\right].\end{aligned}$$

At lag $=0$ the cross-correlation is the correlation.

In figure 4.6 and 4.7 the cross correlation functions for lags from -4 to 4 is plotted together with pairs plots. As already stated, the lag=0 is in many cases significant. Among the bonds the correlation in time has almost vanished because of the splitting into weekly data. Only a few lags other than zero are significant. CSIYHYI has significant negative lags to the stock indices and JPGCCOMP, the reason for this is the autocorrelation left in the log return of CSIYHYI. The CCF for JPGCCOMP and the stock indices have significant lags at -2 and 0. NDEAGVT has negative correlation to the stock indices, and highly correlated to NDEAMO at lag=0. The CCF for DK00S.N.Index and the other indices have no significant lags at all.

Over all we see that the indices are correlated, but the use of weekly log return data removes the correlation in time. Otherwise it is often seen that the European and eastern markets are correlated to the American market with lag = -1 day because of the time differences. The reason why DK00S.N.Index is not correlated to the other indices can be explained by looking at figure **3.1**. The index prices rise because of the liquidity crisis, but it is adjusted by the government and the central bank, and is for this reason isolated. This is also confirmed by looking at the pairs plots where the correlation among the stock indices is represented as a linear pattern in the paris plot. This CCF plot represents the cross correlation for the whole data period, but if two different periods, one with growing prices and one with falling prices were analysed the result would not necessarily be the same. The correlations might change sign or event not be represented in some types of periods.

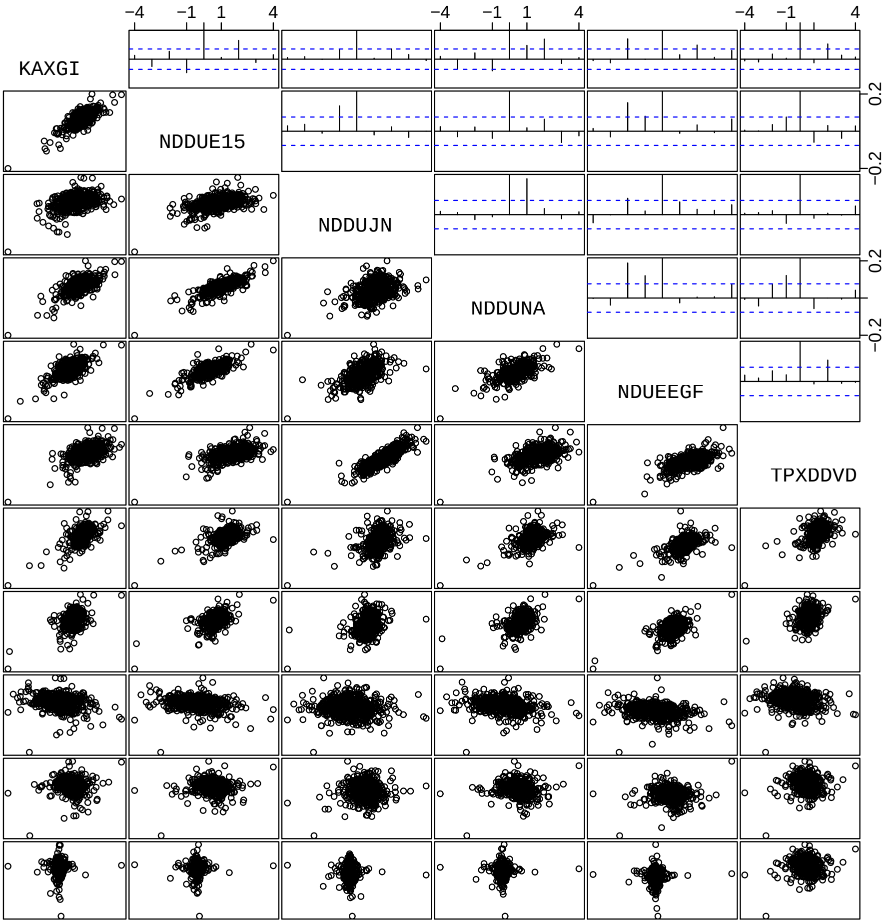


Figure 4.6: In the lower triangle paris plots are found and above the diagonal the CCF's are plotted. The blue line marks a 95% confidence interval.

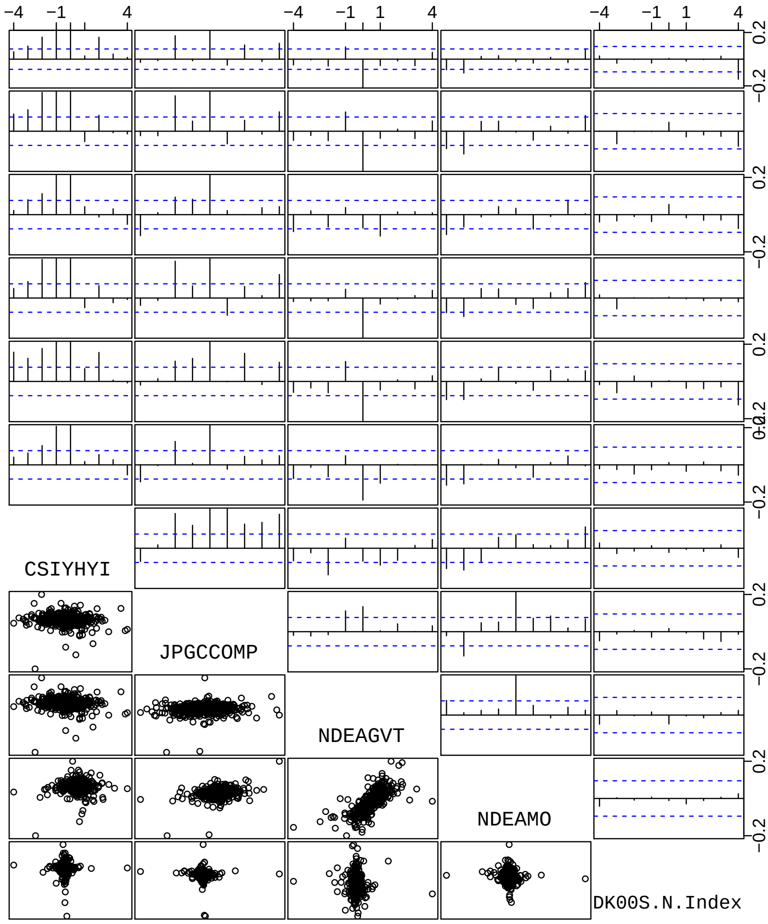


Figure 4.7: In the lower triangle paris plots are found and above the diagonal the CCF's are plotted. The blue line marks a 95% confidence interval.

4.5 Additional descriptive statistics of log return indices

Skewness and kurtosis are both indicators used to describe the shape of a distribution. Skewness measures the degree of asymmetry and is the relationship between the tails of the distribution, and if the skewness is positive the right tail is longer than the left, and opposite if the skewness is negative. Skewness is a rescaled third central moment (the mean is the first, and the variance is the second). The Fisher skewness is defined as [5]:

$$\varsigma = \frac{\mathrm{E}\left(X - \mathrm{E}\left[X\right]\right)^3}{\mathrm{Var}\left(X\right)^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x-\bar{x}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x-\bar{x}\right)^2\right)^{3/2}}. \tag{4.11}$$

The Pearson's skewness is the square of Fisher skewness.

Kurtosis is a measure of the degree of peakedness in the distribution or strictly speaking, a measure of unimodality (one major peak) versus bimodality (two major peaks) in the distribution [3]. If the kurtosis is negative the distribution is said to be platykurtic, which is flat topped for unimodal distributions, and if positive said to be leptokurtic which for a unimodal distribution is peaked. Kurtosis is a rescaled fourth central moment and the Fisher (excess) kurtosis is defined as [5]:

$$\kappa = \frac{\mathrm{E}\left(X - \mathrm{E}\left[X\right]\right)^4}{\mathrm{Var}\left(X\right)^{2}} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x-\bar{x}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x-\bar{x}\right)^2\right)^{2}} - 3. \tag{4.12}$$

Pearson's kurtosis is just the Fisher kurtosis plus three. The excess kurtosis of a normal distribution is equal to zero.

In table 4.1 descriptive statistics for daily and weekly log return is found. The estimated mean of both weekly and daily log return is close to zero, but still positive which indicates an overall positive tendency through the whole period. Only TPXDDVD and DK00S.N.Index has negative mean. The span of log return values for DK00S.N.Index is large compared to the others, and therefore the standard deviation is also larger. In general the standard deviation of weekly log returns is double in value compared to daily log returns. The indices have a small negative skewness, and the weekly log return tend to be a little more left-skewed, except DK00S.N.Index and weekly log return of NDEAGVT having positive skewness. DK00S.N.Index gets from being strongly right-skewed to less right-skewed turning from daily to weekly log returns, and NDEAGVT changes from left-skewed to right-skewed. All indices are leptokurtic, and there is no tendency of what happens in kurtosis when data changes form daily to weekly. JPGCCOMP, CSIYHYI and especially DK00S.N.Index have very large kurtosis compared to the other indices. The kurtosis of these three indices are reduced when using weekly data.

Index	Size	Min.	Max.	$\bar{\mu}$	$\bar{\sigma}$	Skew.	Kurt.
Daily log return							
KAXGI	3291	-0.1058	0.0820	2.664×10^{-4}	0.0114	-0.3863	6.4612
NDDUE15	3291	-0.1018	0.1074	9.569×10^{-5}	0.0142	-0.0760	7.0726
NDDUJN	3291	-0.0951	0.1147	4.414×10^{-5}	0.0149	-0.1450	3.9572
NDDUNA	3291	-0.0950	0.1043	5.094×10^{-5}	0.0132	-0.2320	7.6251
NDUEEGF	3291	-0.0996	0.1007	4.500×10^{-4}	0.0146	-0.3401	5.4146
TPXDDVD	3291	-0.1000	0.1286	-5.670×10^{-5}	0.0140	-0.3619	6.8168
CSIYHYI	3291	-0.0412	0.0267	2.679×10^{-4}	0.0027	-2.5366	40.445
JPGCCOMP	3291	-0.5977	0.0404	4.195×10^{-4}	0.0044	-1.6445	32.961
NDEAGVT	3291	-0.0146	0.0108	1.919×10^{-4}	0.0029	-0.2785	3.6039
NDEAMO	3291	-0.0213	0.0203	2.195×10^{-4}	0.0021	-0.3482	13.946
DK00S.N	2129	-1.5041	2.8622	-2.330×10^{-4}	0.1217	5.5523	170.45
Weekly log return							
KAXGI	659	-0.2107	0.1030	1.330×10^{-3}	0.0273	-1.4432	8.5646
NDDUE15	659	-0.2655	0.1392	4.779×10^{-4}	0.0312	-1.3347	10.278
NDDUJN	659	-0.1640	0.1102	2.204×10^{-4}	0.0292	-0.2528	1.8297
NDDUNA	659	-0.2053	0.1201	2.544×10^{-4}	0.0274	-0.8329	6.7085
NDUEEGF	659	-0.2252	0.1854	2.247×10^{-3}	0.0327	-0.8276	6.8369
TPXDDVD	659	-0.2202	0.0925	-2.832×10^{-4}	0.0288	-0.9112	5.1964
CSIYHYI	659	-0.1025	0.0551	1.338×10^{-3}	0.0100	-2.6222	26.353
JPGCCOMP	659	-0.1298	0.0969	2.095×10^{-3}	0.0213	-2.7719	35.731
NDEAGVT	659	-0.0183	0.0241	9.586×10^{-4}	0.0052	0.1286	1.8096
NDEAMO	659	-0.0394	0.0327	1.096×10^{-3}	0.0055	-0.6926	8.0980
DK00S.N	425	-2.1401	2.6027	-1.211×10^{-3}	0.2204	1.7671	67.627

Table 4.1: Descriptive statistics of daily and week log return indices consisting of data size, minimum, maximum, mean, standard deviation Fisher skewness and Fisher (excess) kurtosis.

In figure 4.8 density curves of weekly log return indices are plotted together with an approximated normal curve. The density at given point, x, for the normal curve is approximated using:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

In this plot it is possible to see, what already has been deduced, that the indices have a positive kurtosis and the main part of the indices are negative skewed. Some of the indices tend to follow a normal distribution more than others. E.g. NDDUJN and TPXDDVD, both Japanese indices, have density curves that are close to the normal curves. Looking at DK00S:N:Index it is very obvious that it behave differently, by the high frequency of very many small log returns. From these different tests, normality in weekly log return indices cannot be proved, though it might be assumed when modelling data.

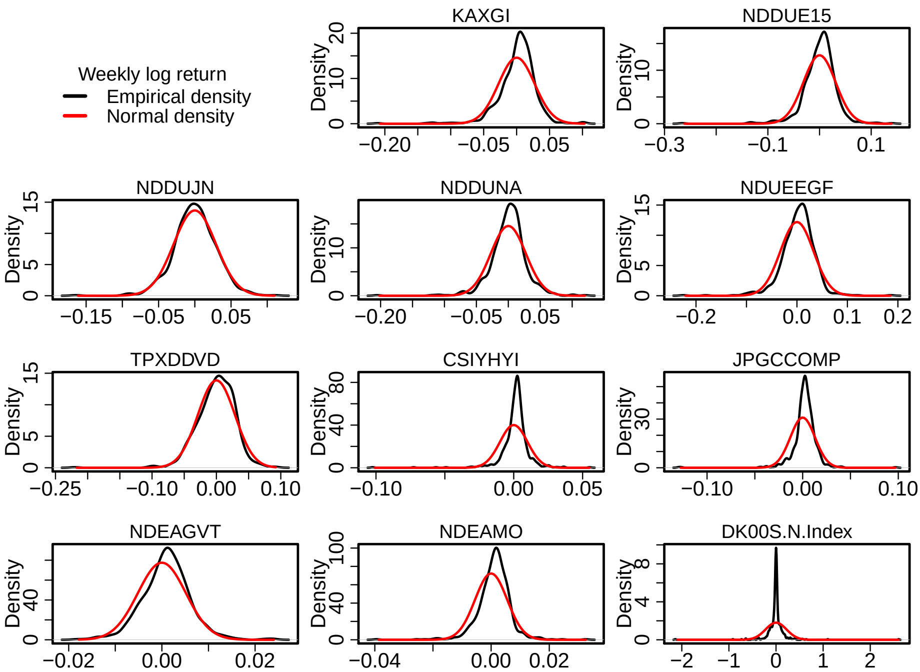


Figure 4.8: Density plot of weekly log return data (black) and an appertaining approximated normal curve (red).

To summarize the daily and weekly log return data has some interesting properties. Outliers are now much easier to identify, and some of them can be explained directly by external events. Using the log return transformation removes much of the one period dependency, and if weekly data is used the data tends to behave almost time independent, but there are still significant lags in the autocorrelation function that needs to be modelled. Therefore the remaining auto correlation is accepted. Test results of stationarity shows that it cannot be rejected that weekly log return indices are stationary. Analysis of the volatility shows that the log return indices are conditional heteroscedastic, and comparing variance in positive and negative periods shows a difference. Plots of the cross correlation functions shows that the indices are correlated, but not in time. The stock indices are highly correlated with each other and more or less also to the bond indices. Using weekly log returns removes the lagged correlation, but the direct correlation is maintained. It is very clear from the analysis that the DK00S.N.Index do not behave like the other indices. The reason for this might be that the rate is to some extent controlled political decisions.

Chapter 5

Financial regimes

It has been demonstrated that the behaviour of the variance of the indices depends on whether the index prices has a positive trend or negative trend. It was also discovered that periods or financial regimes not only are distinguished by positive or negative trend, but also by length, volatility and slope. The indices used here are mainly represented by developed countries, and therefore an indicator for the turning points should reflect the economic conditions in these countries. Therefore OECD's turning points total area [25],[24] are used. The turning points are determined on the basis of OECD Composite Leading Indicators (CLI). CLI are calculated for 29 OECD member countries [1] , 6 non-members and 7 country groupings such as Euro area or G7 [2]. The CLI total area only covers the 29 OECD member countries with weightings corresponding to economic size the year before. E.g. in 2009 Denmark was weighted 0.5%, United States 36.38% and Japan 10.61% in the total area, where the weighting is calculated upon the 2008 gross domestic product (GDP) based on purchasing-power-parity (PPP) valuation of country GDP, in billions of current international dollar. The CLI's are indicators build on business cycles and turning points. These are identified by measuring the deviation from trend series, where Index of Industrial

[1]Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Japan, Korea, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Slovak Republic, Spain, Sweden, Switzerland, Turkey, United Kingdom, and United States[23].

[2]Group of seven major industrialized nations.

Production (IIE) is an often used trend series because of the availability and cyclical sensitivity. GDP is also used to identify turning points. The CLI's are sensitive to exceptions and respond quickly to changes in economic activity.

The turning points within the time span for the data are given in table 5.1.

Date	1999M1	2000M8	2001M12	2008M2	2009M2	2011M3
Event	Trough	Peak	Trough	Peak	Trough	Peak

Table 5.1: OECD's reference turning points, total area. The date of an event is given by the year and (end) month number.

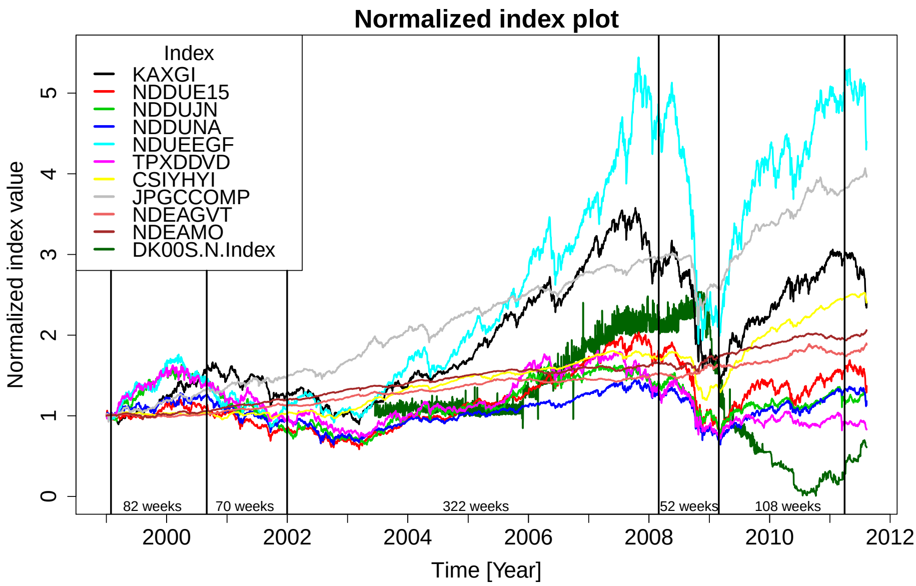


Figure 5.1: Normalized index plot with turning points, and numbers of weeks between them.

In figure 5.1 the turnings points are illustrated together with normalized index plots. The turning points do not lay spot on the dates that the indices indicate, though it is hard to tell because not all the indices are in maximum or minimum at the same time. Over all they seem to estimate reasonably, also when thinking a step further and taking the modelling process into account. In this plot it is again observed that the periods are very different. There is no fixed length of a period. In the time spans studied here, the three periods with the rising prices are 78, 322 and 108 weeks, and the two periods with falling prices are 74 and 52 weeks. The slope and the volatility do also seem to be different. E.g. looking at the periods with falling prices, the first period (mid 2000 - start 2002) does not have such an intense fall and volatility as the second period (start 2008 - start 2009). Therefore the modelling of the indices will be done for each financial regime to obtain a model that fits data the best. Consequently the first (January 1999) and last (April 2011- August 2011) data samples will be left out in the modelling process.

Chapter 6

Theoretical background

Statistical modelling of financial data is the main topic in this thesis. In the previous chapters the indices have been analysed in order to ease the modelling process and support the models and approach used. Only the theory behind the models and approaches used in this thesis will be described, and basic statistics will not be explained. Therefore a certain level of academic knowledge is presumed in order to get full benefit of the rest of this thesis, though it is still possible to understand the outline without any prerequisites.

In this chapter a method for modelling the correlation structure of the indices is presented. Afterwards, the focus turns to the dynamic structure of the indices and their variance and a suitable time series model will be described.

6.1 Principal Component Analysis

Principal component analysis (PCA) is a method used to reduce the numbers of variable in data that clearly are correlated in order to simplify further data modelling. PCA is based on finding the factors that determine the patterns, the correlation or covariance, in data. These independent factors, known as principal components (PC's) are described by a weighted linear combination of explanatory variables. The goal is to replace the variable with a fewer numbers of factors and still describe data without losing to much information of the variation. The principal components are ordered such that PC describing the variation in data the best is listed first. In this section subject relevant to a principal component analysis will be outlined together with a "manual" for deriving PC's. PCA can be done using either the covariance matrix $\mathbf{\Sigma}$ or the correlation matrix $\boldsymbol{\rho}$, some of the advantages, disadvantages and differences will be described.

6.1.1 Assumption underlying PCA

PCA is derived on the basis of different geometrical and statistical assumptions that the data has to fulfil in order to use PCA. Often data sampled from the real world does not fulfil these assumption completely, but that does not mean that PCA cannot be used, just that we have to be careful when concluting. The higher degree of fulfilment, the higher degree of credibility to the conclusions. If PCA is used to describe patterns in covariance/correlation, where hypothesis test also can be used, the degree of fulfilment is not supposed to be as high as if PCA is used in statistical inference. There are three main assumptions [11]:

Multivariate normality
Multivariate normality is hard to prove, but univariate normality of each of the variables is a good indication. Often data is mean adjusted before PCA is applied in order to ensure data being centered. Other method, such as log return is used on data with a more tricky behaviour in order to meet this assumption. Weakly stationarity is a sub-assumption to univariate normality and implies that the mean, variance and autocovariances of the indices are invariant in time. Data has to be weakly stationary, because the covariance matrix and the correlation matrix are unknown, but the sample covariance matrix and correlation matrix can be estimated correct if data is weakly stationary. Assuming data is weekly log return $\{\mathbf{X}_t | t = 1, ...T\}$ the estimates becomes:

$$\widehat{\mathbf{\Sigma}} \equiv [\hat{\sigma}_{i,j}] = \frac{1}{T-1} \sum_{t=1}^{T} \left(X(t) - \overline{X}\right) \left(X(t) - \overline{X}\right)^{\mathrm{T}} \tag{6.1}$$

$$\widehat{\boldsymbol{\rho}} = \widehat{\boldsymbol{D}}^{-1} \widehat{\mathbf{\Sigma}} \widehat{\boldsymbol{D}}^{-1} \tag{6.2}$$

where $\overline{X} = \frac{1}{T} \sum_{t=1}^{T} X(t)$ is the sample mean, and $\widehat{\boldsymbol{D}} = \mathrm{diag}\left\{\sqrt{\hat{\sigma}_{11}}, ..., \sqrt{\hat{\sigma}_{pp}}\right\}$ is a diagonal matrix of standard deviations of the sample. The higher degree of multivariate normality, the higher degree of explanation is achieved in the first principal components and less redundancy is experienced in the later principal components.

Linearity
PCA assumes that the variables have a linear relationship because the principal components are derived on a linear combination of eigenvectors of the variance or covariance matrix and the variables. If the variables possess non-linearity, here related to ARCH-effects, the attempt to describe the variables in a linear relation is hard. The result is often failure or that PCA does not reduce the dimension of data significantly.

Independent random observations
In order to describe the true distribution, the observation has to be drawn and independently. This assumption also ensures that the estimated correlation and covariance is true. This leads to the question of how outliers should be treated because independent random sampling includes outliers. True outliers can be eliminated, but it is important to distinguish between true outliers and extreme values. Extreme values are important in order to describe the distribution. If the sample size is large the influence of outliers vanishes.

6.1.2 Adequate sample size

In order to PCA to work properly, a minimum adequate numbers of object from a sample should be available. The sample size depend on the homogeneity in data, the less inhomogeneity the less the sample size has to be, and therefore there is only rules of thumb for the sample size. Of course the number of object should be at least the number of variable. The variance of the sample should be a good estimate of the true variance of the population, or here the index. In this thesis more than 3000 daily and 659 weekly index values representing each indices (2130 and 426 values for DK00S.N.Index), and if PCA is used in regime separated data, the sample size is still far large enough.

6.1.3 Derivation of principal components

Principal component can be derived either by using the covariance matrix $\mathbf{\Sigma}$ or the correlation matrix $\boldsymbol{\rho}$[11, 9]. First the use of $\mathbf{\Sigma}$ is presented.

Let $\mathbf{x}$ be vector of $p = 11$ variables (indices). The aim is to determine $m < p$ variables, principal components (PC's), that keep as much information on the covariance between the p variables as possible. The k'th PC, $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, ..., \beta_{kp})^{\mathbf{T}}$ is a vector of p elements. The k'th PC is determined by linear combination $\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{x}$, having maximum variance where $\boldsymbol{\beta}_k\mathbf{x}$ has to be uncorrelated with the other $k-1$ earlier determined combinations. Roughly said the first PC is the linear description of data that has maximum variance. The second PC is also a linear description of data with maximum variance under the constraint that it is orthogonal to the first PC. This continues until p'th PC is determined and all variance in $\mathbf{x}$ are explained.

But how is the PC's determined practice? To answer this question the covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{x}$ is used because the relationship that maximizes the variance of $\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{x}$ has to be found. The variance is given by:

$$\mathrm{Var}\left(\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{x},\right) = \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}_k. \tag{6.3}$$

This relationship is only possible to maximize if the principal component is normalized such that it is a unit vector:

$$\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\beta}_k = 1. \tag{6.4}$$

The task is to maximize the right side equation 6.3 subject to6.4. Lagrange multipliers are used for this purpose:

$$\text{Maximize}: \quad \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}_k - \lambda\left(\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\beta}_k - 1\right).$$

Differentiating with respect to $\boldsymbol{\beta}_k$

$$\begin{array}{rcl} \boldsymbol{\Sigma}\boldsymbol{\beta}_k - \lambda\boldsymbol{\beta}_k &=& 0 \\ & \Updownarrow & \\ \left(\boldsymbol{\Sigma} - \lambda\mathbf{I}\right)\boldsymbol{\beta}_k &=& 0, \end{array}$$

where λ now can be seen as an eigenvalue with corresponding eigenvector, $\boldsymbol{\beta}_k$, to the covariance matrix. The k'th largest eigenvector determines the k'th PC because the maximizing reduces to:

$$
\begin{aligned}
\operatorname{Var}\left(\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{x}\right) &= \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}_k \\
&= \boldsymbol{\beta}_k^{\mathrm{T}}\lambda\boldsymbol{\beta}_k \\
&= \lambda\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\beta}_k \\
&= \lambda.
\end{aligned}
$$

The eigenvalues must be sorted by size with the larges values first. Then the k'th PC is the k'th eigenvector. To summarize, let $\mathbf{z}$ be a vector of PC's, then:

$$\mathbf{z} = \mathbf{A}^{\mathrm{T}}\mathbf{x}$$

where $\mathbf{x}$ is a vector of the original variables and $\mathbf{A}$ is the orthogonal matrix with $\boldsymbol{\beta}_k$ in the k'th column, where $\boldsymbol{\beta}_k$ is the k'th ordered eigenvector of $\boldsymbol{\Sigma}$.

The condition of independence between the PC's is fulfilled because $\boldsymbol{\Sigma}$ is symmetric and it is an orthonormal linear transformation of $\mathbf{x}$.

If the principal components instead are derived by using the correlation matrix $\boldsymbol{\rho}$ the principal components are defined as:

$$\mathbf{z} = \mathbf{A}^{\mathrm{T}}\mathbf{x}^*$$

where $\mathbf{A}$ consists of $\boldsymbol{\rho}$'s eigenvectors, ordered in accordance with eigenvalue ordered by size. $\mathbf{x}^*$ is a vector of standardized variables, where the i'th element is calculated as $x_i^* = \frac{x_i}{\sqrt{\sigma_{ii}}}$.

The transformation of $\mathbf{x}$ to $\mathbf{x}^*$ is not an orthogonal transformation, and therefor do the principal components derived by the two methods provide the different information. Consequently the method using the covariance matrices is preferred in some cases instead of correlation matrix. When the variables are of different type the correlation method is preferred, because the element in $\mathbf{x}^*$ is dimensionless and can without any problem be combined to PC scores (the projections of data with respect to the PC). The covariance method would not be the right method to use because the variables are not directly comparable. If the variables belong to the same type and all elements in $\mathbf{x}$ has the same unit, then covariance method is often preferred because the use of $\mathbf{x}^*$ is somewhat equivalent of making an arbitrary choice of measurement units. This argument is only usable when the units are equal, because the covariance of a case where the variables are of different type, choosing units is an even more arbitrary.

Using the correlation methods makes it more difficult to compare the PC's because their coefficients have been normalized differently.

6.1.4 Choosing appropriate number of principal components

How many principal components, m, is necessary to describe the variance in data of p variables properly, is an important question to answer. Many different methods and appertaining approaches exists. Generally the first p PC's with the highest degree of explanation is used, only in some special cases the last PC's are of more interest, but it will not be examined here. Two different methods for determining p [9] is presented here.

Cumulative Percentage of Total Variation
This method, as the name tells, uses the cumulated percentages of each PC's contribution to the total variance. Before doing PCA, an acceptable level for the cumulated percentage is set, and then the number of PC's describing at least that level is used. The level depends of the practical use of the PCA, but often it is within the range 70% to 90%. This method do not depend on whether the covariance or correlation matrix is chosen to compute the PC's.

The scree Graph
This method is a graphical method where the variance or eigenvalue of the k'th PC, l_k is plotted against k. This often forms a graph shaped like an arm, at the point, k, where the "elbow" is, or more mathematically described the points where graph begins a linear tendency, not necessarily parallel to the first axis. This method is not as subjective as the other method.

6.1.5 Limitation of PCA

PCA has many nice features but it also has some limitations. A few is mentioned below, and are all in relation to the derivation of the method.

- Correlation is assumed in order to reduce the dimension.

- The linear combination with the largest variance is assumed to be of most importance.
- The method is not scale invariant if the covariance matrix is used.
- PCA assumes the first two central moments to be time-invariant.
- Only orthogonal rotation of original variables is considered.

In relation to this project and data used it will be a challenge to meet the assumptions of PCA such that an acceptable level is maintained. The analysis in the previous chapters has shown heteroscedastic or ARCH effect which is a non-linear behaviour that is unwanted in PCA. In this project the idea of using PCA is to describe data and reduce the number of variables in order to ease and simplify the further modelling. For this purpose, as already mentioned, larger deviations are accepted in relation to the assumptions. In the next section time series models that takes heteroscedastic behaviour into account will be described in order to find models that can describe the dynamic structure of the volatility in the indices.

6.2 Non-linear time series models

PCA is not enough modelling of financial time series in order to describe the behaviour. A model should include all the information in the series and thereby describe the behaviour completely. A model is accepted when the residuals of the models can be assumed to be noise. For this reason the dynamic behaviour of the time series and the volatility needs to be modelled as well, where the aim is to find a relationship within the series that only leaves residuals out that can be accepted as white noise. Some particular behaviour of indices used in this project has been detected through the data analysis in the previous chapters. Only models taking this behaviour into account might be applicable, and only those used to model data in the next chapter will be described here. It is assumed that weekly log return indices are split into regimes order to have stationary and independent data. It has been shown that log return indices are characterized by :

- Heteroscedasticity and volatility clustering
- Data are white noise.
- Positive excess kurtosis and skewness close to zero

Therefore non-linear parametric discrete time series models taking these characteristic into account can very well be used. The models used in this project are variants of the autoregressive moving average model (ARMA) and therefore this will be introduced briefly.

6.2.1 ARMA (Linear model)

The autoregressive moving average model consists of an auto regressive process and a moving average process. It is defined as [7, 27] :

$$X_t = \alpha_0 + \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + \beta_1 \varepsilon_{t-1} ... \beta_q \varepsilon_{t-q} + \varepsilon_t, \qquad (6.5)$$

where $\{\varepsilon_t\}$ is white noise (i.i.d. and $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$), α_0 is a constant determing the level, α and β being parameters of the model, and p and q are the order of the AR and MA parts respectively. Introducing the autoregressive function $\alpha(\mathrm{L}) = 1 - \alpha_1 L - ... - \alpha_p L^p$ and the moving average function $\beta(\mathrm{L}) = 1 + \beta_1 L - ... - \beta_q L^q$ the model can be written as:

$$\alpha(\mathrm{L}) X_t = \nu + \beta(\mathrm{L}) \varepsilon_t. \qquad (6.6)$$

ARMA models do not require data do be stationary to fit, because they can both be stationary or non-stationary, which depends on the parameters in the model. But an ARMA model does not suit financial data of the type used here, because the heteroscedasticity is not taken into account. The following models that will be presented allows the variance to be conditional and non-constant.

6.2.2 ARCH

The autoregressive conditional heteroscedasticity model (ARCH) has the property of modelling variance. Variance in financial time series might be unconditional and constant at some time points in the series, and at other points the variance might be highly conditional and non-constant. This behaviour can be modelled with a model of the ARCH family. The idea behind the ARCH model is to have a mean adjusted return, X, that is uncorrelated in time but dependent, e.g. $\mathrm{E}[X_t | \mathcal{F}_{t-1}] = 0$, where $\mathcal{F}_{t-1}$ is a set of information at time $t-1$. X is sometimes called a shock. The ARCH model is defined as [7]:

$$\begin{aligned} X_t &= \sigma_t \omega_t & (6.7) \\ \sigma_t^2 &= \alpha_0 + \alpha_1 X_{t-1}^2 + ... + \alpha_p X_{t-p}^2 & (6.8) \\ &= \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2, & (6.9) \end{aligned}$$

with constraints on the parameters:

$$\begin{aligned} \alpha_0 &> 0 \\ \alpha_i &\geq 0 \;\; for\ i \in \{1; p\} \\ \sum_1^p \alpha_i &< 1. \end{aligned}$$

$\{\omega_t\}$ is i.i.d with $\omega_t \sim N(0, 1)$, α_0 is a constant variance drift, α_i is the parameter in the model of order p. The first two constraints ensures positive variances and the third one ensures that the unconditional variance of X_t is finite. Volatility clustering is a behaviour in data and if the ARCH model is inspected closer, it is discovered that if variance yesterday σ_{t-1}^2 was high, the probability of getting a high value again today is also high.

The properties of an ARCH(p) process is now studied closer. The unconditional mean X_t is:

$$\mathrm{E}(X_t) = \mathrm{E}\left[\mathrm{E}(X_t | \mathcal{F}_{t-1})\right] = \mathrm{E}\left[\sigma_t(\mathrm{E}(\omega_t)\right] = 0,$$

The unconditional variance of X_t is:

$$
\begin{aligned}
\mathrm{Var}(X_t) = \mathrm{E}\left[\sigma_t^2\right] &= \mathrm{E}(X_t^2) \\
&= \mathrm{E}\left[\mathrm{E}(X_t^2|\mathcal{F}_{t-1})\right] \\
&= \mathrm{E}\left[\alpha_0 + \alpha_1 X_{t-1}^2 + ... + \alpha_p X_{t-p}^2\right] \\
&= \alpha_0 + \alpha_1 \mathrm{E}\left[X_{t-1}^2\right] + ... + \alpha_p \mathrm{E}\left[X_{t-p}^2\right] \\
&= \alpha_0 + \alpha_1 \mathrm{E}\left[\sigma_{t-1}^2\right]\mathrm{E}\left[\omega_{t-1}^2\right] + ... + \alpha_p \mathrm{E}\left[\sigma_{t-p}^2\right]\mathrm{E}\left[\omega_{t-p}^2\right] \\
&= \alpha_0 + \alpha_1 \mathrm{E}\left[\sigma_{t-1}^2\right] + ... + \alpha_p \mathrm{E}\left[\sigma_{t-p}^2\right] \\
\mathrm{E}\left[\sigma_t^2\right](1 - \alpha_1 ... - \alpha_p) &= \alpha_0 \\
\mathrm{Var}(X_t) = \mathrm{E}\left[\sigma_t^2\right] &= \frac{\alpha_0}{1 - \alpha_1 ... - \alpha_p}. \qquad (6.10)
\end{aligned}
$$

Above is has been used that $\mathrm{E}\left(\omega^2\right) = \mathrm{Var}\left(\omega^2\right) = 1$ and because X_t is stationary, and $\mathrm{E}(X_t) = 0 \Rightarrow \mathrm{E}\left(\sigma_t^2\right) = \mathrm{E}\left(\sigma_{t-1}^2\right)$ because σ_t and ω_t are independent. σ_t depends on ω_{t-1},...,ω_{t-p} and ω_t is drawn at time t. From equation 6.10 it is concluded that the unconditional expectation of σ_t^2 is the same for all t, and $\sum_1^p \alpha_i < 1$ must be fulfilled in order to have positive unconditional finite variance and if the conditional mean is stationary then the ARCH model is covariance stationary. ARCH (and GARCH) processes must have positive autocorrelation. In an ARCH process with negative autocorrelation, a shock might be sufficiently large to cause the conditional variance to be negative which is undesirable.

If $\mathrm{Var}(X_t|F_{t-1}) = \sigma_t^2$ and X_t/σ_t is i.i.d. the ARCH is said to be strong. If only $\mathrm{Var}(X_t|\mathcal{F}_{t-1}) = \sigma_t^2$ the ARCH is said to be semi-strong.

Higher order moment of X_t is now studied. First the element of the skewness is investigated:

$$
\mathrm{E}\left(X_t^3\right) = \mathrm{E}\left[\mathrm{E}\left(X_t^3|\mathcal{F}_{t-1}\right)\right] = \mathrm{E}\left[\sigma_t^3(\mathrm{E}(\omega_t^3)\right] = 0.
$$

Therefore the skewness of an ARCH process is zero and this result is actually true for all odd moments of X_t. The kurtosis is:

$$\begin{aligned}
\kappa &= \frac{\mathrm{E}\left(X_t^4\right)}{\mathrm{E}\left(X_t^2\right)^2} \qquad (6.11)\\
&= \frac{\mathrm{E}\left[\mathrm{E}\left(X_t^4|\mathcal{F}_{t-1}\right)\right]}{\mathrm{E}\left[\mathrm{E}\left(\sigma_t^2\omega_t^2|\mathcal{F}_{t-1}\right)\right]^2}\\
&= \frac{\mathrm{E}\left[\sigma_t^4\mathrm{E}\left(\omega_t^4|\mathcal{F}_{t-1}\right)\right]}{\mathrm{E}\left[\sigma_t^2\mathrm{E}\left(\omega_t^2|\mathcal{F}_{t-1}\right)\right]^2}\\
&= \frac{\mathrm{E}\left[3\sigma_t^4\right]}{\mathrm{E}\left[\sigma_t^2\right]^2}\\
&= \frac{3\mathrm{E}\left[\sigma_t^4\right]}{\mathrm{E}\left[\sigma_t^2\right]^2}\\
&> 3
\end{aligned}$$

where it is used $\mathrm{E}\left[\omega_t^4\right] = 3$, because $\omega_t \sim N(0,1)$ is from a normal distribution. The last step is true because $Var\left[\sigma_t^2\right] = \mathrm{E}\left[\sigma_t^4\right] - \mathrm{E}\left[\sigma_t^2\right]^2 \geq 0$. The kurtosis of an ARCH process is larger than three and therefore leptokurtic ,meaning that is has higher kurtosis than a normal distribution.

ARCH models can be estimates using ordinary least squares, but often maximum likelihood estimation with the least squares estimates as initial value is preferred. Regarding this project, the estimation is done in R, which uses maximum likelihood methods [30].

6.2.3 GARCH

The generalized ARCH (GARCH) is an extension of the ARCH model with an autoregressive conditional volatility part. It is defined as [7, 27]:

$$\begin{aligned}
X_t &= \sigma_t\omega_t \qquad (6.12)\\
\sigma_t^2 &= \alpha_0 + \sum_{i=1}^{p}\alpha_i X_{t-i}^2 + \sum_{j=1}^{q}\beta_j\sigma_{t-j}^2. \qquad (6.13)
\end{aligned}$$

with the constraints:

$$\begin{aligned} \alpha_0 &> 0 \\ \sum_{1}^{max(p,q)} \alpha_i + \beta_i &< 1 \end{aligned}$$

α_i, β_i is further more constrained such that σ_t^2 is uniform positive for all t.$\{\omega_t\}$ is i.i.d with $\omega_t \sim N(0,1)$, α_0 is a constant underlying variance, α_i and β_i are the parameter in the model of order p and q.

The second constraint ensures the unconditional variance of X_t is finite, though the conditional variance σ_t^2 evolves over timer. If $p = q = 1$ then σ_t^2 is uniform positive if $\alpha_i, \beta_i \geq 0$ for $i > 0$ and the first condition is fulfilled. If $\text{Var}(X_t|F_{t-1}) = \sigma_t^2$ and X_t/σ_t is i.i.d. the GARCH is said to be strong. If only $\text{Var}(X_t|\mathcal{F}_{t-1}) = \sigma_t^2$ the GARCH is said to be semi-strong.

The properties of a GARCH(p,q) process is very similar to an ARCH(p) process. From the definition of the GARCH model it is seen that it features the volatility clumping as the ARCH model. The unconditional mean of X_t is:

$$\text{E}(X_t) = \text{E}\left[\text{E}(X_t|\mathcal{F}_{t-1})\right] = \text{E}\left[\sigma_t(\text{E}(\omega_t)\right] = 0,$$

The unconditional variance of X_t can be calculated as in 6.10 and therefore the derivation is omitted and only the result is presented:

$$\text{Var}(X_t) = \text{E}\left[\sigma_t^2\right] = \frac{\alpha_0}{1 - \sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_i} = \frac{\alpha_0}{1 - \sum_{i=1}^{max(p,q)} \alpha_i + \beta_i}.$$

By this it is clear that the constraints to 6.13 ensures the GARCH process to be stationary. As for the ARCH process the skewness of a GARCH process is zero by the same reasons. The calculation of the kurtosis follows the same steps as in 6.11 and the result remains $\kappa > 3$.

Again GARCH models is determined by maximum likelihood estimates, which in R is found by a build-in Quasi-Newton optimizer [30].

GARCH models can be used to forecast by letting t_γ be the forecast origin, then one-step forecasting can be obtained as:

$$\sigma^2_{t_\gamma+1} = \sigma^2_{t_\gamma}(1) = \alpha_0 + \sum_{i=1}^{p} \alpha_i X^2_{t_\gamma+1-i} + \sum_{j=1}^{q} \beta_j \sigma^2_{t_\gamma+1-j}, \qquad (6.14)$$

and a multiple ℓ-step forecast as:

$$\sigma^2_{t_\gamma+\ell} \quad = \quad \sigma^2_{t_\gamma}(\ell) = \alpha_0 + \sum_{i=1}^{p} \alpha_i X^2_{t_\gamma+\ell-i} + \sum_{j=1}^{q} \beta_j \sigma^2_{t_\gamma+\ell-j}. \qquad (6.15)$$

It can be shown for e.g. GARCH(1,1) that $\ell \to \infty$ then $\sigma^2_{t_\gamma}(\ell) \to \frac{\alpha_0}{1-\alpha_1-\beta_1} = \mathrm{Var}(X)$. This means that if GARCH model and the forecast horizon goes to infinity the variance of the model converges to the unconditional variance of X. Consequently it is $\alpha_1 + \beta_1$ that determine how fast the variance forecast converges to $\mathrm{Var}(X)$.

6.2.4 Limitation of ARCH and GARCH

Both ARCH and GARCH models have volatility defined by the size of the shock, X_t, but not whether it is positive or negative caused by the squaring of shocks in the model definition. In real world this is often not true, and the leverage effect is often observed. The leverage effect is the fact that the volatility increases more after negative news than after positive news.

ARCH an GARCH models tends to overestimate the volatility if large shocks are isolated, because they take this into account in the following points.

The effect of these limitations in relation to this project will be evaluated in chapter 10.

Chapter 7

Modelling of data

In the recent chapters eleven stock indices have been analysed. Now it is time to make use of these result in order to model the data by using the methods described in chapter 6. For this purpose R is a strong tool and is therefore used widely to solve and calculate problems that else would have taken far too much time in practice. The first task is to reduce the dimension of data by using PCA. Afterward data will be fitted with GARCH models in order to describe the reduced data space quantitatively in a mathematical sense.

The data that will be used here is weekly log returns because they have shown the best characteristics that fits the underlying assumptions in the model. For three reasons the DK00S/N will be left out from the modelling. First of all, the index shows a behaviour that is significantly different from the other indices. It is definitely not normal distributed, and that is a behaviour that should be fulfilled to a certain extent. If PCA is done with DK00S.N.Index despite the high degree of non-normality it would be represented by its own PC, because of the low correlation to the other indices. Secondly, it has not been possible to get index values before June 2003, consequently this index does not contain as much historical data as the other indices, which is a desirable property for the simulation part. Thirdly, it is presumably not possible to invest in the index directly, though it in some sense represents the money market. The ten indices left will be split into the regimes, defined in chapter 5. Each regime will be modelled individually in order to meet the problems concerned with modelling the asymmetry in the variance, e.g. the asymmetric reaction to positive and negative shocks.

7.1 PCA

The purpose of PCA in this project is mainly to describe data and thereby reduce the number of variables, therefore the meeting of the assumptions underlying PCA is not of high importance. The assumption of multivariate normality and independent random observations can roughly be accepted. Concerning the linearity of the variables, non-linear (ARCH) effects have been observed in data. This assumption cannot be met, but with that in mind a PCA will be performed. The R-function `prcomp` from the `stats`-package is used to find the principal components, and the method built on the correlation matrix is used because of the difference in variance in the indices. Data are also mean adjusted, though the mean is close to zero, just to ensure symmetry.

7.1.1 Period 1 1999M1 - 2000M8

This is a period with growing prices. In table 7.1 the principal components of period is listed together with standard deviation, proportion of variance and cumulative proportion of variance.

The eigenvectors in the first principal component are almost equally large, and therefore an almost equally weighted linear combination of all the indices. PC1 shows the general tendency in the market. In the PC2 the Japanese stock indices and the Danish bond indices have the highest weightings. This component shows that they drag in different ways. This is already known from the correlations in table 3.1. PC3 is mainly weighted by the stocks from Europe and USA and the Danish bonds.

76% of the total variance in data can be described by the first four PC's. In figure 7.1 is a scree plot for the PCA in period 1. Using this to determine the appropriate numbers of PC, the "elbow" is at $k = 4$. So using the four first PC's seems reasonable.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
KAXGI	0.32	0.03	-0.25	0.56	0.30	-0.44	0.47	-0.14	-0.05	0.04
NDDUE15	0.36	-0.06	-0.47	-0.21	-0.09	-0.06	-0.02	0.77	-0.01	0.02
NDDUJN	0.23	-0.54	0.34	-0.21	0.18	-0.09	0.08	0.02	-0.56	-0.38
NDDUNA	0.34	-0.04	-0.53	-0.17	-0.16	-0.09	-0.45	-0.56	-0.11	-0.13
NDUEEGF	0.16	-0.35	0.14	0.50	-0.76	0.09	-0.03	0.05	0.04	-0.01
TPXDDVD	0.28	-0.57	0.16	-0.14	0.26	-0.04	-0.10	-0.09	0.56	0.44
CSIYHYI	0.33	0.14	0.10	0.45	0.40	0.55	-0.42	0.12	-0.08	-0.07
JPGCCOMP	0.38	0.13	-0.05	-0.26	-0.13	0.58	0.60	-0.22	0.02	0.04
NDEAGVT	0.35	0.38	0.37	-0.12	-0.16	-0.25	-0.14	0.002	-0.38	0.58
NDEAMO	0.37	0.34	0.37	-0.11	-0.08	-0.27	-0.05	0.032	0.46	-0.55
SD	1.89	1.38	1.10	0.94	0.87	0.80	0.65	0.61	0.39	0.31
PoV	0.36	0.19	0.12	0.09	0.08	0.06	0.04	0.04	0.02	0.00
CP	0.36	0.55	0.67	0.76	0.83	0.90	0.94	0.98	1.00	1.00

Table 7.1: Principal components for period 1 each with Standard deviation (SD), Proportion of Variance (PoV) and Cumulative Proportion (CP).

Figure 7.1: Scree plot of PCA in period 1.

7.1.2 Period 2 2000M9 - 2001M12

Period 2 is a period defined as an overall downturn, but as seen earlier, the bond indices actually do not decrease. The can be seen by the first principal components. PC1 is primary weighted by the stocks and PC2 is primary weighted by bonds, so PC1 is a stock component and PC2 is a bond component. Using four principal components 81% of the total variance is described.

	PC1	PC2	PC3	PC4
KAXGI	-0.43	0.01	-0.02	0.30
NDDUE15	-0.40	-0.16	-0.12	0.19
NDDUJN	-0.30	-0.16	0.55	-0.22
NDDUNA	-0.42	0.05	-0.06	0.38
NDUEEGF	-0.39	0.12	-0.23	-0.33
TPXDDVD	-0.33	-0.10	0.54	-0.24
CSIYHYI	-0.29	0.09	-0.35	0.20
JPGCCOMP	-0.18	-0.23	-0.46	-0.66
NDEAGVT	0.14	-0.63	-0.07	0.22
NDEAMO	0.01	-0.68	-0.04	0.06
SD	1.98	1.33	1.26	0.90
PoV	0.39	0.18	0.16	0.08
CP	0.39	0.57	0.73	0.81

Table 7.2: First four principal components for period 2 and Standard deviation (SD), Proportion of Variance (PoV) and Cumulative Proportion (CP) for each.

7.1.3 Period 3 2002M1 - 2008M2

The third period is a very long period with positive trend. Again PC1 is a stock component and PC2 is a bond component and the first four PC explains 84% of the variance.

	PC1	PC2	PC3	PC4
KAXGI	-0.35	0.01	0.19	-0.24
NDDUE15	-0.40	-0.01	0.22	-0.38
NDDUJN	-0.35	-0.10	-0.6	0.07
NDDUNA	-0.37	0.05	0.26	-0.38
NDUEEGF	-0.40	-0.08	0.00	0.00
TPXDDVD	-0.36	-0.01	-0.56	0.09
CSIYHYI	-0.26	-0.30	0.27	0.53
JPGCCOMP	-0.18	-0.47	0.27	0.39
NDEAGVT	0.21	-0.57	-0.11	-0.27
NDEAMO	0.16	-0.59	-0.11	-0.36
SD	2.14	1.41	1.01	0.87
PoV	0.46	0.20	0.10	0.08
CP	0.46	0.66	0.76	0.84

Table 7.3: First four principal components for period 3 and Standard deviation (SD), Proportion of Variance (PoV) and Cumulative Proportion (CP) for each.

7.1.4 Period 4 2008M3 - 2009M2

The fourth period is the latest financial crisis. The stock indices have experienced massive falls, and even JPGCCOMP and CSIYHYI have also fallen. This is also reflected in the first principal component as a crisis component. The two Danish bond indices have not been affected by the crisis as much, and it is possible to see in PC1 where they are weighted almost zero. On the contrary the second PC shows their behaviour through the crisis, where they are heavily weighted. The two first PC's explain 82% of the variation in data, but for ease in the later modelling four PC's is used. PC4 is a CSIYHYI component, by its high weighting.

	PC1	PC2	PC3	PC4
KAXGI	0.37	-0.11	0.18	0.11
NDDUE15	0.37	-0.09	0.16	-0.15
NDDUJN	0.33	-0.03	-0.65	-0.38
NDDUNA	0.35	-0.10	0.19	-0.05
NDUEEGF	0.37	-0.02	0.32	-0.12
TPXDDVD	0.36	-0.04	-0.27	-0.28
CSIYHYI	0.30	-0.04	-0.41	0.83
JPGCCOMP	0.34	0.18	0.35	0.15
NDEAGVT	0.01	0.70	-0.13	-0.08
NDEAMO	0.13	0.67	0.07	0.03
SD	2.54	1.33	0.70	0.68
PoV	0.65	0.18	0.05	0.05
CP	0.65	0.82	0.87	0.92

Table 7.4: First four principal components for period 4 and Standard deviation (SD), Proportion of Variance (PoV) and Cumulative Proportion (CP) for each.

7.1.5 Period 5 2009M3 - 2011M3

Period 5 is a kind of reaction on the big losses in period 4. Therefore the PC1 consists of all the indices almost equally weighted, where the Danish bonds points in the opposite direction. 88% of the volatility in data is explained by the first four principal components.

	PC1	PC2	PC3	PC4
KAXGI	0.34	-0.04	0.13	-0.58
NDDUE15	0.38	-0.06	0.17	0.20
NDDUJN	0.26	0.34	-0.57	0.05
NDDUNA	0.37	-0.08	0.11	0.30
NDUEEGF	0.38	0.07	0.11	0.40
TPXDDVD	0.31	0.19	-0.55	0.03
CSIYHYI	0.34	0.17	0.13	-0.57
JPGCCOMP	0.30	0.25	0.47	0.10
NDEAGVT	-0.23	0.59	0.04	-0.10
NDEAMO	-0.18	0.62	0.25	0.15
SD	2.37	1.25	1.04	0.72
PoV	0.56	0.16	0.11	0.05
CP	0.56	0.72	0.83	0.88

Table 7.5: First four principal components for period 5 and Standard deviation (SD), Proportion of Variance (PoV) and Cumulative Proportion (CP) for each.

7.1.6 Communalities

Communality h_{ij} of index i in one regime is calculated as the sum of squared weighting a_{ij} in the PC's used [1]:

$$h_{ij} = \sum_{j=1}^{k} a_{ij}^2$$

Communality is a measure of the how much variance in the index can be explained by the k component model. Using the R function `prcomp` the weighting cannot be accessed directly, but multiplying the PC value (or eigenvector element) u_{ij} with the standard deviation (or square root eigenvalue) λ_j the weightings are found:

$$h_i = \sum_{j=1}^{k} u_{ij}^2 \lambda_j$$

Period	1	2	3	4	5
KAXGI	0.71	0.79	0.66	0.92	0.84
NDDUE15	0.77	0.73	0.89	0.93	0.86
NDDUJN	0.91	0.90	0.94	0.98	0.93
NDDUNA	0.77	0.81	0.80	0.85	0.86
NDUEEGF	0.56	0.79	0.77	0.92	0.92
TPXDDVD	0.88	0.95	0.93	0.93	0.94
CSIYHYI	0.61	0.57	0.78	1.00	0.87
JPGCCOMP	0.60	0.90	0.77	0.86	0.85
NDEAGVT	0.88	0.81	0.93	0.87	0.86
NDEAMO	0.89	0.83	0.92	0.90	0.87

Table 7.6: Communalities for each index in each period using four principal components.

The communalities for each period are found in table 7.6. In general a lot of the variance in each index can be explained by the four PC's. The lack of explanation in CSIYHYI in the first periods might very well be explained by the significant lags in the autocorrelation. In period four CSIYHYI has it "own" component, resulting in a communality on 1, by which all variance are explained. In figure 7.2 NDDUJN and CSIYHYI has been reconstructed using four principal components. These two indices are chosen because they are representing indices with highest and lowest communalities. Both indices shows a nice behaviour close the original data and that CSIYHYI fits the original data completely in the fourth period because of the high communality in that very period. From the communalities and the plot it is concluded that four principal component is enough to describe data reasonable.

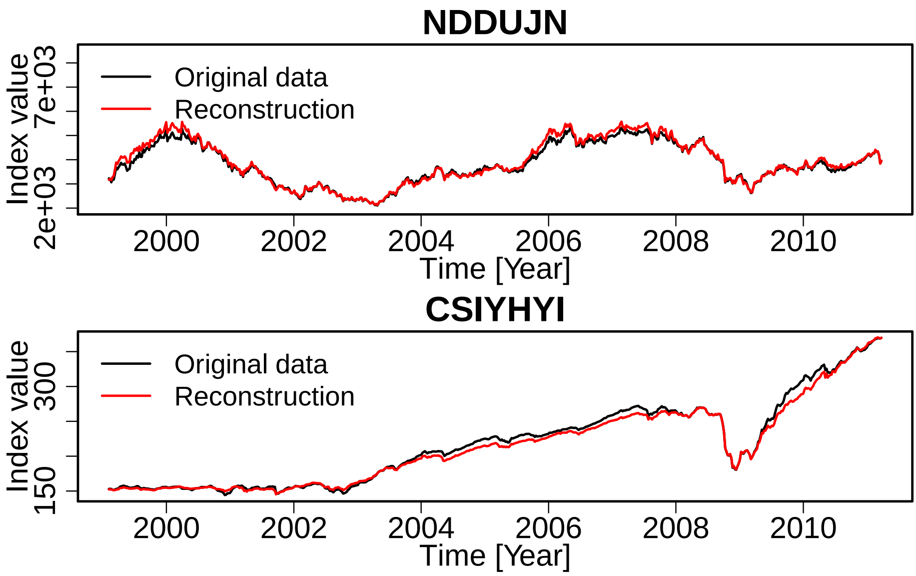


Figure 7.2: Reconstruction of NDDUJN and CSIYHYI when using four principal components.

7.1.7 Outline of PCA

If the PC's for the periods are compared, it can be concluded that they are very different and change over time. This is also expected, as it has already been concluded that there are different volatilities depending on the trend. There are also different types of periods with growing prices and falling prices. In the periods with falling prices the two periods modelled here show a bit of the same behaviour. This can be seen by the similarities in the PC's of period 2 and 4.

It is anticipated that the behaviour in the PC scores has the same behaviour of the data. In figure 7.3 a QQ plot of the PC scores from the PC's in period 1 related to each index are found. The PC scores do not look normal, but since it was a little rough assumption to data, a better result would not have been expected after transforming data. It will not be appropriate to make any more statistical tests on the PCA, because weekly log return did not meet the assumption on PCA totally. Thereby, statistical tests might not present the accurate picture and they would be useless.

It was shown that using four PC's is enough to explain at least 76 % of the total variance, in some periods even more. 76 % and high communalities is an acceptable level and therefore it will be the first four PC's representing each regime that is modelled with GARCH or ARCH models in the next section. As this is a description of data, it is acceptable to use this information in further modelling despite lack of fulfilling the assumptions.

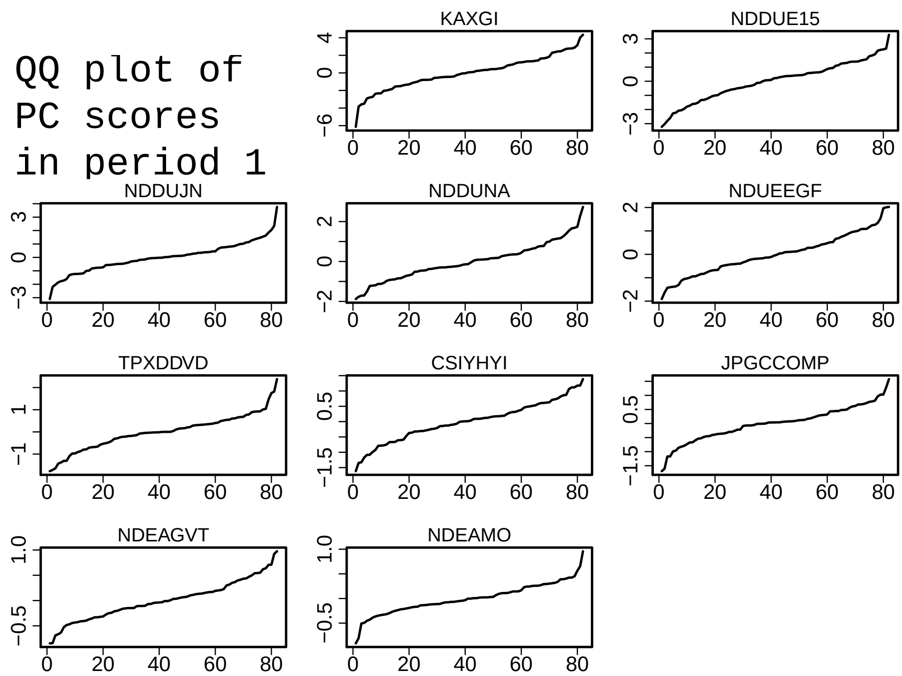


Figure 7.3: QQ plot of the PC scores in period 1.

7.2 GARCH

GARCH models are used to model the dynamic structure in the "new data", data in PCA space, derived using four principal components. They will be named PCA data. Actually it is the PCA data that should fulfil the requirements of using GARCH, but if the log return indices do not fulfil the requirement, neither will PCA data, and therefore we are satisfied with the analysis of weekly log return indices. The PCA data is mean adjusted from the PCA transformation, and it is assumed that there are no significant lags in log return indices on figure 4.3a. Also weekly log returns are assumed to be independent, and it was shown that the skewness is almost zero and the kurtosis is larger than 3, just as expected for a GARCH process. The volatility clustering has not been tested, only observed on plots. If the clustering is not represented significantly in data it will not be significant in the model, so tests for clustering are not necessary. Because the models are built on PC-data, they do not represent a single index, but some general tendencies and aspects in the financial market. E.g. some of the models based on the first PC-data are models for the overall stock market and overall the bond market. The indices' behaviour can then be estimated upon their dependency on the behaviour of general market, that is turning from PCA space to data space.

To estimate the parameters and order of each PCA data series the R-function `garch` from the `tseries`-packace is used. It has some tests build-in that will be explained briefly. The Jarque Bera test [32] tests the residuals for normality, with:

$$H_0 : \begin{cases} \varsigma_{Fisher} = 0 \\ \kappa_{Excess} = 0 \end{cases} .$$

The test statistics is:

$$JB = \frac{n}{6}\left(\varsigma^2 + \frac{1}{4}\kappa^2\right),$$

where n is the number of observations. If the residuals are normal distributed the test statistics follows a $\chi^2(2)$ distribution.

The Box-Ljung test [35] tests the squared residuals for the null hypothesis that they are independently distributed (random data), with the test statistics:

$$Q = n(n+2)\sum_{i=1}^{h}\frac{\hat{\rho}_i^2}{n-i},$$

where n is the number of observations and h is the number of lags being tested in the autocorrelation with parameter$\hat{\rho}_i$. If the squared residuals are random, the test statistics follows a $\chi^2(k)$ distribution.

When modelling data with GARCH models it is desirable that the residuals are Gaussian white noise, therefore accepting H_0 in the Jarque Bera test and the Box-Ljung test (large p-value), it cannot be rejected that the residuals are Gaussian white noise.

To see which model order fits the best, the summery output of different order combinations are studied. It is an iterative process when determining the model order, therefore the modelling of the first period of the first PCA data series is studied closer to show the considerations. First a model of order $(p, q) = (0, 1)$ is modelled. The summary from R is:

```
> summary(gp1[[1]])
Call:
garch(x = data_pca[, 1], order = c(0, 1))
Model:
GARCH(0,1)
Residuals:
Min 1Q Median 3Q Max
-3.32511 -0.58060 0.04404 0.70820 2.34910
Coefficient(s):
Estimate Std.  Error t value Pr(>|t|)
a0 3.408e+00 5.583e-01 6.104 1.04e-09 ***
a1 3.722e-14 1.570e-01 0.000 1
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1
Diagnostic Tests:
Jarque Bera Test
data:  Residuals
X-squared = 2.0539, df = 2, p-value = 0.3581
Box-Ljung test
data:  Squared.Residuals
X-squared = 0.1639, df = 1, p-value = 0.6856
```

The summary contains info on the residuals, the coefficients and the p-values from Jarque Bera test and Box-Ljung test. The $a0$ (α_0 in 6.13) is significant but $a1$ is not, but its size is close to zero so it will be ignored. The Jarque Bera test fails to reject H_0 but the null hypothesis of the squared residual being random cannot be rejected in the Box-Ljung test. The constraints in 6.13 are obeyed. Another model might perform better, therefore a $(1, 1)$ model is tested. The R-summary is:

```
> summary(gp1[[1]])
Call:
garch(x = data_pca[, 1], order = c(1, 1))
Model:
GARCH(1,1)
Residuals:
Min 1Q Median 3Q Max
-3.3142 -0.5787 0.0439 0.7059 2.3414
Coefficient(s):
Estimate Std.  Error t value Pr(>|t|)
a0 3.229e+00 2.640e+02 0.012 0.990
a1 3.115e-14 1.620e-01 0.000 1.000
b1 5.883e-02 7.696e+01 0.001 0.999
Diagnostic Tests:
Jarque Bera Test
data:  Residuals
X-squared = 2.0539, df = 2, p-value = 0.3581
Box-Ljung test
data:  Squared.Residuals
X-squared = 0.1639, df = 1, p-value = 0.6856
```

None of the coefficient are now significant, and therefore this model is useless. The $(0, 2)$ is tried instead:

```
> summary(gp1[[1]])
Call:
garch(x = data_pca[, 1], order = c(0, 2))
Model:
GARCH(0,2)
Residuals:
Min 1Q Median 3Q Max
-3.2058 -0.6091 0.0479 0.6971 2.4004
Coefficient(s):
Estimate Std.  Error t value Pr(>|t|)
a0 3.230e+00 9.670e-01 3.341 0.000836 ***
a1 3.618e-14 1.594e-01 0.000 1.000000
a2 5.830e-02 1.530e-01 0.381 0.703122
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1
Diagnostic Tests:
Jarque Bera Test
data:  Residuals
X-squared = 1.4659, df = 2, p-value = 0.4805
Box-Ljung test
data:  Squared.Residuals
X-squared = 0.2078, df = 1, p-value = 0.6485
```

Of course these result are close to the (0,1) model. There is still only one significant parameter, and the residuals do not behave any better, therefore the (0,1) model is preferred. Sometimes testing for higher order suddenly gives a nice result. These tests for higher order have also been done in this project, but it has no interest to the reader unless an appropriate model is found, therefore useless models are not presented.

In table 7.7 all the estimated GARCH parameters together with p-values, Box-Ljung, Jarque Bera test results are found.

Period	PC-data	α_0	p-value	α_1	p-value	J-B	B-L
1	1	3.408	1.0e-09	3.7e-14	1	0.358	0.69
1	2	1.712	1.8e-04	0.088	0.698	0.616	0.96
1	3	1.210	2.7e-13	0.004	0.972	0.108	0.99
1	4	0.842	5.8e-10	8.8e-15	1	0.410	0.37
2	1	2.638	3.0e-04	0.331	0.041	0.203	0.81
2	2	1.499	7.7e-08	0.076	0.571	0.050	0.66
2	3	1.506	3.6e-12	2.8e-15	1	0.132	0.33
2	4	0.390	0.008	0.646	0.044	0.261	0.71
3	1	4.336	$<$2e-16	0.055	0.27	2.9e-06	0.78
3	2	1.842	$<$2e-16	0.074	0.136	4.6e-07	0.90
3	3	0.865	$<$2e-16	0.162	0.082	0.230	0.75
3	4	0.679	$<$2e-16	0.098	0.182	$<$2e-16	0.78
4	1	1.728	0.002	1.080	0.002	0.019	0.31
4	2	1.564	7.3e-04	0.108	0.692	0.598	0.90
4	3	0.302	2.1e-04	0.334	0.214	0.088	0.55
4	4	0.440	6.1e-05	0.012	0.96	0.025	0.99
5	1	3.139	2.2e-06	0.494	0.048	4.5e-10	0.46
5	2	1.508	3.8e-06	0.041	0.772	0.698	0.97
5	3	1.037	1.5e-08	0.034	0.759	0.561	0.96
5	4	0.446	$<$2e-16	0.022	0.689	1.2e-12	0.49

Table 7.7: Estimates of model parameters of the best GARCH fit, with p-value for each parameter and test results of the Jarque Bera test and the Box-Ljung test.

In table 7.7 it appears that the constraints of ARCH models are not met in period 4 PC-data 1, where $\alpha_1 > 1$. I order to have finite variance, another model needs to be found. It turns out that the model with the lowest order that meets all constraints is an ARCH(3). The model parameter is found in table 7.8.

Period	PC-data	α_0	p-value	α_1	p-value
4	1	5.268	0.0448	7.636e-01	0.0423
α_2	p-value	α_3	p-value	J-B	B-L
4.804e-02	0.8028	9.993e-15	1	6.623e-05	0.1228

Table 7.8: Model parameters and test statistics for PC-data 1 in period 4.

As it is seen, it turns out that all the best fits are ARCH models of order 1, except the one mention above. Actually a lot of the models should be ARCH(0) models, because α_1 is not significant. ARCH(0) processes are Gaussian white noise processes because:

$$
\begin{aligned}
X_t &= \sigma_t \omega_t && (7.1)\\
\sigma_t^2 &= \alpha_0, && (7.2)
\end{aligned}
$$

or

$$
X_t = \alpha_0^2 \omega_t \qquad (7.3)
$$

Despite that those models with insignificant α_1, the size of α_1 is relatively small and therefore they are still modelled as ARCH(1) processes. Some of the models have residuals that are not normal distributed, though all squared residuals are independently distributed. The use of weekly data seems resonable to obtain independent residuals, but having non-normal residuals is almost a matter of course when input data are non-normal with outliers. The test for normality easily fails if only a few residual are outliers, and therefore this will be ignored.

An ARCH(1):

$$
\begin{aligned}
X_t &= \sigma_t \omega_t && (7.4)\\
\sigma_t^2 &= \alpha_0 + \alpha_1 X_{t-1}^2, && (7.5)
\end{aligned}
$$

can be represented as an AR(1) process for X_t^2:

$$
\begin{aligned}
\sigma_t^2 &= \alpha_0 + \alpha_1 X_{t-1}^2 \\
\sigma_t^2 + X_t^2 - \sigma_t^2 &= \alpha_0 + \alpha_1 X_{t-1}^2 + X_t^2 - \sigma_t^2 \\
X_t^2 &= \alpha_0 + \alpha_1 X_{t-1}^2 + X_t^2 - \sigma_t^2 \\
X_t^2 &= \alpha_0 + \alpha_1 X_{t-1}^2 + \sigma_t^2 \left(\omega_t^2 - 1\right) \\
X_t^2 &= \alpha_0 + \alpha_1 X_{t-1}^2 + v_t,
\end{aligned}
$$

where $v_t = X_t^2 - \sigma_t^2$ is the surprise in volatility. ω_t is i.i.d with $\mathrm{E}(\omega_t^2) = 1$, therefore v_t is a white noise process with $\mathrm{E}(v_t) = 0$. ℓ-step forecasting in ARCH(1) is by [27]:

$$X_{t+\ell} = \sigma_{t+\ell}\omega_{t+\ell} \tag{7.6}$$

$$\sigma_{t+\ell}^2 = \alpha_1^\ell X_t^2 + \sum_{i=0}^{\ell-1} \alpha_0 \alpha_1^i. \tag{7.7}$$

This will be useful when generating scenarios, though it is a function in R that is used to simulate data.

It is actually quite surprising that none of the models are GARCH models, and in some cases not even ARCH models, despite the fact stated in the data analysis chapter where a test for data being white noise could not be rejected. In chapter 8 the models from table 7.7 and 7.8 will be used to generate scenarios. Scenarios, based on other methods, will be generated in order to evaluate the use of non-significant models. The subjective choice of models will be discussed in chapter 10, and the model will be tested in chapter 9.

Chapter 8

Scenario generation

In the recent chapters data has been analysed, transformed and at last modelled into four models in each of the five regimes. Now it is time to use these models to generate scenarios. The scenarios should not be predictions of the future index values, but represent a wide range of possible future index values. The approach to generate scenarios might be very subjective, but in order to get good scenarios different thing and aspect should be considered at first, such as the use of the scenarios and which characteristics in data are necessary in order to get realistic scenarios, and afterward build in to the algorithm for the generation.

The purpose of generating scenarios in this project is to have scenarios that can be used in the investment process, more specified in the asset allocation process where optimization and risk management are key factors. For this reason, not only "realistic" scenarios are a must but also extreme events are a desirable quality. Scenarios should be constructed such that they can be tested for correctness and accuracy. There might be patterns and characteristics in data that are unique or special for exactly that class of data. In this project, volatility clustering, trend, decreasing volatility when prices rise and increasing volatility when prices fall and correlation to other asset classes have been observed. It is also important that the scenarios take basic economic assumptions into account, such as no arbitrage principle etc. Two different approaches have been used to generate scenarios. The first presented in this chapter is based on data in principal component space modelled with ARCH models. The other method is a simple bootstrap.

8.1 Scenario generation using ARCH models

The regime divided ARCH models on principal component are chosen in order to keep as much of the characteristics in data as possible. ARCH models are used in order to keep the behaviour in the indies e.g. trend and volatility clustering, however some models did not have this quality after splitting data into regimes. PCA ensure the correlation between the indices, and reduces the number of models to be made. Splitting data into regimes ensures the different behaviour between up and down periods.

8.1.1 Approach for scenario generation

The approach used here generates scenarios of tendencies in the market, and then these are translated into specific indices values based upon its reaction to these tendencies, cf. the use of principal component data. The data modelled ends 25 March 2011, but data is available until 12 August 2011, without any intermediate turning points. Therefore the simulations are assumed to start at 13 August 2011 in a period with falling prices. The idea in this approach is first to choose a regime randomly, however still make use of the fact that after a downswing, a regime with growing prices will come. The length of a regime depends on which regime is chosen, because it is Erlang distributed with mean equal to the regime length sampled. The use of the Erlang distribution is carefully selected because it is said to be memory less:

$$P\left(T > \tau + t | T > \tau\right) = P(T > t) \;\; \forall\; \tau, t \geq 0.$$

This means that e.g. the first scenario is not affected by the downswing already has lasted for twenty weeks, the probability of waiting another twenty weeks is the same. This also ensures diversity in the scenarios because they will all have different regime lengths even if the same regime is drawn. Now a regime and a length, τ, are sampled, thereby the ARCH model is known and therefore the R function `garch.sim` from the `TSA`-package is used to simulate/forecast the model τ time steps ahead in accordance with equation 7.7. A part of a scenario is now generated, but the sampling of regime types and waiting times keeps going on until the scenario length exceed 5years= 260weeks. After the scenario has been generated the same procedure is repeated until a desired amount of scenarios is generated. Because simulated values are in PCA space the data need to be turned back in to first log return, and afterward into data space. The scenario approach in pseudo code:

```
N = numbers of scenarios
for n in 1 to N {
regime type = up
scenario length = 0
Step 1:
        if regime type = up {
        Chose either regime number 2 or 4 randomly
        Set regime type = down
        Chose period length:
        τ ~ Erlang (k = Period length (regime number), λ = 1)
        }else{
        Chose either regime number 1, 3 or 5 randomly
        Set regime type = up
        Chose period length:
        τ ~ Erlang (k = Period length (regime number), λ = 1)
        }

Step 2:
        for 1 to numbers of series in PCA space {
        Simulation = forecast using garch.sim τ time steps ahead
        }

Step 3:
        Scenario length = Scenario length + τ
        if Scenario length < 5 years {
        Go to step 1
        }else{

        Step 4:
                Transform data from PCA space to log return space
                Transform data from log return space to index space
        }
}
```

After scenarios have been generated, analytic methods are used to find scenarios with interesting behaviour. Maximum drawdown (MDD) is widely used in risk management. Taking one series at the time, drawdown at time, t_τ, is a measure form a historical maximum to a historical minimum within a time interval $[t_{\tau-1}, t_\tau]$. In this project the interval is one week. Therefore drawn down is only measured at time points where index prices are falling. Maximum drawdown at time, t_τ, is the largest drawdown in the interval $[t_0, t_\tau]$. It can be used as an estimate of the maximal volatility in the index, if such parameters should be estimated. Normally relative drawdowns are of interest, and the drawdowns given in this thesis are mainly the percentage change from the peak to the trough. The scenario containing the largest drawdown among all the indices can be used as a "worst case" scenario in the risk-return appraisal for the index, even though is not certain that it is the scenario ending at the lowest index price.

8.1.2 Scenarios

On the following pages the generated scenarios are found in figure 8.1, 8.2, 8.3 and 8.4. For each index, 1000 scenarios have been generated, such that the sampled regimes and lengths are parallel for all indices. Twenty scenarios are plotted with 5%, 25%, 50% (median), 75% and 95% quantiles. The scenario containing the largest maximal drawdown among all the scenarios is highlighted in red. Some scenarios take very extreme values, and for this reason, the medians are used instead of the mean to give a true picture of the "general" scenario. Single events cannot explain the "wavy" behaviour of the quantiles. This is caused by a large amount of scenarios with high volatility. It is clearly to see that all the simulations start in a regime where prices are falling, but is the lengths of the regime are different. Some scenarios hit a turning point fast, others waits for a longer time. Looking at the 50% quantile for the stock indices there seems to be a similar behaviour. There is a trough is around 2012 M10, and then a peak around 2014 M8 and then again a trough at 2015 M12. These observations might be used cautiously as estimate of turning points, which corresponds to regime length of 83, 95 and 70 weeks. The lengths of the "negative" regimes are a little longer than the ones in historical data, and the length of the "positive" regime seems very reasonable. This is an interesting feature that the model might be able to estimate turning points, though lac of historic knowledge of regime lengths make this very uncertain. In general both the stock indices and the bond indices seems to be highly correlated also across asset classes. This feature is also expected as it is represented in principal components. cf. PC1 is the market component, sometimes the stock component and PC2 is somtimes the bond component. The impact of the long third positive period in historical data is observable in many scenarios with positive trend and small volatility, though the regime here seems to be much shorter.

The stock indices have higher volatility than the bond indices, and therefore also a wider span on the quantiles. Despite majority of random walk models, the ARCH(1) components still have influence on the scenarios since it is possible to detect volatility clustering in some scenarios. The largest MDD scenario seems to be the same for all the indices except NDEAGVT. This scenario might be characterized as having a regime known as a bear market which is a market changing form high positive atmosphere to overall negativity. For the stock indices this scenario gets close to zero, but the bonds are not that badly hit. In chapter 9 the scenarios are studied more closely by considering validity and usefulness.

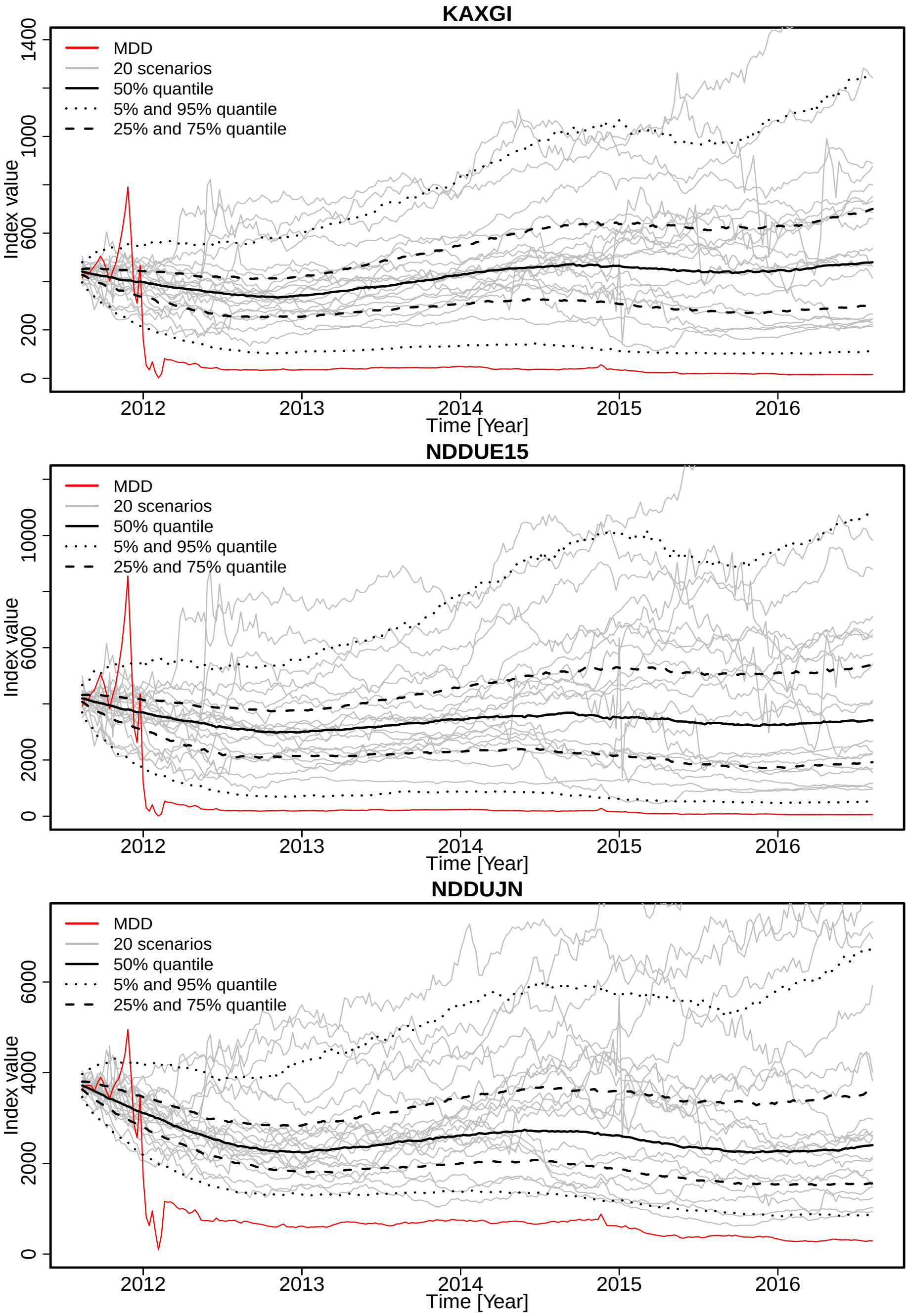


Figure 8.1: Scenarios for KAXGI, NDDUE15 and NDDUJN generated using PCA and GARCH models.

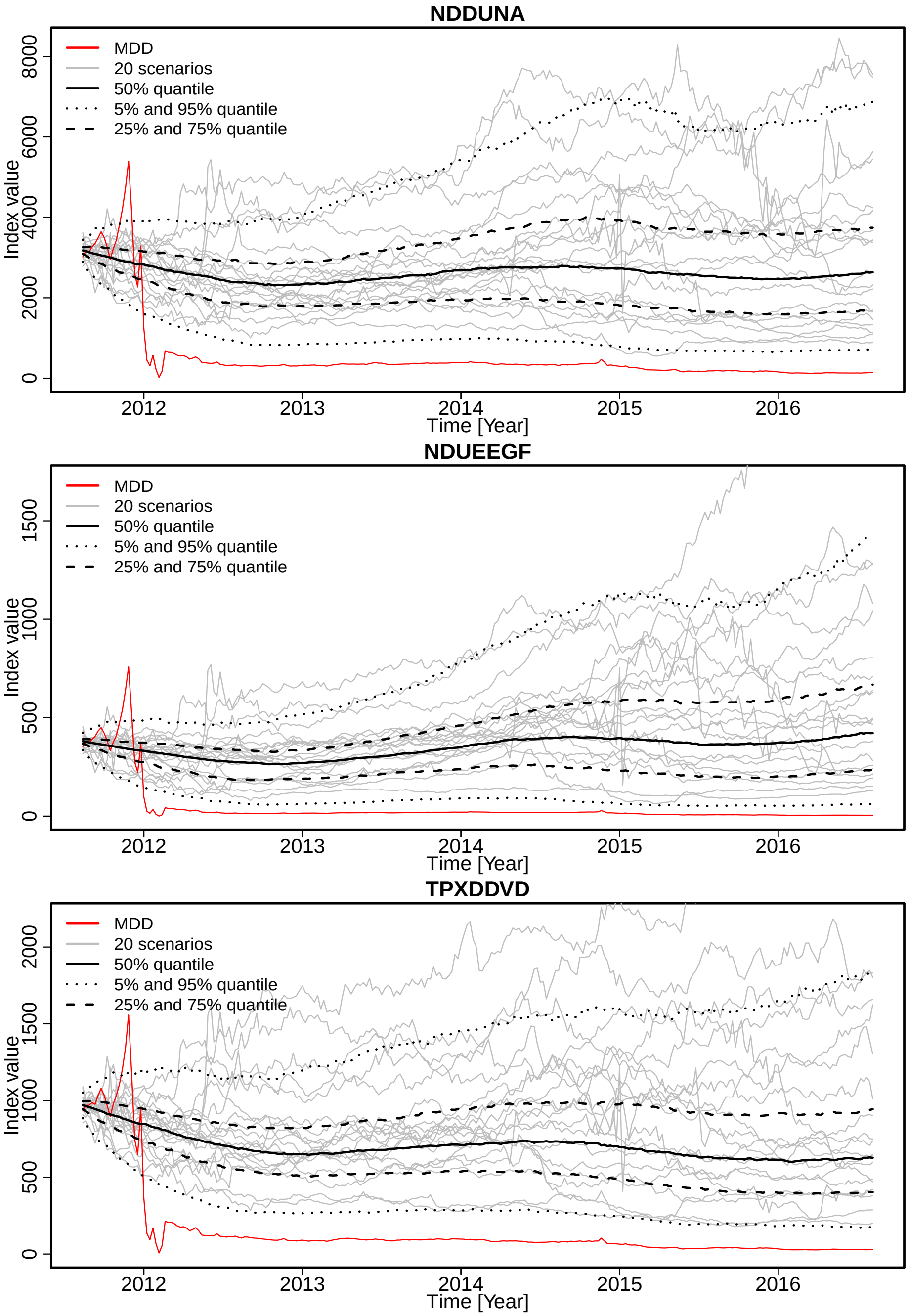


Figure 8.2: Scenarios for NDDUNA, NDUEEGF and TPXDDVD generated using PCA and GARCH models.

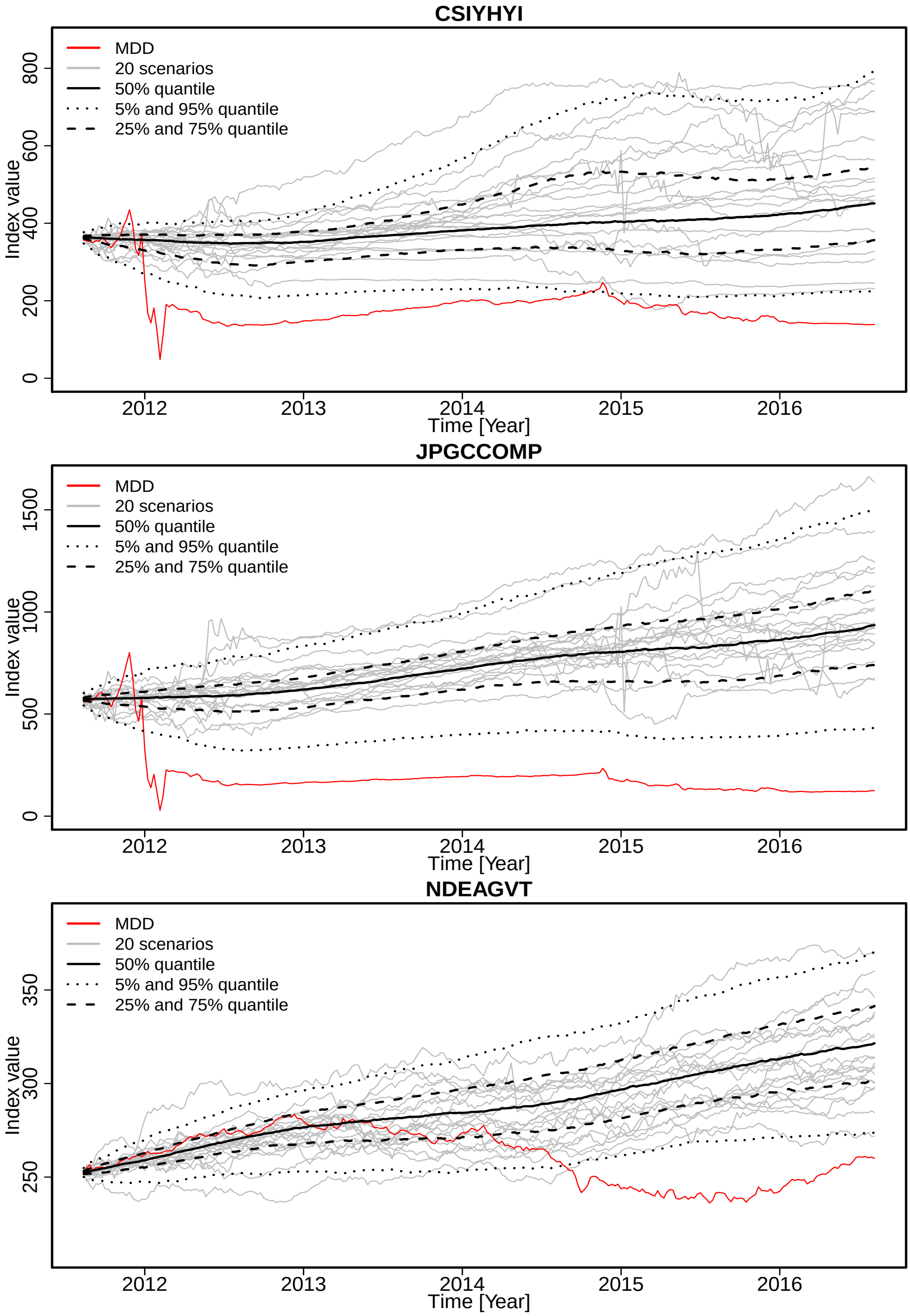


Figure 8.3: Scenarios for CSIYHYI, JPGCCOMP and NDEAGVT generated using PCA and GARCH models.

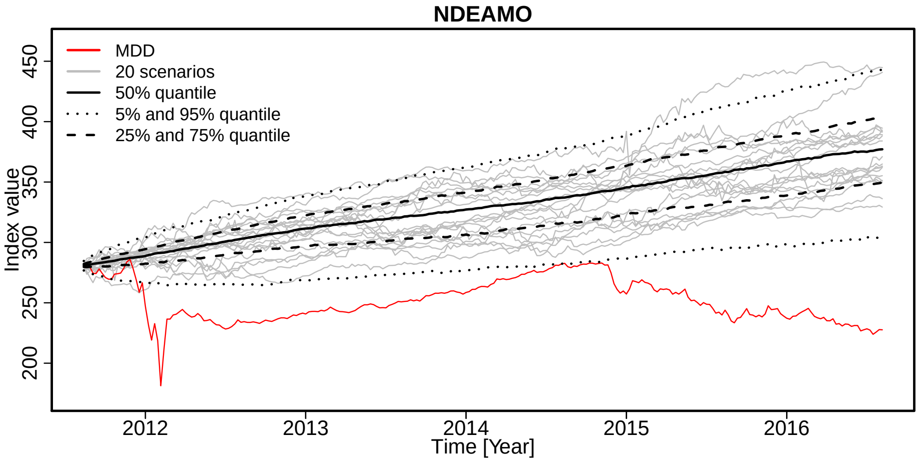


Figure 8.4: Scenario for NDEAMO generated using PCA and GARCH models.

8.1.3 Analysis of scenarios

One of the factors in risk management is the risk-return ratio. In figure 8.5 the normalized median of each index is plotted. In this plot the correlation in the indices is very clear, especially within an asset class. The stock indices are highly correlated and the corporate bond and the high yield bond have some of the same patterns. The Danish bonds are highly correlated, and have almost identical behaviour in the first regime. This is naturally caused by the homogeneity of the weighting in the principal components in falling regimes. This plot can be used to compare the indices and determine which one has the highest potential return at a given time among all the indices. Overall, the JPGCCOMP index performs the best, and in general the bond indices have the highest relative return after five years. Depending on the investing strategy, it might be an idea to invest in NDEAMO until 2013 and then switch to JPGCCOMP in order to get the highest return throughout the whole period. This plot only gives information about the return, and does not give any idea about the risk attached to the index. For this reason it would be interesting to look at the distribution of the end values of the indices, in order to get an idea of the risk.

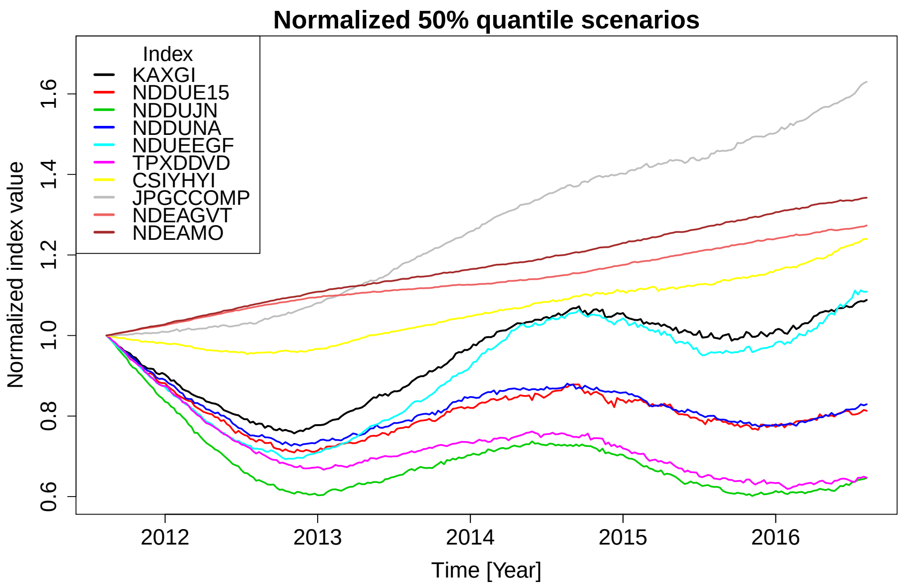


Figure 8.5: Normalized 50% quantile scenario for each index.

Figure 8.6 shows histograms of the end values for each index, together with markings of 5%, 50% and 95% quantiles. A 50% quantile at zero corresponds to a zero return on the median. The end value distributions look similar with in an asset class. The stock indices are positively skewed with a higher degree of deviation, because they are having some extreme scenarios with very high return. Again it is stated that the bonds indices have the highest return, some of the stock indices even have a lot scenarios with negative 5-year returns. This plot together with table 8.1 is a nice tool for risk-return valuation for 5-year investments. In the table the standard deviation on the 5-year return is given together with the median return. As an example of the risk-return considerations, the JPGCCOMP can be used. It has the highest median return at 63.1% change over 5 years corresponding to an annual effective rate on $\sqrt[5]{1+0.631}-1 = 0.103 = 10.3\%$. But the standard deviation of the relative changes is $\hat{\sigma} = 0.671$, which is the highest among the bond indices. Whether the investor should choose to accept this higher risk in order to get a higher return depends on the overall strategy. It is unusual, is that some of the indices with high standard deviation on the 5-year relative return, have a negative median return. The extreme scenarios, the fact that the simulation starts in a bearish market and that the simulation horizon only is five years lead to this unfavourable risk return relationship. If the horizon was longer, the overall positive trend in historical data would probably have affected results, such that

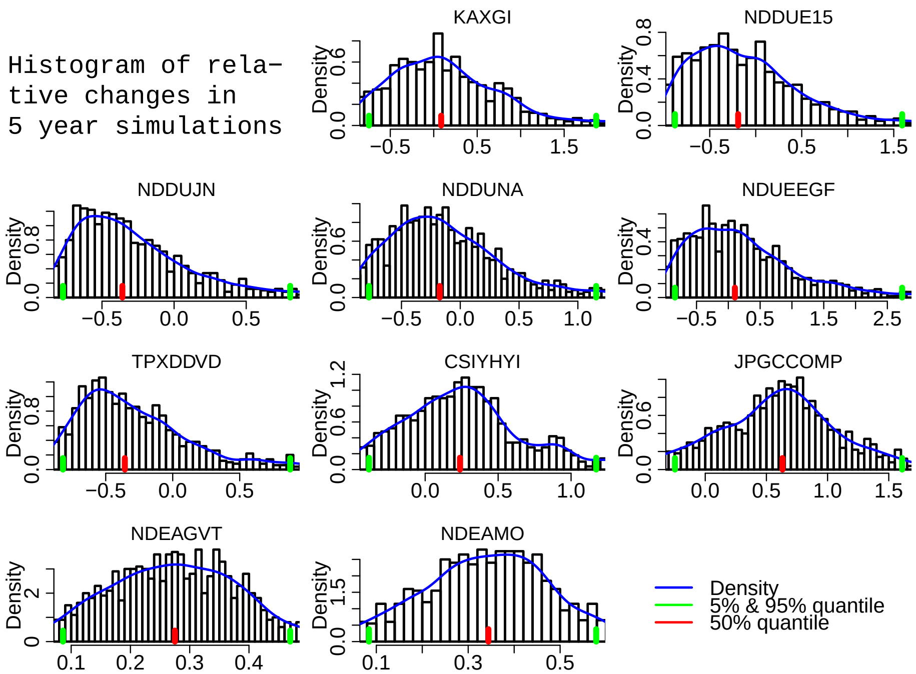


Figure 8.6: Histogram of the relative end values for the series.

the risk on the stock indices still would have been higher, but the return would have exceeded the return on bond indices. But this behaviour is anyway still desirable and most likely because of the market situation at the starting point.

Index	KAXGI	NDDUE15	NDDUJN	NDDUNA	NDUEEGF
Median	0.085	-0.191	-0.359	-0.174	0.103
$\hat{\sigma}$	1.090	1.314	0.565	0.803	1.948
Index	TPXDDVD	CSIYHYI	JPGCCOMP	NDEAGVT	NDEAMO
Median	-0.357	0.239	0.631	0.275	0.344
$\hat{\sigma}$	0.680	0.475	0.671	0.119	0.152

Table 8.1: Median 5-year relative return and the standard deviation of relative returns for all the scenarios.

Figure 8.7 shows a plot of the normalized scenarios containing the largest maximum drawdown. This can be used in risk management as well , an in this case it is clear that the Danish bonds do not experience such extreme maximum drawdowns as the other indices, actually the five year return on the scenarios for these indices are slightly negative, where the loses on the other scenarios are considerably large. It is also clear that it is the same "market" scenario that holds the largest maximum drawdown for all the indices except for NDEAGVT. This means that the indices have a high correlation in bear market as they are supposed to.

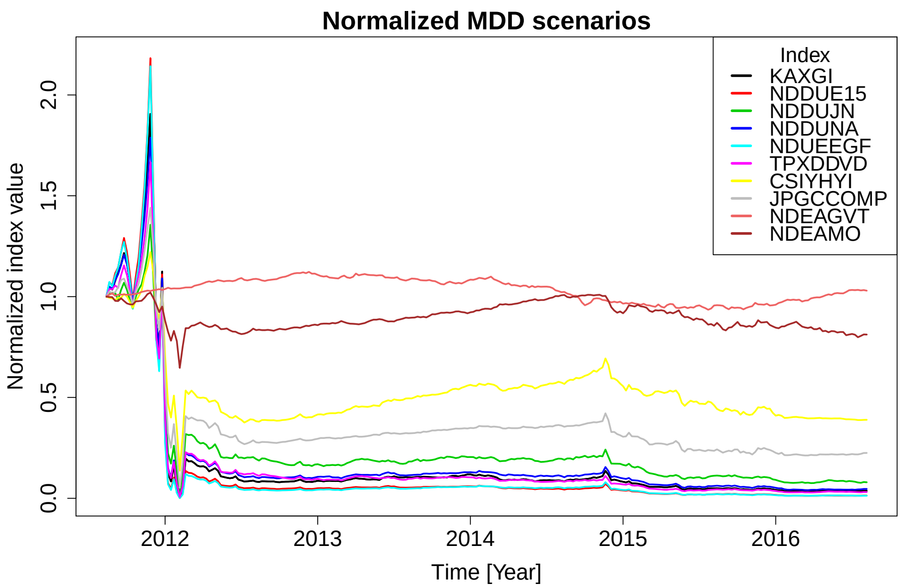


Figure 8.7: Normalized scenario with MDD for all the indices.

A plot of the drawdown and maximum drawdown from simulation start t_0 until t_τ might give a clue of how fast the maximum drawdown occur. In figure 8.8 it is seen that it occurs as early as stated above. The plot can also be used to estimate when a crisis at least would occur. For instance, let a crisis be defined as a 40% drawdown, then the median states when there is 50% chance that a crisis already has occurred. On the other way around, the chance that a crisis will hit before a given date, can also be stated looking at the quantiles. For example there is 50% chance that a crisis (or actually a drawdown equals to 40% cf. the crises definition) on the Copenhagen Stock Exchange (KAXGI) has occurred in 2014. At the same time it is more than 95% certain that there will not be a drawdown larger than 8% in NDEAGVT and 15% in NDEAMO. It is very certain ($\sim$ 95%) that the stock indices will experience a drawdown larger than 15% within the five-year period, but it is very uncertain that the bond indices will experience drawdowns at that size. This is fully consistent with the earlier statements and the exceptions for the asset classes.

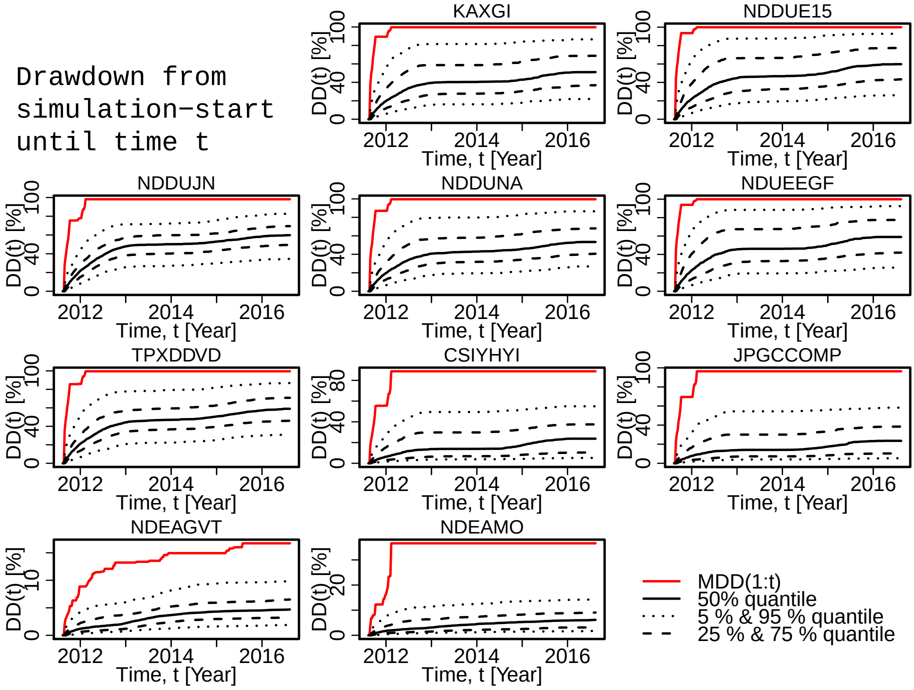


Figure 8.8: Plot of quantiles based on the drawdown for time t_0 until time t_τ for each scenario. The red line represents MDD for all the scenarios until time t_τ which is the same as the 100% quantile.

8.2 Scenario generation via bootstrapping

Bootstrapping is a method where returns are generated by sampling among historical (weekly) log returns. This way of generating scenarios is quite easy and simpler than the method above. Modelling of data is not needed, and the sampling ensures that the sizes of the returns are true, but this is not enough to ensure nice scenarios. An analysis of data is very useful in order to identify characteristics in data. As stated earlier the scenarios should represent some of these characteristics in order to be reasonable. Therefore a well-considered approach for bootstrapping is needed in order to sample such that the majority of the characteristics will be represented in the scenarios.

8.2.1 Approach for scenario generation

Bootstrapping among weekly log returns, makes it a straightforward procedure to translate the sampled values into continuous (actually discrete in time) time series for each index. The idea behind this bootstrap is to keep it simple. The approach for this bootstrapping is to sample from the whole period and not within regimes only. The sampling should start at the point where data ends, 13 August 2011. In order to make regime patterns, the sampling should not cover the whole period uniformly, but instead sample returns at time t close to returns sampled at time $t-1$. Different distributions can be used for this, and this approach uses a uniform distribution. The mean of the distribution is the week sampled from at time $t-1$, and the standard deviation is set to approximately 5 weeks. This is a very arbitrary and subjective choice, and it will of course affect the result. The bootstrapping has been tested with different standard deviations, and 5 weeks seems reasonable. The R function `ruinf` is used to generate random numbers from the continuous uniform distribution, but because the indices are discrete in time, the generated values need to be round off. The function input is the minimum α, and maximum β of possible outcomes, so the input needs to be estimated using the relationship between α, β and σ. Assuming discrete uniform distribution, the following is valid [36]:

$$
\begin{aligned}
X &\sim U(\alpha, \beta) \\
\hat{\mu}(X) &= \frac{1}{2}(\alpha + \beta) \\
\sigma^2 &= \frac{(\beta - \alpha + 1)^2 - 1}{12}
\end{aligned}
$$

In order to have $\hat{\mu}(X) = 0 \Rightarrow -\alpha = \beta$ and $\sigma = 5 \Rightarrow -\alpha = \beta = \frac{\sqrt{5^2 \cdot 12 + 1} - 1}{2} = 8.175$. Using these limits and a round-off of the result from `runif` in R, gives $\hat{\sigma} = 4.73$. Using $-\alpha = \beta = 8.65$ instead, gives $\hat{\sigma} = 4.99$. As this is a subjective chosen constant, the precision does not matter that much. In this bootstrapping 1000 scenarios are generated and the scenario length is 260 weeks = 5 years. The bootstrapping in pseudo code is:

Set time $t_i = t_{end}$, where t_{end} is the last week in data set.
N = numbers of scenarios
Simulation length = I
for n in 1 to N {
    for i in 1 to I {
    Set $t = t_{end} + 1$
    while $t > t_{end}$ or $t \leq 0$ {
        Draw an integer t from $X \sim U(\alpha, \beta)$ s.t.$\hat{\sigma}(X) = 5$ and $\hat{\mu} = t_i$
    }
    Set $t_i = t$
    Log return Scenario (I,N)=log return data(t_i)
    }
Scenario(I,N)= exp (cummulated sum of Log return Scenario (I,N)) $\cdot$ data(t_{end})
}

8.2.2 Scenarios

The scenarios from bootstrapping are found in figure 8.9, 8.10, 8.11 and 8.12. All the scenarios start with a negative trend as stated. Slowly some scenarios start to sample from earlier data, e.g. the positive regime number five, and as a consequence the median tends to flatten out, and for some indices, especially the bond indices, the trend turns positive. The quantiles are pretty smooth, and only a few scenarios take unrealistic large values. The stock indices are of course more volatile than the bond indices, and it is possible to identify scenarios with more extreme behaviours. The volatility clumping is not represented as much as desired, but the indices still seem to be highly correlated, especially within asset classes. It is not possible to identify turning points, but it possible to see regimes within a scenario, because of the uniform sampling. If the standard deviation on the sampling was larger the regimes within a scenario would have been shorter resulting in even more flat curves. It would have taken longer time to get out of a regime if the standard deviation was smaller and that would have resulted in very long negative periods to start with and maybe followed by a positive period. Among the 1000 scenarios this bootstrapping method never samples from the first half of data, therefore there is a majority of positive data, but because of the sampling method and length of the scenario, the median is not influenced by that. If the period was longer, all the median scenarios would have been positive.

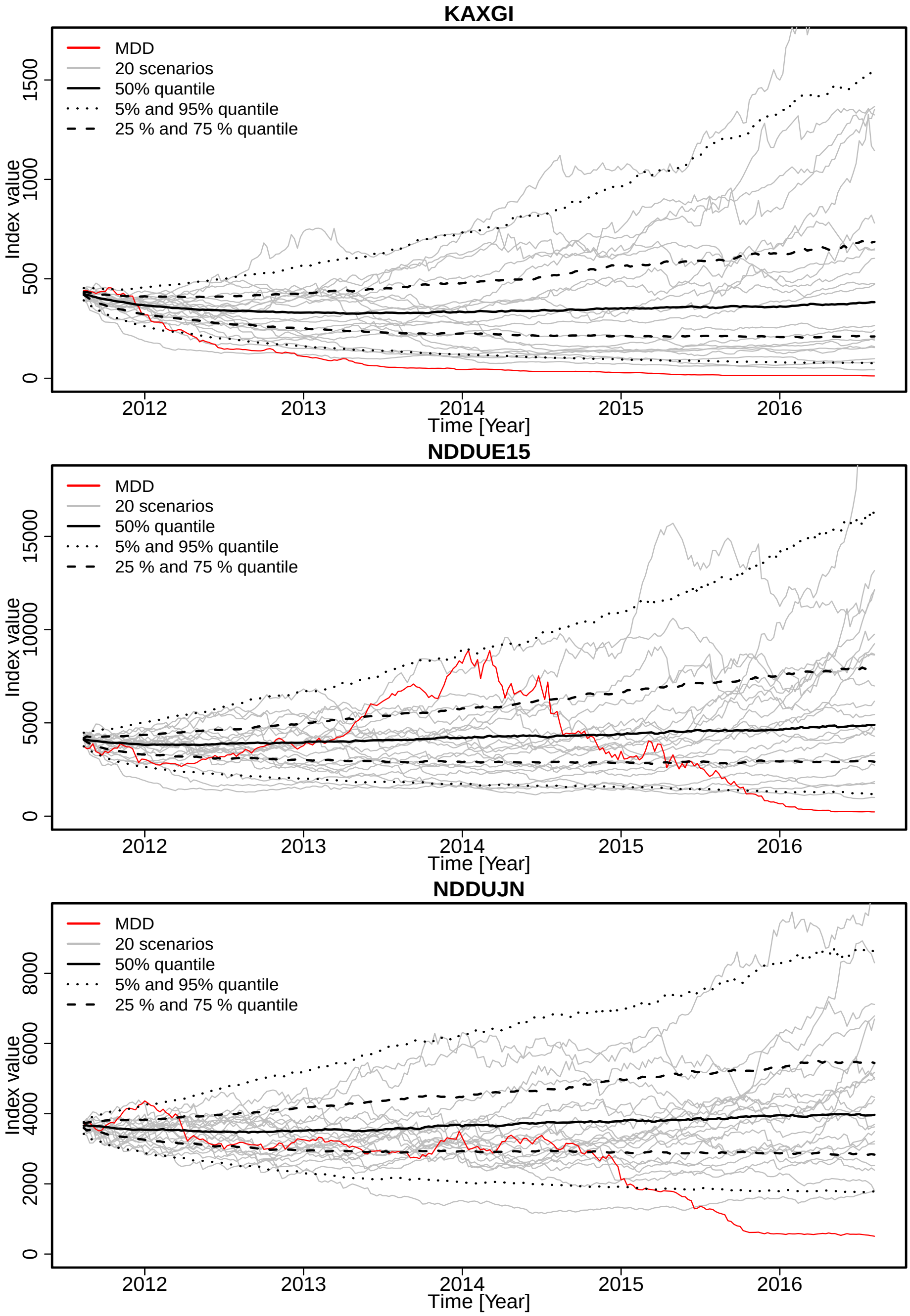


Figure 8.9: Scenarios for KAXGI, NDDUE15 and NDDUJN generated using bootstrap method.

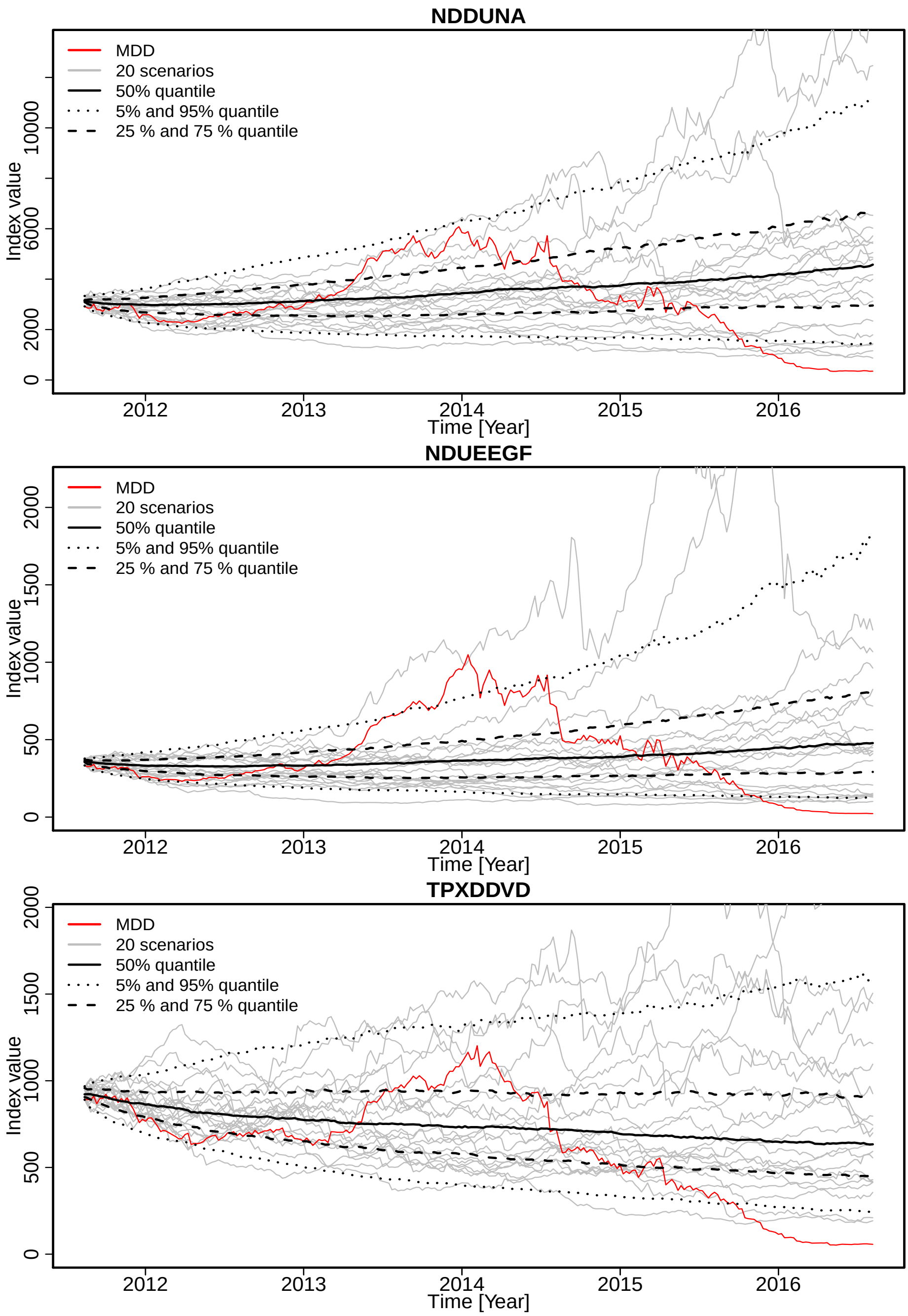


Figure 8.10: Scenarios for NDDUNA, NDUEGF and TPXDDVD generated using bootstrap method.

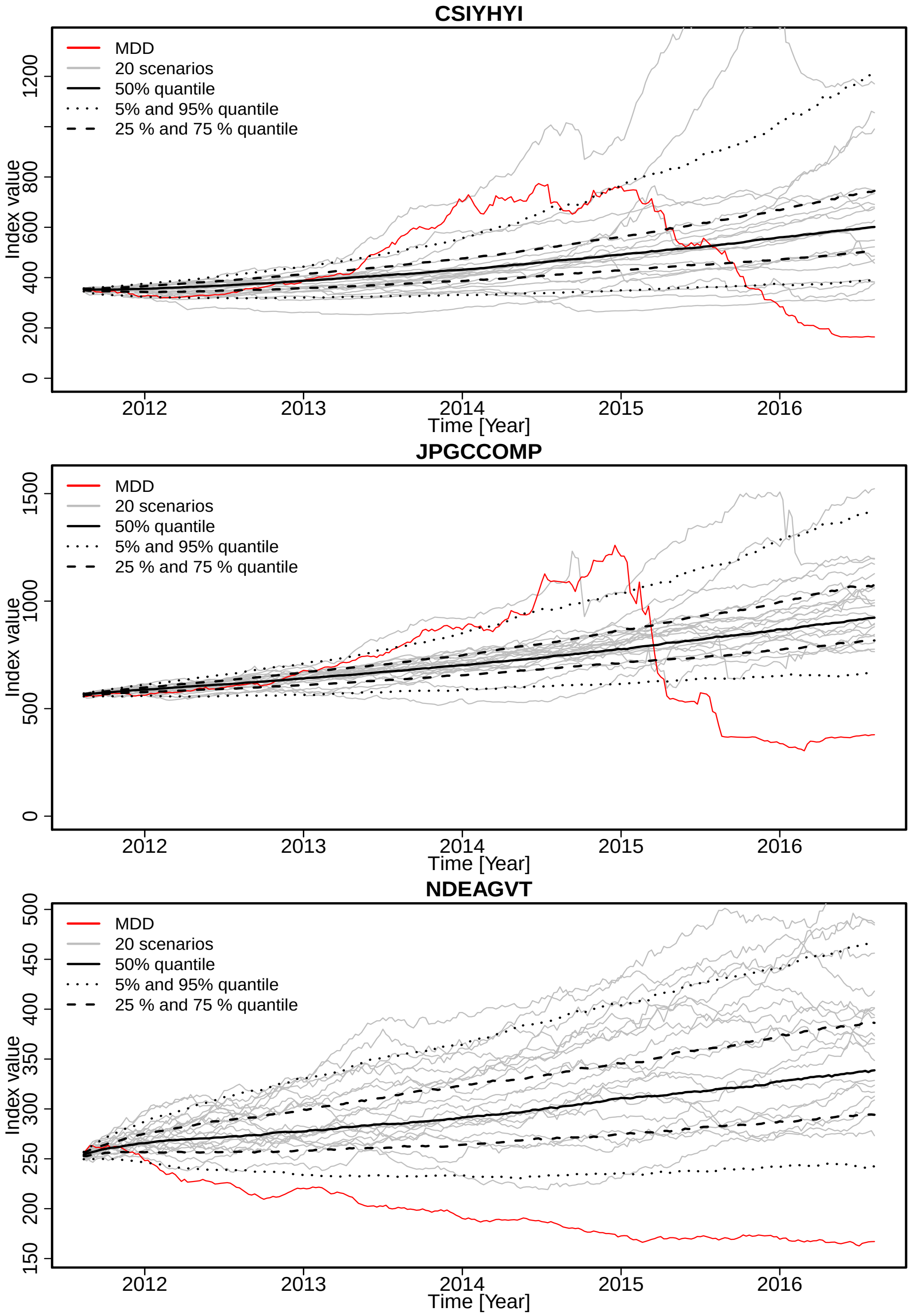


Figure 8.11: Scenarios for CSIYHYI, JPGCCOMP and NDEAGVT generated using bootstrap method.

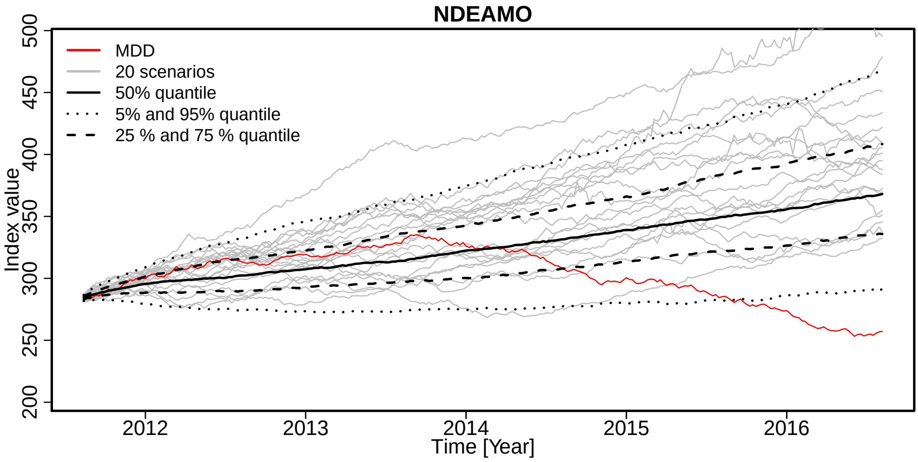


Figure 8.12: Scenario for NDEAMO generated using bootstrap method. 20 Scenarios have been plotted together with quantiles for 1000 scenarios and the scenario containing the largest MDD.

8.2.3 Analysis of scenarios

Figure 8.13 is histograms of relative changes in 5-year scenarios for each index. All the histograms are more or less positive skewed and all the indices have positive median scenario return except KAXGI and TPXDDVD. The standard deviation on the end values and the 5-year mean return are given in table 8.2. The standard deviations on the end values are largest for the stock indices and smallest for the Danish bonds. The NDDUE15 index performs the best with a 5-year return on 65.1%, the same as a 10.5 % annual rate. Again the searching for a high return results in a high risk. Again the asset allocation depends on what risk the investor is willing to accept.

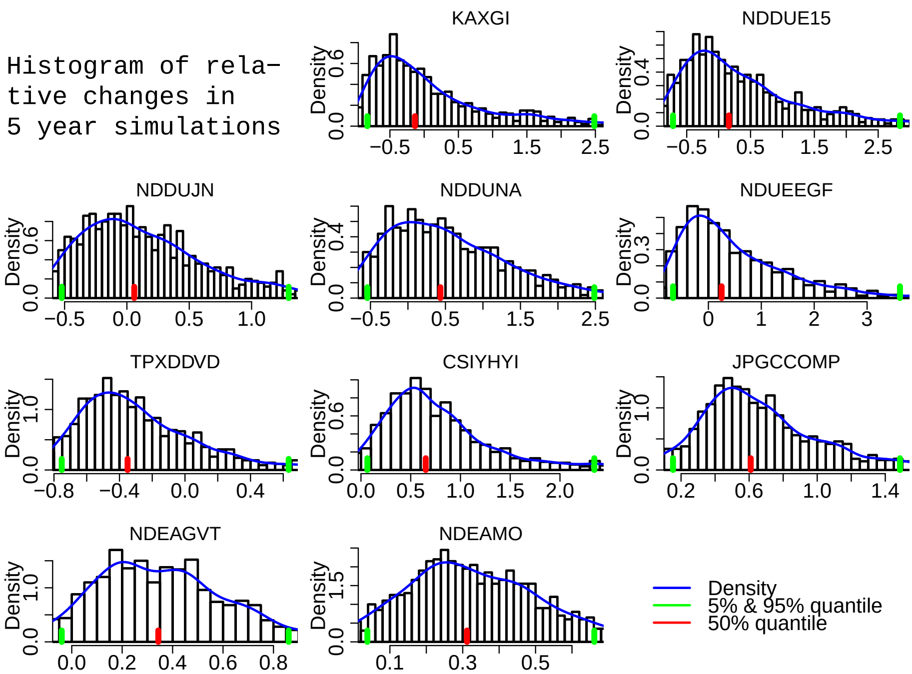


Figure 8.13: Histogram of the relative end values for the series.

Index	KAXGI	NDDUE15	NDDUJN	NDDUNA	NDUEEGF
Median	-0.134	0.160	0.059	0.433	0.250
$\hat{\sigma}$	1.192	1.357	0.630	1.117	1.974
Index	TPXDDVD	CSIYHYI	JPGCCOMP	NDEAGVT	NDEAMO
Median	-0.351	0.651	0.610	0.343	0.312
$\hat{\sigma}$	0.441	0.778	0.443	0.278	0.194

Table 8.2: Median 5-year relative return and the standard deviation of relative returns for all the scenarios.

Looking at the normalized scenarios with maximum drawdown in figure 8.14, it is seen that the maximum drawdown occur differently. Again the bond indices perform the best among the maximum drawdown scenarios.

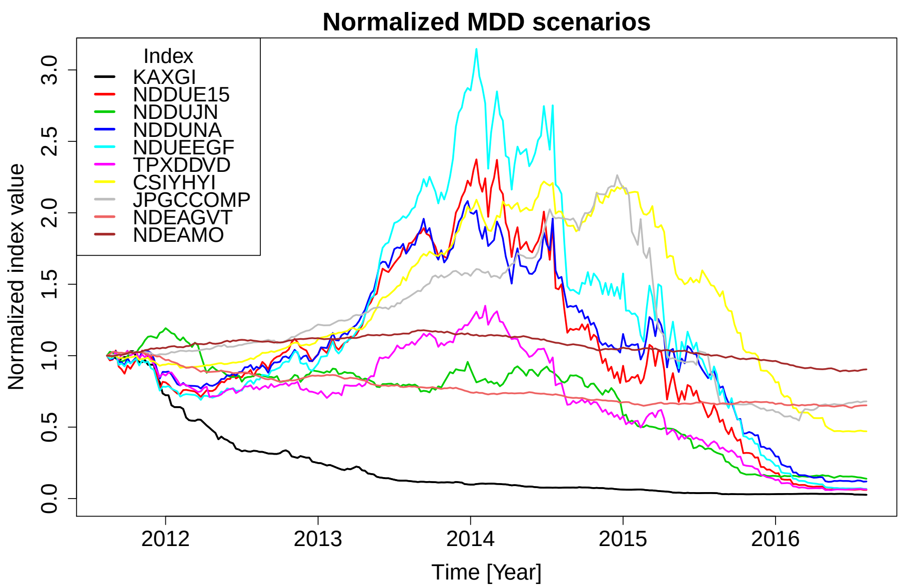


Figure 8.14: Normalized scenario with MDD for all the indices.

Figure 8.15 shows the drawdown and maximum drawdown from simulation start t_0 until t_τ. Now it takes way more time before the maximum drawdown occurs, and the possibility of getting a drawdown larger than 40% in the 5-year period is also smaller. This is caused by the uniform sampling method that needs quite a few iterations before it starts to sample form regime 4 where the largest drawdowns occur. It is not even certain that it will ever sample from periods with large drawdowns.

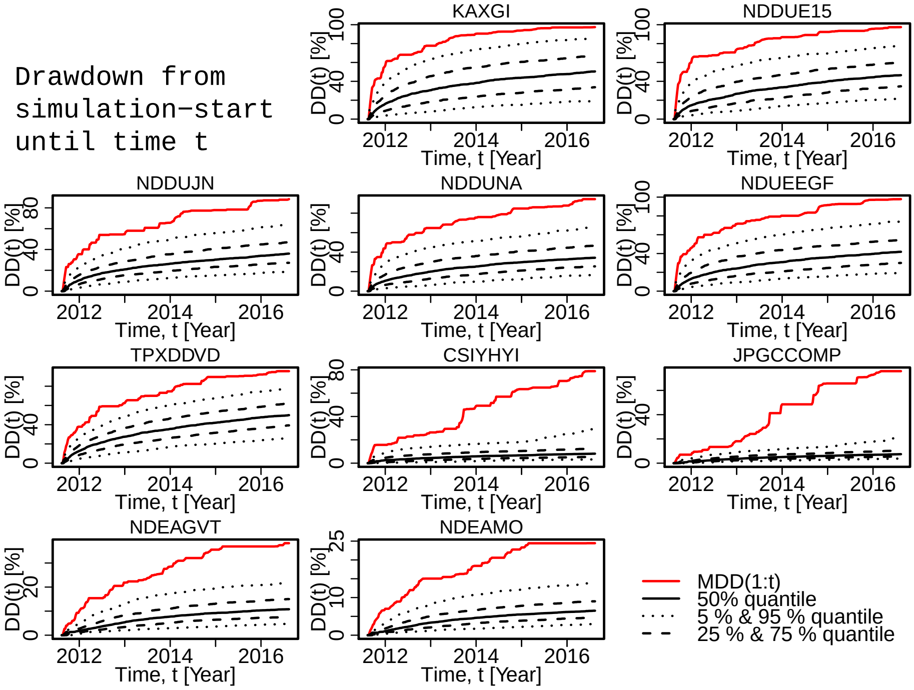


Figure 8.15: Plot of quantiles based on the drawdown for time t_0 until time t_τ for each scenario. The red line represents MDD for all the scenarios until time t_τ which is the same as the 100% quantile.

8.3 Comparing Scenario generation methods

Both methods give end distributions that seem reasonable, compared to the historical data. The standard deviations on the two methods are very similar and in general the bond indices have lower standard deviation than for the stock indices. This pattern is also represented in the scenarios through the period. The bootstrapping method gives more positive 5-year median return, and it is caused by only few scenarios having long negative periods. The shape of the scenarios are more smooth using bootstrapping, and volatility clustering is to a certain extent represented in both methods. Looking at the quantile scenarios, especially for the stock indices the ARCH methods have more non-linear quantiles. This is a result of the regime separation, and a similar pattern cannot be observed in the quantiles using bootstrapping, but on a single scenario level, periods with positive and negative trend is observable, indicating regimes. Only very extreme scenarios occur in the ARCH method, and the reason for this is obviously caused by the fact thatbootstrapping only samples among real size returns. For the same reason the size of the drawdowns is larger using ARCH methods. The sampling method in bootstrapping determines the later occurrence of the relative smaller maximum drawdown.

In the next chapter the scenarios are tested and the performance of the models is analysed.

Chapter 9

Testing scenarios

A scenario is as mentioned before, not predictions of true index values, but a possible outcome used to estimate the risk. Therefore the quality of scenarios are not a measure of how well the scenarios approximate the true value, but rather how close the distribution of uncertainty is to the true one. To get a picture of the quality of the scenarios accuracy, correctness and consistency are studied [26]. These are all quantities, that to some extent can be measured, but also require a degree of subjectivity.

9.1 Moment matching

First moment matching is considered. Moment matching is a measure of how many statistical properties are represented in the distribution of the scenarios compared to the true historical distribution.

9.1.1 Accuracy

Accuracy is a concerned with comparison of the first four moments: mean, standard deviation, skewness, kurtosis. In table 9.1 the relative annual return

Annual relative return [%]			
	Data	GARCH	Bootstrapping
KAXGI	7.20	1.65	-2.84
NDDUE15	2.53	-4.15	3.01
NDDUJN	1.16	-8.50	1.15
NDDUNA	1.34	-3.76	7.46
NDUEEGF	12.46	1.99	4.56
TPXDDVD	-1.47	-8.46	-8.28
CSIYHYI	7.24	4.37	10.55
JPGCCOMP	11.57	10.3	9.99
NDEAGVT	5.14	4.98	6.08
NDEAMO	5.90	6.09	5.58

Table 9.1: Annual relative return for the 12.6-year data period and 5-year simulation medians.

for each index is given for data and for median scenarios. The return is calculated for the whole data or simulation period. None of the methods have the same returns, but they are not incorrect. Both methods start in negative period, and therefore underestimate the returns, in particular the GARCH method. For this reason the annual return across regimes are not comparable between the simulations and the historical data, though the bond returns are performing the best for both methods, because the bonds do not experience significant losses in the first part of the scenarios. In general the higher volatility on the index the larger deviation on the annual return. Comparing the two methods, it is clear that the bootstrapping generates scenarios with a more positive trend, except for KAXGI.

It is also difficult to compare the higher order moments. The median scenario is useless because it does not behave like a real scenario. The only way is to study a single scenario at the time, but as this project focuses on the preface in scenario generation this extensive study will be left over to further research. Other test is more applicable, because they test the modelling in the preface more directly.

9.1.2 Consistency

When generating scenarios for more than one variable, the internal dependency should be maintained, e.g. the correlation matrix should be the same for data and scenarios. The GARCH method is built on keeping the correlation in the scenarios by the use of PCA, and that the returns for the indices are generated parallel, such that the added noise are the same for all the indices. The numbers of principal components and the cumulated proportion directly explain the correlation in the simulated indices. I figure 8.5 and 8.7 it is clear that the indices are highly correlated, and as stated for the historical data correlations are strongest within an asset class.

The correlation is also maintained in the bootstrapping, because the sampling of returns for all the indices is done parallel. For this reason no correlation is lost.

9.1.3 Correctness

Correctness is concerned with the properties known from historical data or theory. Both models do not allow negative index values. Regimes are represented in both methods, but the turning points in the bootstrapping are differently from scenario to scenario, where the GARCH scenarios have a overall tendency on where the turning points occur. Both methods shows volatility clustering, but the bootstrapping has a higher propensity to generate more smooth curves where the small jumps have vanished. When separating the model into regimes in the GARCH modelling, a lot of the volatility clustering have "disappeared", and therefore not represented in the model, though the use of insignificant parameters might anyway give some volatility clustering, but the correctness of these is hard to evaluate. The behaviour of the volatility in different types of period is kept. in both methods, such that the volatility is smaller in positive periods than in negative periods.

Extreme events are also a part of the correctness, because it is shows that the model not is a direct reconstruction of historical data in the future. The GARCH method generates some quite extreme scenarios, but also more realistic "high-volatility-scenarios" which is not observed in data. The bootstrapping method shows the same feature to some degree.

9.2 Stability of scenarios

The stability of the scenarios will not be tested here, but only discussed briefly as this is more interesting when optimizing scenarios. Stability of scenarios deals with stability in the sample and out of the sample.

In the sample stability for both of the method depends of the random numbers drawn. For the GARCH model a random number is drawn, in order to decide which regime to simulate from, the length and the noise on this simulation. Running enough simulation the quantile scenarios become stable, but some of the extreme events might change, and therefore also the maximum drawdown might change. Random numbers are also drawn when generating scenarios via bootstrapping. It is very unlikely that two scenarios will be identical, but if enough scenarios are generated the quantiles will remain the same. As this model does not generate such extreme scenarios as the GARCH method their change and change in maximum drawdown might not be as large.

The fact that both methods involve a degree of randomness just ensures diversity among the scenarios, which is a desirable feature.

Out of the sample stability is not considered here as it deals with stability of sampling from the "true" / benchmark distribution.

9.3 Back testing

Back testing is especially relevant for the GARCH model, as it is a measure of how well the PCA and GARCH modelling approximate historical data. The aim of using these methods is to describe the behaviour of the indices mathematically. In figure 9.1 the real-values (black line) are plotted together with the median (blue line), 5%, 25% , 75% and 95% quantiles of 1000 "reconstructions". The reconstructions are made such that values for each regimes are simulated separately with the true length of the regimes. That the reconstructions have sharp changes at the turnings points, just shows that the models for the regimes are differently as they should be. Mostly the true data is within the 25% and 75% quantiles, showing that the model is an acceptable approximation of historical data, especially considering sudden changes at certain time points only will occur in some scenarios and therefore not shown in the median.

The fact that the model performs well on the historical data it is generated from is not necessarily a quality mark for the scenarios. This is just an assumption on what happened in the past might happen in the future. This assumption is a uncertainty and error to the scenario generator. Therefor the sampling among regimes and their length makes the influence on this potential error less important, because scenarios will probably not become reconstructions of historical data when the generator has built-in randomness.

Back testing the bootstrapping method is not relevant because it weights the latest return highest, and almost no weights to the earliest data. This model

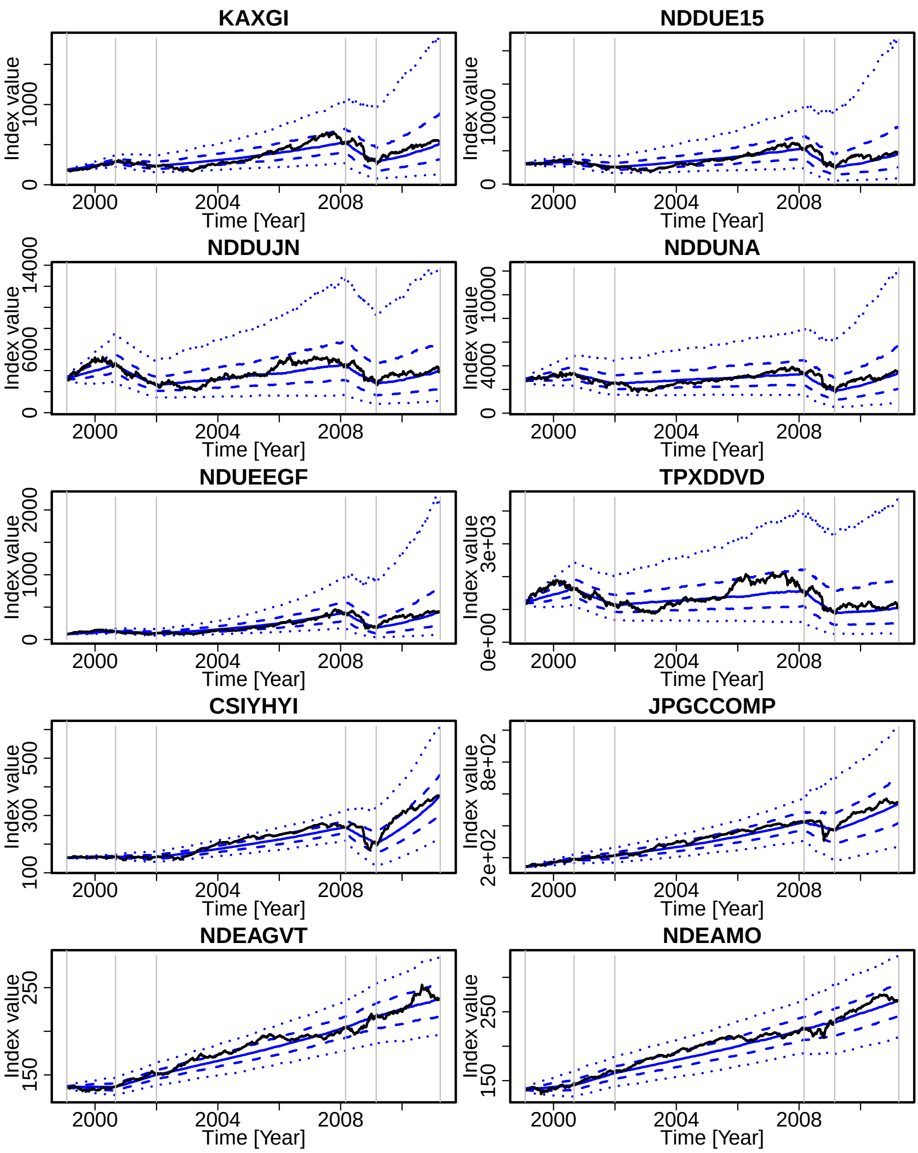


Figure 9.1: Back testing using the PCA and GARCH model. The blue line is the 50% quantile, the dashed line is the 25% and 75% quantiles and the dotted line is the 5% and 95% quantile of 1000 "reconstructions" of data, the black line.

also assumes what happened in the past will happen in the future, especially that what happened last week might very well happen again this week. This assumption does not seem unreasonable, but in order to get scenarios with behaviours not observed before, this might be a risk to the scenario generator. Adding randomness to the sampling eliminate some of this error.

9.4 Outline of scenario testing

Individual scenarios have not been tested quantitatively, instead a general expression, the quantiles, of all the scenarios have been tested. Both methods seem to have reasonable trend, and the volatility seem to have a realistic size, despite some very extreme events occurring in some GARCH scenarios. Volatility clustering can also be found, maybe not as much in historical data, but still enough to observe high volatility and low volatility periods in some scenarios. Overall regime-pattern can be seen in the GARCH scenarios, and on a single-scenarios-level also found in the bootstrapping. The scenarios seem to have an acceptable degree of correctness and consistency. The correlation between the indices is maintained in the scenarios and both models generate scenarios that are in sample stable, though some extreme events occur. Testing the GARCH model on the historical data shows that the model describes the behaviour at an acceptable level. Therefore the quality of the scenarios for both methods seem acceptable, though there are pros and cons on both methods. Testing scenarios once at the time is recommended if scenarios should be used in an optimization process.

Chapter 10

Discussion

Data has been analysed and characteristic as, non-normality, high correlation between indices, autocorrelation and conditional heteroscedastic behaviour are observed. In order to find a suitable time series model some of these characteristics have been removed by transforming data. The Danish LIBOR index has shown way different behaviour, and is for this reason leaved out from the dataset. Principal components have been used to reduce the data-space from being 10-dimensional to 4-dimensional. Random walk and ARCH(1) models (within the GARCH(p, q) family) were shown to be the most appropriate models for the 4-dimensional regime-divided data. Scenarios have been generated using a regime changing generator built on the ARCH models. These scenarios seem to behave acceptable with some of the same characteristics as observed in data. Using this scenario generator, economic turning points can be estimated, though they must be treated carefully. Scenarios have also been generated using bootstrapping with a uniform distributed sampler. These scenarios also seem to have a fine quality, though less volatility clustering is represented.

The characteristics of the data are not surprising and expected from similar studies[27]. It is natural that both autocorrelation and cross correlation are found in the indices. The volatility clustering is mostly known from "high-risk" indices, and for this reason it is easiest to observe in the stock indices. The positive correlation between stock and bond indices was also expected because of the low inflation and interest rate they are not seen as alternative investment opportunities. The Danish LIBOR index behaves differently, and that is the reason why it is left out from the data, because the PCA is a preferred method to reduce the dimension of the data space. LIBOR data was not available for the whole period, and therefore would have resulted in scenarios built on other terms. The LIBOR index could have been modelled separately, but again this would have given scenarios with other conditions. Nor is it possible to invest in the LIBOR directly, and therefore it is acceptable to leave it out from data. The use of weekly log return has the great feature that it makes the data independent, such that the modelling is done more easily. Of course some daily variance vanishes, but as scenarios are 5 years, this small variance is trivial. In order to get as realistic variance as possible Friday sample is used to represent the week sample.

That data behaves differently within periods is also observed. In order to reproduce a likely behaviour in the scenarios the data is spitted in to regimes using OECD CLI turning points. The CLI turning points is a representation of the economics changes and not financial changes. This might be a little error not to use the financial turning points because they are often occurring a few month earlier. This might have changed the size of the parameters in the ARCH model and maybe also the model order. The reason why a lot of the models should have been random walks with a drift, is probably caused by the splitting into regimes. Within a regime the volatility is fairly constant, and because of the length of the regime the conditional heteroscedasticity is not statistical significant in many of the ARCH models. But the variance is still conditional, because it depends on which regime is sampled.

Knowing that the indices are highly correlated, the use of PCA is obvious. Using four PC's explain around 80% of the variation within all the indices. The data derived from the PC's is used in scenarios and not precise predictions of index values, therefore it is reasonable that not all of the variation is explained. PCA is only used to describe the relation of indices, and therefore it is acceptable that not all the requirements for the use of PCA are met. From the communalities in table 7.6, the CSIYHYI had the lowest communalities overall, but looking at the back testing of the index in 9.1, the model for CSIYHYI actually perform nicely within the quantile of reconstructions. There are significant lags of autocorrelation left in the weekly log return of CSIYHYI, and that might disturb the PCA. It does not seem to be any strange behaviour in the back testing, so the scenarios might be as good as for the other indices.

It is a bit surprising that it turns out, that a lot of the series from the PC's can be describes as random walks with drift. The conditional heteroscedastic behaviour has been observed, but modelling regimes separately the volatility clustering becomes statistically insignificant. This results in a use of ARCH(1) parameter that, from a statistical point of view, cannot be justified, but their relative weightings are small and it does not have any considerable influence to the result. The models where both parameters, α_0 and α_1 are significant are often the models derived from the first PC's for a regime. The reason for this is that the first PC's have the highest variance, and therefore the conditional heteroscedastic behaviour is more distinct. The choice of model has been a rather subjective process, but statistical test of the residuals has been used in order choose a proper model. The residuals in the GARCH model should ideally be normally distributed. After testing the residuals for normality for some models, the residuals are proved not to be normal. This might indicate that the models used do not catch all the patterns in data. Known from the data analysis, a few outliers are represented, and they might very well cause the result of the test. Only a few non-normal residuals are enough to makes the test fail.

The different behaviour of the variance in negative and positive period has, as already mentioned, vanished to some extent. A limitation in an ARCH model is that it does not count for the sign if the index change /shock, for this reason EGARCH models might have been used if data was not divided into regimes. ARCH models also over estimates the variance if a large index change occurs in non-volatile periods. The model definition is also the reason why some very extreme occur, because when the model gets a high variance, the next variance is also likely to be large. However this have only limited influence in this project because of the small α_1 parameters.

The choice of the period lengths only depends on the observed data. It might have been an idea to use other estimates of the regime lengths. For instance the length of the regimes before 1999 could have been taking into account. It is obvious that the third regime has a very large positive on the trend of the scenarios because of its length. Using other regime lengths could have resulted in more equal period lengths. On the other hand, the variance of data in a period might quite well be described by the length of the period. Again the question of how much the simulations should reflect the historical data is brought up. For short-period simulations a high degree of the scenarios should be able to detect in recent history. If the scenarios are longer they should cover a larger amount of unobserved events. The use of historical data in both scenario generation techniques ensures a realistic variance, but will not generate new events. To ensure new events, randomness is “added” to both methods. The GARCH model samples among the periods and afterwards among their length. The bootstrapping samples uniformly around “yesterdays” return. The use of historical data should also be limited, because a lot have changed in the finical market through recent decades. It is for this reason quite unlikely some pattern will show up again and it might be more likely that new events occur. The combination of historical data and methods that takes new events into account is nicely represented in the GARCH method when generating 5-year scenarios.

The bootstrapping is more applicable when the scenarios are shorter. It has the disadvantage of generating scenarios with too short "true" length of the negative periods and therefore it over performs when the time horizon gets too long.

This project is done in collaboration with Peter Nystrup. He has generated scenarios based on slightly different methods [22]. He has chosen to paste all the positive and negative regimes into two regimes. This result in more significant ARCH(1) parameters, implying conditional heteroscedasticity is modelled. Pasting the scenarios together also has a disadvantage, because the model represents a smoothing of the periods. Therefore the scenarios are also more smooth, though extreme events and scenarios also occur. The trend of the medians in the bond indices are positive in both models but the trend in the median scenario in Peter's stock indices are negative where those in this thesis depends on the index. The distribution of the end values are quite equal for the bond indices, with equal variation. The distribution of the end values in the stock indices are more positive skewed in Peter's model, caused by more "loss scenarios" than the end values using the model in this thesis. Peter's bootstrapping method uses a normal distribution as sampling parameter instead of a uniform. The scenarios are very equal and the distributions as well and therefore it is hard to tell which method preform the best.

To summarize all generation techniques generate realistic scenarios, the GARCH method seem to generate more realistic scenarios with the "right" behaviour, but it is also this method that generates a couple very extreme events, however this does not matter because in risk management extreme events are used as a tool in the optimization process. Adjusting and improvement of the GARCH and bootstrapping model are discussed in chapter 12.

Chapter 11

Conclusion

Eleven indices representing three asset classes have been analysed from a statistical point of view. A few outliers have been identified, and a few errors in data set have been corrected, but the rest of the outliers are unchanges. At first index values are studied, and characteristic such as autocorrelation, non-constant volatility and high correlation are observed. Especially correlations within an asset class are high. It is concluded that the index values are non-normal distributed, and an Augmented Dickey Fuller-test states that the index values might follow a random walk. Afterwards log return index values are analysed, and because they still have autocorrelation, weekly log return is used in order to get independent data. Despite a few significant lags in CSIY-HYI, significant autocorrelation has been removed. The standard deviation on the weekly log returns are reconstructed recursively, and it is again stated that there is conditional heteroscedastic behaviour in the indices. A plot of the cross correlations shows that weekly log return indices are correlated, and some indices are cross correlated, especially CSIYHYI has many significant lags to the stock indices and JPGCCOMP. The Danish bond indices are highly correlated with each other, but not as much with the other indices. The Danish LIBOR index, DK00S.N.Index, has no significant correlation to the other indices, and is therefore left out in the modelling and scenario generation. Weekly log return indices are weakly negative skewed, the excess kurtosis is positive and it cannot be rejected that they are stationary. Normality cannot be proved, but it is assumed in the further modelling. In the data analysis it is observed that the behaviour of the volatility depends on the trend in the index prices. Therefore data is spitted into regimes defined by OECD CLI turning points. Each regime is transformed using PCA, and it is shown that using four principal components explains at least 76% of the total variance and in some regimes even more. A reconstruction of data using four principal components shows that there is a minor loss, and for this reason four principal components is used in the further modelling, where GARCH models are used to model the data derived from the principal components, that is the dynamic behavior of the variance. It turns out that the ARCH(1) models and random walks with a drift are the best fit for data. This is a bit surprising because conditional heteroscedasticy is not represented in a random walk, but on the other hand, a test stated that it could not be rejected that data follows a random walk. ARCH(1) models, and in one single case ARCH(3) models are used to model principal component data. Often the parameters controlling the conditional heteroscedastic behaviour, in the models that should have been random walks from a statistical point of view, are so small it that might vanish when scenarios are generated. The model is tested on the real data, at it performs well, often the residuals are accepted as independent and normal distributed. Scenarios are generated where the type of regime is picked at random, and its length is sampled from an Erlang distribution with mean equal the period length picked. This procedure continues until the duration of the scenario is at least 5 years. This method generates scenarios that behave acceptable, both volatility clustering, autocorrelation and cross correlation is represented together with extreme events. The scenarios generated with this method show an overall pattern that gives indication on economic turning points.

Bootstrapping is also used to generate scenarios. These scenarios also behave reasonable, but the method performs better if the scenarios have a shorter time horizon.

The GARCH method generates scenarios with a higher degree of diversity, but it is not a negative quality in this case. Both methods generate scenarios that state the risk on the asset reasonably, and because of that the scenarios can be used in risk management which is a part of the asset allocation process. The GARCH method should be preferred when generating scenarios with a longer time horizon.

Chapter 12

Further Research

The subject of modelling financial indices is very comprehensive. Many other models might be used to model the indices, and many other techniques for generation scenarios exist. Through this project ideas have arisen, and a few have also been tried out and thrown away again. Some of the more interesting untested ideas will be presented in this chapter as suggestions for further research. There are two main suggestions, one concerning the modelling and one concerning test and use of scenarios.

It would be interesting to place the turning points more subjectively in order to see the changes in the model and the scenarios. This could also be used in a more profound stability test of the scenarios. As there are only five regimes in the data set the length of the regimes might not be representative, and therefor historical period lengths might be implemented. It would also be interesting to see if the GARCH model is estimated on weekly log return data will have significant ARCH parameters. The modelling of data can be done with other time series models, e.g. other variation of the GARCH model. Assuming conditional heteroscedasticity, models like EGARCH, GJR-GARCH and TGARCH could be tested. These models allow volatility shocks to reach differently on positive and negative input, and there are no restrictions in the parameters on the EGARCH. The models might be applied directly on single indices or to PCA data, and maybe leaving the subdivision of regimes out.

Concerning the bootstrapping it would have been interesting to implement the regime part from the GARCH model, where a regime is chosen randomly, and afterwards the length. This might remove the tendency of overestimation, and sampling to much among the same values. Volatility clustering might also very well be represented in such scenarios.

The scenarios could also be studied more carefully, in order to optimise and finish the asset allocation step with recommendations on allocation within this two asset classes. For this purpose scenarios should be tested closely, e.g. by changing parameter in the GARCH model or analysing some of the scenarios one at the time. This could be a whole project in itself.

Other asset classes could also be implement such that the asset allocation becomes more comprehensive and versatile.

Appendix A

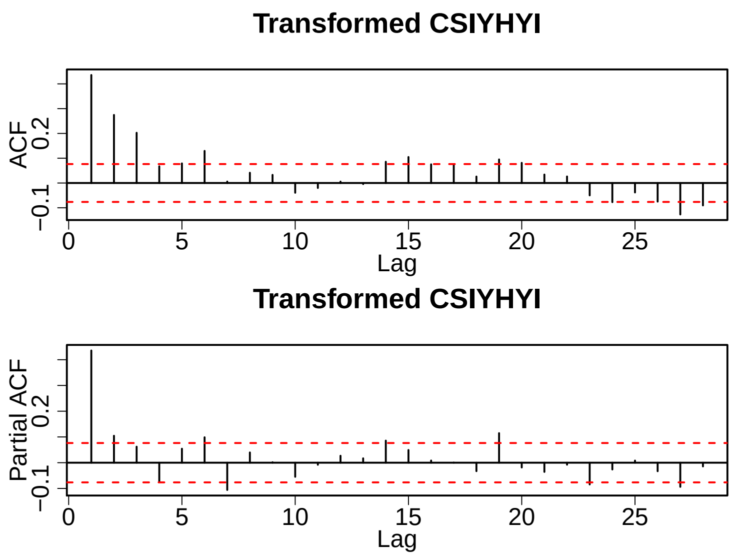


Figure A.1: ACF and PACF in weekly square root simple gross return of CSIYHYI.

Appendix B

R-script for analysis af data in chapter 3 and 4

```
##################################################################
## Filename: Analysis of data.R   Date: 20-01-2012             ##
## Author: Emil Ahlmann Østergaard s082632                      ##
## Description: R-script for chapter 3+4+5 in B.Sc. thesis:     ##
## "Scenario gerneration for financial market indices"          ##
##################################################################

setwd("C:/Users/Emil/Documents/Skole/Bachelorprojekt")
library('fGarch')
library('tseries')
library('TSA')
library('chron')
library('fields')
library('gplots')
library('SDMTools')

#Reading in data
data<-read.csv2('Data endelig version.csv',header=T)
data<-data[,1:12]
```

```
data$Date<-as.Date(data$Date,"%d-%m-%Y")
data$DK00S.N.Index[1:1161]<-NA
data$DK00S.N.Index<-as.numeric(data$DK00S.N.Index)
attach(data)

#Plot of data
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/data_single_plot.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(2,2.2,2,1.5),mgp=c(3,1,0),cex.axis=1.6,
cex.main=1.8)
textplot('Indexsplot')
for(i in 2:11){
plot(Date,data[,i],type='l',lwd=2,main=names(data)[i],)
axis.Date(1,at=seq(as.Date("1999/1/1"),as.Date("2011/1/1"),"years")
,labels=F)
}
plot(Date,data[,12],type='l',lwd=2,main=names(data)[12])
axis.Date(1,at=seq(as.Date("1999/1/1"),as.Date("2011/1/1"),"years")
,labels=F)
dev.off()

#Generating data for NDUEEGF
#end_month is where a month ends.
original.data.NDUEEGF=data$NDUEEGF
set.seed(200)
end_month=c(1,21,41,64,86,107,129,151,173,195,216,238,261,
282,303,326,346,369,391,412,435,456,478,500,521)
b=approx(x=end_month,y=data$NDUEEGF[end_month],n=521)
data$NDUEEGF[1:520]=b$y[1:520]+rnorm(520,mean=0,
sd=sd(diff(data$NDUEEGF[end_month-1]))/(sqrt(22)))

#plot of new data for NDUEEGF
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/new_data_NDUEEGF.pdf",width=8,height=5)
par(mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0),cex.axis=1.6,cex.lab=1.6,
cex.main=1.8,lwd=2)
plot(Date[1:521],original.data.NDUEEGF[1:521],type='l',
main='NDUEEGF',xlab='Time [Year]',ylab='Index value')
axis.Date(1,at=seq(as.Date("1999/1/1"),as.Date("2011/1/1"),
"months"),labels=F)
lines(Date[1:520],data$NDUEEGF[1:520],col='blue')
lines(Date[1:521],b$y,col='red',lwd=2)
legend('topleft',legend=c('Original data','Linear interpolation',
'New data'),col=c('black','red','blue'),bty='n',cex=1.4,
```

```
lwd=c(2,2,2))
dev.off()

#KPSS-test for stationarity in data
index.number=2
kpss.test(data[,index.number])

#Augmented Dickey-Fuller test for unit root
adf.test(na.omit(data[,index.number]))

#Shapiro-Wilk test for normal distributed data
shapiro.test(data[,index.number])

#Plot of autocorrelation in data
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/ACF_data_plot.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(3.3,3.5,1.5,0.5),mgp=c(2,0.8,0),
cex.axis=1.6, cex.main=1.8, cex.lab =1.6,lwd=2)
textplot('ACF',valign="top",cex=5)
for(i in 2:12){
acf(na.omit(data[,i]),main='',lag.max=15,ylim=c(0,1),ci.col='red')
mtext(names(data)[i],3,line=0.2)
}
dev.off()

#Plot of partial autocorrelation in data
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/PACF_data_plot.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(3.3,3.5,1.5,0.5),mgp=c(2,0.8,0),
cex.axis=1.6, cex.main=1.8, cex.lab =1.6,lwd=2)
textplot('PACF',valign="top",cex=5)
for(i in 2:12){
pacf(na.omit(data[,i]),main='',lag.max=15,ylim=c(0,0.3),
ci.col='red')
mtext(names(data)[i],3,line=0.2)
}
dev.off()

#Log return data
logr_data<-apply(log(data[,2:12]),2,diff)
logr_data<-rbind(c(rep(0,10),0),logr_data)

# Mean of index in log return space
index_number=1 #number in 1:11
```

```
#Plot of log return data
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/log_return_data_single_plot.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(2,2.2,2,1.5),mgp=c(3,1,0),cex.axis=1.6,
cex.main=1.8)
textplot('logR indexsplot')
for(i in 1:11){
plot(Date,logr_data[,i],type='l',lwd=1,main=names(data)[i+1],)
axis.Date(1,at=seq(as.Date("1999/1/1"),as.Date("2011/1/1"),"years"),
labels=F)
}
dev.off()

#Plot of autocorrelation in log return data
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/ACF_log_return_data_plot.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(3.3,3.5,1.5,0.5),mgp=c(2,0.8,0),
cex.axis=1.6, cex.main=1.8, cex.lab =1.6,lwd=2)
textplot('ACF',valign="top",cex=5)
for(i in 1:11){
acf(na.omit(logr_data[2:3291,i]),main='',ci.col='red')
mtext(names(data)[i+1],3,line=0.2)
}
dev.off()

#Plot of partial autocorrelation in log return data
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/PACF_log_return_data_plot.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(3.3,3.5,1.5,0.5),mgp=c(2,0.8,0),
cex.axis=1.6, cex.main=1.8, cex.lab =1.6,lwd=2)
textplot('PACF',valign="top",cex=5)
for(i in 1:11){
pacf(na.omit(logr_data[2:3291,i]),main='',ci.col='red')
mtext(names(data)[i+1],3,line=0.2)
}
dev.off()

#Weekly (friday) sample
data_week=data[seq(1,length(data[,1]),5),]

#Weekly (friday) log return sample
logr_data_week=apply(log(data_week[,2:12]),2,diff)
logr_data_week=rbind(c(rep(0,10),NA),logr_data_week)
```

```
#Recursive estimation of mean weekly log return data
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/recursiv_estimat_of_MEAN_week_logR_data.pdf",width=8,
height=6)
par(mfrow=c(4,3),mar=c(2,3,1.5,1.5),mgp=c(2,0.8,0),cex.axis=1.6,
cex.main=1.8, cex.lab =1.6,lwd=2)
n=length(logr_data_week[,1])
textplot("Recursiv estimation\nof mean in weekly log
\nreturn with \nlambda=0.90",valign="top")
for (i in 1:10){
rvar=numeric(n)
for (j in 1:n){
w=0.90^((j-1):0)
rvar[j]=wt.mean(logr_data_week[1:j,i],w)
}
plot(data_week[,1],rvar,type='l',lwd=2,main=names(data)[i+1],
xlab='',ylab='')
axis.Date(1,at=seq(as.Date("1999/1/1"),as.Date("2011/1/1"),
"years"),labels=F)
}
for (j in 1:length(logr_data_week[,11])){
w=0.90^((j-1):0)
rvar[j]=wt.mean(logr_data_week[1:j,11],w)
}
plot(data_week[,1],rvar,type='l',lwd=2,main=names(data)[12]
,xlab='',ylab='')
axis.Date(1,at=seq(as.Date("1999/1/1"),as.Date("2011/1/1")
,"years"),labels=F)
dev.off()


#Plot of autocorrelation in weekly log return data
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/ACF_log_return_weekly_data_plot.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(3.3,3.5,1.5,0.5),mgp=c(2,0.8,0),
cex.axis=1.6, cex.main=1.8, cex.lab =1.6,lwd=2)
textplot('ACF',valign="top",cex=5)
for(i in 1:11){
acf(na.omit(logr_data_week[2:659,i]),main='',ci.col='red')
mtext(names(data)[i+1],3,line=0.2)
}
dev.off()
```

```
#Plot of partial autocorrelation in weekly log return data
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/PACF_log_return_weekly_data_plot.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(3.3,3.5,1.5,0.5),mgp=c(2,0.8,0),
cex.axis=1.6, cex.main=1.8, cex.lab =1.6,lwd=2)
textplot('PACF',valign="top",cex=5)
for(i in 1:11){
pacf(na.omit(logr_data_week[2:659,i]),main='',ci.col='red')
mtext(names(data)[i+1],3,line=0.2)
}
dev.off()

#Weekly (friday) log return (index 2:7 and 9:12)
#and sqrt return (index 8) sample
transformed_data_week=array(logr_data_week,c(659,11))
transformed_data_week[1,7]=1
for (i in 1:(length(data_week[,8])-1)){
transformed_data_week[i+1,7]=sqrt(data_week[i+1,8]/data_week[i,8])
}

#Plot of ACF and PACF in weekly transformed CSIYHYI data
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/ACF_and_PACF_transformed_return_weekly_CSIYHYI_plot.pdf",
width=8,height=6)
par(mfrow=c(2,1),mar=c(3.3,3.5,3.5,0.5),mgp=c(2,0.8,0),
cex.axis=1.6, cex.main=1.8, cex.lab =1.6,lwd=2)
acf(na.omit(transformed_data_week[2:659,7]),
main='Transformed CSIYHYI',ci.col='red')
pacf(na.omit(transformed_data_week[2:659,7]),
main='Transformed CSIYHYI',ci.col='red')
dev.off()

# Recursive estimation of SD in weekly log return data
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/recursiv_estimat_of_SD_week_logR_data.pdf",width=8,
height=6)
par(mfrow=c(4,3),mar=c(2,3,1.5,1.5),mgp=c(2,0.8,0),
cex.axis=1.6, cex.main=1.8, cex.lab =1.6,lwd=2)
n=length(logr_data_week[,1])
textplot("Recursiv estimation\nof SD in weekly log
\nreturn with \nlambda=0.90",valign="top")
for (i in 1:10){
rvar=numeric(n)
for (j in 1:n){
```

```
w=0.90^((j-1):0)
rvar[j]=wt.sd(logr_data_week[1:j,i],w)
}
plot(data_week[,1],rvar,type='l',lwd=2,main=names(data)[i+1],
xlab='',ylab='')
axis.Date(1,at=seq(as.Date("1999/1/1"),as.Date("2011/1/1"),
"years"),labels=F)
}
for (j in 1:length(logr_data_week[,11])){
w=0.90^((j-1):0)
rvar[j]=wt.sd(logr_data_week[1:j,11],w)
}
plot(data_week[,1],rvar,type='l',lwd=2,main=names(data)[12],
xlab='',ylab='')
axis.Date(1,at=seq(as.Date("1999/1/1"),as.Date("2011/1/1"),
"years"),labels=F)
dev.off()

########
lagmax=4
#CCF & pairsplot in weekly log return
lagmax=4
yl=0.2
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/CCF_logR-weekly_data.pdf",width=8,height=8)
par(mfrow=c(11,6),mar=c(0,2,2,0),mgp=c(3,0.6,0),cex.axis=1.4,
cex.main=1, cex.lab=1.3,lwd=1)
par(mar=c(0,0.2,1.7,0))
textplot(names(data[2]),valign="center",cex=1.7)
par(mar=c(0,0.2,1.7,0))
ccf(logr_data_week[,1],logr_data_week[,2],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
axis(3, at = c(-lagmax,-1,0,1,lagmax), labels = TRUE, tick = TRUE,)
par(mar=c(0,0.2,1.7,0))
ccf(logr_data_week[,1],logr_data_week[,3],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,1.7,0))
ccf(logr_data_week[,1],logr_data_week[,4],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,1.7,0))
axis(3, at = c(-lagmax,-1,0,1,lagmax), labels = TRUE, tick = TRUE,)
ccf(logr_data_week[,1],logr_data_week[,5],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,1.7,2))
```

```
ccf(logr_data_week[,1],logr_data_week[,6],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
axis(3, at = c(-lagmax,-1,0,1,lagmax), labels = TRUE, tick = TRUE,)
par(mar=c(0,2,0.2,0))
plot(logr_data_week[,2],logr_data_week[,1],xlab='',ylab='',
main='',xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
textplot(names(data[3]),valign='top',cex=1.6)
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,2],logr_data_week[,3],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,2],logr_data_week[,4],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,2],logr_data_week[,5],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,2))
ccf(logr_data_week[,2],logr_data_week[,6],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
axis(4, at = c(-yl,yl), labels = c(-yl,yl), tick = TRUE,)
par(mar=c(0,2,0.2,0))
plot(logr_data_week[,3],logr_data_week[,1],xlab='',ylab='',
main='',xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,3],logr_data_week[,2],xlab='',ylab='',
main='',xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
textplot(names(data[4]),valign='top',cex=1.6)
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,3],logr_data_week[,4],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,3],logr_data_week[,5],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,2))
ccf(logr_data_week[,3],logr_data_week[,6],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,2,0.2,0))
plot(logr_data_week[,4],logr_data_week[,1],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,4],logr_data_week[,2],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
```

```
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,4],logr_data_week[,3],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
textplot(names(data[5]),valign='top',cex=1.6)
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,4],logr_data_week[,5],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,2))
ccf(logr_data_week[,4],logr_data_week[,6],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
axis(4, at = c(-yl,0,yl), labels = TRUE, tick = TRUE,)
par(mar=c(0,2,0.2,0))
plot(logr_data_week[,5],logr_data_week[,1],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,5],logr_data_week[,2],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,5],logr_data_week[,3],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,5],logr_data_week[,4],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
textplot(names(data[6]),valign='top',cex=1.6)
par(mar=c(0,0.2,0.2,2))
ccf(logr_data_week[,5],logr_data_week[,6],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,2,0.2,0))
plot(logr_data_week[,6],logr_data_week[,1],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,6],logr_data_week[,2],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,6],logr_data_week[,3],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,6],logr_data_week[,4],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,6],logr_data_week[,5],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
```

```
par(mar=c(0,0.2,0.2,2))
textplot(names(data[7]),valign='top',cex=1.6)
par(mar=c(0,2,0.2,0))
plot(logr_data_week[,7],logr_data_week[,1],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,7],logr_data_week[,2],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,7],logr_data_week[,3],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,7],logr_data_week[,4],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,7],logr_data_week[,5],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,2))
plot(logr_data_week[,7],logr_data_week[,6],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,2,0.2,0))
plot(logr_data_week[,8],logr_data_week[,1],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,8],logr_data_week[,2],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,8],logr_data_week[,3],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,8],logr_data_week[,4],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,8],logr_data_week[,5],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,2))
plot(logr_data_week[,8],logr_data_week[,6],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,2,0.2,0))
plot(logr_data_week[,9],logr_data_week[,1],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,9],logr_data_week[,2],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
```

```
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,9],logr_data_week[,3],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,9],logr_data_week[,4],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,9],logr_data_week[,5],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,2))
plot(logr_data_week[,9],logr_data_week[,6],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,2,0.2,0))
plot(logr_data_week[,10],logr_data_week[,1],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,10],logr_data_week[,2],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,10],logr_data_week[,3],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,10],logr_data_week[,4],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,10],logr_data_week[,5],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,2))
plot(logr_data_week[,10],logr_data_week[,6],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0.2,2,0.2,0))
plot(logr_data_week[,11],logr_data_week[,1],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0.2,0.2,0.2,0))
plot(logr_data_week[,11],logr_data_week[,2],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0.2,0.2,0.2,0))
plot(logr_data_week[,11],logr_data_week[,3],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0.2,0.2,0.2,0))
plot(logr_data_week[,11],logr_data_week[,4],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
par(mar=c(0.2,0.2,0.2,0))
plot(logr_data_week[,11],logr_data_week[,5],xlab='',ylab='',main='',
```

```
xaxt='n',yaxt='n')
par(mar=c(0.2,0.2,0.2,2))
plot(logr_data_week[,10],logr_data_week[,6],xlab='',ylab='',main='',
xaxt='n',yaxt='n')
dev.off()

lagmax=4
#CCF in logR and weekly logR
lagmax=4
yl=0.2
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/CCF_logR-weekly_data_2.pdf",width=8,height=8)
par(mfrow=c(11,6),mar=c(0,2,2,0),mgp=c(3,0.6,0),cex.axis=1.4,
cex.main=1, cex.lab=1.3,lwd=1)
par(mar=c(0,2,1.7,0))
ccf(logr_data_week[,1],logr_data_week[,7],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,1.7,0))
axis(3, at = c(-lagmax,-1,0,1,lagmax), labels = TRUE, tick = TRUE,)
ccf(logr_data_week[,1],logr_data_week[,8],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,1.7,0))
ccf(logr_data_week[,1],logr_data_week[,9],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,1.7,0))
axis(3, at = c(-lagmax,-1,0,1,lagmax), labels = TRUE, tick = TRUE,)
ccf(logr_data_week[,1],logr_data_week[,10],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,1.7,0))
ccf(logr_data_week[,1],na.omit(logr_data_week[,11]),lag.max=lagmax,
xlab='',ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,1.7,0))
axis(4, at = c(-yl,yl), labels = c(-yl,yl), tick = TRUE,)
axis(3, at = c(-lagmax,-1,0,1,lagmax), labels = TRUE, tick = TRUE,)
textplot('')
par(mar=c(0,2,0.2,0))
ccf(logr_data_week[,2],logr_data_week[,7],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,2],logr_data_week[,8],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,2],logr_data_week[,9],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
```

```
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,2],logr_data_week[,10],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,2],na.omit(logr_data_week[,11]),
lag.max=lagmax,xlab='',ylab='',main='',xaxt='n',yaxt='n',
ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
textplot('')
par(mar=c(0,2,0.2,0))
ccf(logr_data_week[,3],logr_data_week[,7],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,3],logr_data_week[,8],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,3],logr_data_week[,9],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,3],logr_data_week[,10],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,3],na.omit(logr_data_week[,11]),lag.max=lagmax,
xlab='',ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
axis(4, at = c(-yl,yl), labels = c(-yl,yl), tick = TRUE,)
par(mar=c(0,0.2,0.2,0))
textplot('')
par(mar=c(0,2,0.2,0))
ccf(logr_data_week[,4],logr_data_week[,7],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,4],logr_data_week[,8],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,4],logr_data_week[,9],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,4],logr_data_week[,10],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,4],na.omit(logr_data_week[,11]),
lag.max=lagmax,xlab='',ylab='',main='',xaxt='n',yaxt='n',
ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
```

```
textplot('')
par(mar=c(0,2,0.2,0))
ccf(logr_data_week[,5],logr_data_week[,7],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,5],logr_data_week[,8],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,5],logr_data_week[,9],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,5],logr_data_week[,10],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,5],na.omit(logr_data_week[,11]),
lag.max=lagmax,xlab='',ylab='',main='',xaxt='n',yaxt='n',
ylim=c(-yl,yl))
axis(4, at = c(-yl,yl), labels = c(-yl,yl), tick = TRUE,)
par(mar=c(0,0.2,0.2,0))
textplot('')
par(mar=c(0,2,0.2,0))
ccf(logr_data_week[,6],logr_data_week[,7],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,6],logr_data_week[,8],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,6],logr_data_week[,9],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,6],logr_data_week[,10],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,6],na.omit(logr_data_week[,11]),
lag.max=lagmax,xlab='',ylab='',main='',xaxt='n',yaxt='n',
ylim=c(-yl,yl))
axis(4, at = c(-yl,yl), labels = c(-yl,yl), tick = TRUE,)
par(mar=c(0,0.2,0.2,0))
textplot('')
par(mar=c(0,2,0.2,0))
textplot("  CSIYHYI",halign="center",cex=1.7)
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,7],logr_data_week[,8],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
```

```
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,7],logr_data_week[,9],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,7],logr_data_week[,10],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,7],na.omit(logr_data_week[,11]),
lag.max=lagmax,xlab='',ylab='',main='',xaxt='n',yaxt='n',
ylim=c(-yl,yl))
textplot('')
par(mar=c(0,2,0.2,0))
plot(logr_data_week[,9],logr_data_week[,7],xlab='',ylab='',
main='',xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
textplot(names(data[9]),valign="center",cex=1.7)
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,8],logr_data_week[,9],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,8],logr_data_week[,10],lag.max=lagmax,xlab='',
ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,8],na.omit(logr_data_week[,11]),
lag.max=lagmax,xlab='',ylab='',main='',xaxt='n',yaxt='n',
ylim=c(-yl,yl))
axis(4, at = c(-yl,yl), labels = c(-yl,yl), tick = TRUE,)
textplot('')
par(mar=c(0,2,0.2,0))
plot(logr_data_week[,9],logr_data_week[,7],xlab='',ylab='',
main='',xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,9],logr_data_week[,8],xlab='',ylab='',
main='',xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
textplot(names(data[10]),valign="center",cex=1.7)
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,9],logr_data_week[,10],lag.max=lagmax,
xlab='',ylab='',main='',xaxt='n',yaxt='n',ylim=c(-yl,yl))
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,9],na.omit(logr_data_week[,11]),
lag.max=lagmax,xlab='',ylab='',main='',xaxt='n',yaxt='n',
ylim=c(-yl,yl))
textplot('')
```

```
par(mar=c(0,2,0.2,0))
plot(logr_data_week[,10],logr_data_week[,7],xlab='',ylab='',
main='',xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,10],logr_data_week[,8],xlab='',ylab='',
main='',xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
plot(logr_data_week[,10],logr_data_week[,9],xlab='',ylab='',
main='',xaxt='n',yaxt='n')
par(mar=c(0,0.2,0.2,0))
textplot(names(data[11]),valign="center",cex=1.7)
par(mar=c(0,0.2,0.2,0))
ccf(logr_data_week[,10],na.omit(logr_data_week[,11]),
lag.max=lagmax,xlab='',ylab='',main='',xaxt='n',yaxt='n',
ylim=c(-yl,yl))
axis(4, at = c(-yl,yl), labels = c(-yl,yl), tick = TRUE,)
textplot('')
par(mar=c(0.2,2,0.2,0))
plot(logr_data_week[,11],logr_data_week[,7],xlab='',ylab='',
main='',xaxt='n',yaxt='n')
par(mar=c(0.2,0.2,0.2,0))
plot(logr_data_week[,11],logr_data_week[,8],xlab='',ylab='',
main='',xaxt='n',yaxt='n')
par(mar=c(0.2,0.2,0.2,0))
plot(logr_data_week[,11],logr_data_week[,9],xlab='',ylab='',
main='',xaxt='n',yaxt='n')
par(mar=c(0.2,0.2,0.2,0))
plot(logr_data_week[,11],logr_data_week[,10],xlab='',ylab='',
main='',xaxt='n',yaxt='n')
par(mar=c(0.2,0.2,0.2,0))
textplot(names(data[12]),valign="center",cex=1.5)
textplot('')
dev.off()
##############

#log R daily data info
index_number=11 #number in 1:11
length(na.omit(logr_data[,index_number]))
min(na.omit(logr_data[,index_number]))
max(na.omit(logr_data[,index_number]))
mean(na.omit(logr_data[,index_number]))
sd(na.omit(logr_data[,index_number]))
skewness(na.omit(logr_data[,index_number]))
kurtosis(na.omit(logr_data[,index_number]))
```

```
#log R weekly data info
index_number=11 #number in 1:11
length(na.omit(logr_data_week[,index_number]))
min(na.omit(logr_data_week[,index_number]))
max(na.omit(logr_data_week[,index_number]))
mean(na.omit(logr_data_week[,index_number]))
sd(na.omit(logr_data_week[,index_number]))
skewness(na.omit(logr_data_week[,index_number]))
kurtosis(na.omit(logr_data_week[,index_number]))

jarque.bera.test(na.omit(logr_data_week[,index_number]))

#function for normal approximation
f<-function(x){exp(-x^2/2)/sqrt(2*pi)}

#Density plot weekly log R data
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/density_plot_week_data.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(3.3,3.5,1.5,0.5),mgp=c(2,0.8,0),
cex.axis=1.6, cex.main=1.8, cex.lab =1.6,lwd=2)
textplot('')
legend('top',legend=c('Empirical density',
'Normal density'),title='Weekly log return',lty=1,lwd=c(3,3),
col=c('black','red'),bty='n',cex=1.5,seg.len=1,y.intersp=0.7)
for(index_number in 1:11){
plot(density(na.omit(logr_data_week[,index_number])),xlab='',
lwd=2,main='')
mtext(names(data)[index_number+1],3,line=0.2)
lines(c(min(na.omit(logr_data_week[,index_number]))+
(max(na.omit(logr_data_week[,index_number]))-
min(na.omit(logr_data_week[,index_number])))*(1:100)/100),
f((c(min(na.omit(logr_data_week[,index_number]))+
(max(na.omit(logr_data_week[,index_number]))-
min(na.omit(logr_data_week[,index_number])))
*(1:100)/100))/sd(na.omit(logr_data_week[,index_number])))/
sd(na.omit(logr_data_week[,index_number])),
type="l",lwd=2,col='red',xlab='Residuals')
}
dev.off()

#KPSS-test for stationarity in weekly log return data.
index.number=1  #(1:11)
kpss.test(logr_data_week[,index.number])
```

```
#Normalized index-plot with turning points
MY.colors=c("black","red","green3","blue","cyan","magenta",
"yellow","gray","indianred2","brown","darkgreen")
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/Normalized_data_plot_with_turning_points_.pdf",width=11,
height=7)
par(mar=c(4.3,4.3,2,1.5),mgp=c(3,1,0),cex.axis=1.6, cex.main=1.8,
cex.lab=1.6)
plot(Date,data[,12]/data[1162,12],type='l',
main='Normalized index plot',col=MY.colors[11],ylim=c(0,5.5),
lwd=2,ylab='Normalized index value',xlab='Time [Year]')
axis.Date(1,at=seq(as.Date("1999/1/1"),as.Date("2011/1/1"),
"years"),labels=F)
for(i in 2:11){
lines(Date,data[,i]/data[1,i],lwd=2,main=names(data)[i],
col=MY.colors[i-1])
}
lines(c(Date[21],Date[21]),c(-0.22,2.8),lwd=2)
lines(c(Date[435],Date[435]),c(-0.22,2.8),lwd=2)
lines(c(Date[782],Date[782]),c(-0.22,2.8),lwd=2)
lines(c(Date[2391],Date[2391]),c(-0.22,6),lwd=2)
lines(c(Date[2651],Date[2651]),c(-0.22,6),lwd=2)
lines(c(Date[3195],Date[3195]),c(-0.22,6),lwd=2)
legend('topleft',legend=c(names(data[2:11]),"DK00S.N.Index   "),
title='Index',lty=1,lwd=c(3,3),col=MY.colors,cex=1.5,seg.len=1,
y.intersp=0.7)
mtext('82 weeks\n',side=1,at=c(Date[200],1))
mtext('70 weeks\n',side=1,at=c(Date[600],1))
mtext('322 weeks\n',side=1,at=c(Date[1600],1))
mtext('52 weeks\n',side=1,at=c(Date[2522],1))
mtext('108 weeks\n',side=1,at=c(Date[2950],1))
dev.off()
```

Appendix C

R-script for PCA and GARCH modelling and scenario generation

```
##################################################################
## Filename: Modeling and Scenario generation.R                 ##
## Date: 20-01-2012                                             ##
## Author: Emil Ahlmann Østergaard s082632                      ##
## Description: R-script for chapter 7+8 in B.Sc. thesis:       ##
## "Scenario gerneration for financial market indices"          ##
##################################################################

setwd("C:/Users/Emil/Documents/Skole/Bachelorprojekt")
library('fGarch')
library('tseries')
library('TSA')
library('chron')
library('fields')
library('gplots')
library('SDMTools')

#Reading in data
```

```
data<-read.csv2('Data endelig version.csv',header=T)
data<-data[,1:12]
data$Date<-as.Date(data$Date,"%d-%m-%Y")
data$DK00S.N.Index[1:1161]<-NA
data$DK00S.N.Index<-as.numeric(data$DK00S.N.Index)
attach(data)

#Generating data for NDUEEGF
#end_month is where a month ends.
original.data.NDUEEGF=data$NDUEEGF
set.seed(200)
end_month=c(1,21,41,64,86,107,129,151,173,195,216,238,261,282,303,
326,346,369,391,412,435,456,478,500,521)
b=approx(x=end_month,y=data$NDUEEGF[end_month],n=521)
data$NDUEEGF[1:520]=b$y[1:520]+rnorm(520,mean=0,
sd=sd(diff(data$NDUEEGF[end_month-1]))/(sqrt(22)))

#Log return data
logr_data<-apply(log(data[,2:11]),2,diff)
logr_data<-rbind(c(rep(0,10)),logr_data)

#Weekly (friday) sample
data_week=data[seq(1,length(data[,1]),5),]

#Weekly (friday) log return sample
logr_data_week=apply(log(data_week[,2:11]),2,diff)
logr_data_week=rbind(c(rep(0,10)),logr_data_week)

#Function to get data back in log return
#Arguments: PCA, data in PCA space and number of PC's
restore=function(rotated_data,pca,n){
r=na.omit(rotated_data[,1:n]%*%t(pca$rotation[,1:n]))
r=t(apply(r,1,function(x)x*pca$scale))
r=t(apply(r,1,function(x)x+pca$center))
return(r)}

#Period 1 (up) 1999M1-2000M8
period1=logr_data_week[6:87,]
pca_period1=prcomp(period1,scale=T,center=T)
data_pca=predict(pca_period1)
gp1=list()
n_pc=4
gp1[[1]]=garch(data_pca[,1],order=c(0,1))
gp1[[2]]=garch(data_pca[,2],order=c(0,1))
```

```
gp1[[3]]=garch(data_pca[,3],order=c(0,1))
gp1[[4]]=garch(data_pca[,4],order=c(0,1))

#Scree plot of period 1
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/scree_plot_period1.pdf",width=8,height=5)
par(mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0),cex.axis=1.6,cex.lab=1.6,
cex.main=1.8,lwd=2)
screeplot(pca_period1,type='l',main = 'Scree plot period 1')
mtext('PC',1,line=1.5,cex=1.6)
dev.off()

#Period 2 (down) 2000M9-2001M12
period2=logr_data_week[88:157,]
pca_period2=prcomp(period2,scale=T,center=T)
data_pca=predict(pca_period2)
gp2=list()
gp2[[1]]=garch(data_pca[,1],order=c(0,1))
gp2[[2]]=garch(data_pca[,2],order=c(0,1))
gp2[[3]]=garch(data_pca[,3],order=c(0,1))
gp2[[4]]=garch(data_pca[,4],order=c(0,1))

#Period 3 (up) 2002M1-2008M2
period3=logr_data_week[158:479,]
pca_period3=prcomp(period3,scale=T,center=T)
data_pca=predict(pca_period3)
gp3=list()
gp3[[1]]=garch(data_pca[,1],order=c(0,1))
gp3[[2]]=garch(data_pca[,2],order=c(0,1))
gp3[[3]]=garch(data_pca[,3],order=c(0,1))
gp3[[4]]=garch(data_pca[,4],order=c(0,1))

#Period 4 (down) 2008M3-2009M2
period4=logr_data_week[480:531,]
pca_period4=prcomp(period4,scale=T,center=T)
data_pca=predict(pca_period4)
gp4=list()
gp4[[1]]=garch(data_pca[,1],order=c(0,3))
gp4[[2]]=garch(data_pca[,2],order=c(0,1))
gp4[[3]]=garch(data_pca[,3],order=c(0,1))
gp4[[4]]=garch(data_pca[,4],order=c(0,1))

#Period 5 (up) 2009M3-2011M3
period5=logr_data_week[532:639,]
```

```
pca_period5=prcomp(period5,scale=T,center=T)
data_pca=predict(pca_period5)
gp5=list()
gp5[[1]]=garch(data_pca[,1],order=c(0,1))
gp5[[2]]=garch(data_pca[,2],order=c(0,1))
gp5[[3]]=garch(data_pca[,3],order=c(0,1))
gp5[[4]]=garch(data_pca[,4],order=c(0,1))

#QQ PCA-scores plot
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/qq_scores_pc1.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(3.3,3.5,1.5,0.5),mgp=c(2,0.8,0),
cex.axis=1.6, cex.main=1.8, cex.lab =1.6,lwd=2)
textplot('QQ plot of\nPC scores\nin period 1')
for(i in 1:10){
qqplot(c(1:82),pca_period1$x[,i],type='l',xlab='',lwd=2,main='',
ylab='')
mtext(names(data)[i+1],3,line=0.2)
}
dev.off()

#Calculation of communalities
pc=pca_period5 #choose period
A=matrix(data = NA, nrow = 10, ncol = 4)
for (i in 1:10){
for (j in 1:4){
A[i,j]=pc$sd[j]*pc$rotation[i,j]
}}
B=matrix(data = 0, nrow = 10, ncol = 1)
for (i in 1:10){
B[i,]=round(sum(na.omit(A[i,]^2)),digits=2)
}
##################################################################

#Reconstruction of da data using n_pc PC
#OBS! Variables and Constant has similar names in the simulation.
n_sim=1000
data_sim_pca=array(numeric(n_pc*length(data_week[1,])),
c(length(data_week[,1]),n_pc))
data_sim_logr=array(numeric(10*length(data_week[1,])),
c(length(data_week[,1]),10))
data_sim=array(numeric(10*length(data_week[1,])*n_sim),
c(length(data_week[,1]),10,n_sim))
n_pc=4
```

```
set.seed(200)
for (k  in 1:n_sim) {
data_pca=predict(pca_period1)
pca=pca_period1
period_length=length(period1[,1])
data_rek_logr1=restore(data_pca,pca=pca,n=n_pc) #restore
data_rek1=exp(apply(data_rek_logr1,2,cumsum))
model=gp1
for  (i in 1:n_pc){
data_sim_pca[6:87,i]=t(garch.sim(model[[i]]$coef,
      n=period_length))
}
data_sim_logr[6:87,]=restore(data_sim_pca[6:87,],pca=pca,n=n_pc)
data_pca=predict(pca_period2)
pca=pca_period2
period_length=length(period2[,1])
data_rek_logr2=restore(data_pca,pca=pca,n=n_pc) #restore
data_rek2=exp(apply(data_rek_logr2,2,cumsum))
model=gp2
for  (i in 1:n_pc){
data_sim_pca[88:157,i]=t(garch.sim(model[[i]]$coef,
      n=period_length))
}
data_sim_logr[88:157,]=restore(data_sim_pca[88:157,],
pca=pca,n=n_pc) #restore
data_pca=predict(pca_period3)
pca=pca_period3
period_length=length(period3[,1])
data_rek_logr3=restore(data_pca,pca=pca,n=n_pc) #restore
data_rek3=exp(apply(data_rek_logr3,2,cumsum))
model=gp3
for  (i in 1:n_pc){
data_sim_pca[158:479,i]=t(garch.sim(model[[i]]$coef,
n=period_length))
}
data_sim_logr[158:479,]=restore(data_sim_pca[158:479,],
pca=pca,n=n_pc) #restore
data_pca=predict(pca_period4)
pca=pca_period4
period_length=length(period4[,1])
data_rek_logr4=restore(data_pca,pca=pca,n=n_pc) #restore
data_rek4=exp(apply(data_rek_logr4,2,cumsum))
model=gp4
for  (i in 1:n_pc){
```

```
data_sim_pca[480:531,i]=
      t(garch.sim(model[[i]]$coef,n=period_length))
}
data_sim_logr[480:531,]=restore(data_sim_pca[480:531,],
pca=pca,n=n_pc) #restore
data_pca=predict(pca_period5)
pca=pca_period5
period_length=length(period5[,1])
data_rek_logr5=restore(data_pca,pca=pca,n=n_pc) #restore
data_rek5=exp(apply(data_rek_logr5,2,cumsum))
model=gp5
for  (i in 1:n_pc){
data_sim_pca[532:639,i]=t(garch.sim(model[[i]]$coef,
      n=period_length))
}
data_sim_logr[532:639,]=restore(data_sim_pca[532:639,],
pca=pca,n=n_pc) #restore
data_sim[,,k]=exp(apply(data_sim_logr,2,cumsum))
}
mean_data_sim=apply(data_sim[,,1:n_sim],c(1,2),mean)
quantile_data_sim=apply(data_sim[,,1:n_sim],c(1,2),quantile,
probs = c(0.05,0.25,0.5,0.75,0.95))

#Plot of PCA reconstruction
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/PCA_reconstruction.pdf",width=8,height=5)
par(mfrow=c(2,1),mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0),cex.axis=1.6,
cex.lab=1.6,cex.main=1.8,lwd=2)
plot(data_week[6:639,1],data_week[6:639,4],type='l',
main=names(data)[4],xlab='Time [Year]',ylab='Index value',
ylim=c(2000,8500))
lines(data_week[6:87,1],data_rek1[,3]*data_week[5,4],col='red')
lines(data_week[87:157,1],c(1,data_rek2[,3])*data_week[87,4],
col='red')
lines(data_week[157:479,1],c(1,data_rek3[,3])*data_week[157,4],
col='red')
lines(data_week[479:531,1],c(1,data_rek4[,3])*data_week[479,4],
col='red')
lines(data_week[531:639,1],c(1,data_rek5[,3])*data_week[531,4],
col='red')
legend('topleft',legend=c('Original data','Reconstruction'),
col=c('black','red'),bty='n',cex=1.4,
lwd=c(2,2,2))
plot(data_week[6:639,1],data_week[6:639,8],type='l',
```

```
main=names(data)[8],xlab='Time [Year]',ylab='Index value')
lines(data_week[6:87,1],data_rek1[,7]*data_week[5,8],
col='red')
lines(data_week[87:157,1],c(1,data_rek2[,7])*data_week[87,8],
col='red')
lines(data_week[157:479,1],c(1,data_rek3[,7])*data_week[157,8],
col='red')
lines(data_week[479:531,1],c(1,data_rek4[,7])*data_week[479,8],
col='red')
lines(data_week[531:639,1],c(1,data_rek5[,7])*data_week[531,8],
col='red')
legend('topleft',legend=c('Original data','Reconstruction'),
col=c('black','red'),bty='n',cex=1.4,
lwd=c(2,2,2))
dev.off()

#Plot of simulated garch models/backtesting
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/GARCH_BACKTEST_SIM_nolegend.pdf",width=8,height=10)
par(mfrow=c(5,2),mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0),cex.axis=1.6,
cex.lab=1.6,cex.main=1.8,lwd=2)
for (index.number in 2:11){
plot(data_week[6:639,1],data_week[6:639,index.number],type='l',
main=names(data)[index.number],xlab='Time [Year]',ylab='Index value',
ylim=c(min(data_sim[6:639,index.number-1,1:20]*
data_week[5,index.number]),
max(quantile_data_sim[5,6:639,index.number-1]*
data_week[5,index.number])),xlim=c(Date[21],Date[3200]))
lines(data_week[6:639,1],quantile_data_sim[3,6:639,index.number-1]*
data_week[5,index.number],lwd=2,col='blue')    # mean
lines(data_week[6:639,1],quantile_data_sim[2,6:639,index.number-1]*
data_week[5,index.number],lwd=2,lty=2,col='blue') #25% quantile
lines(data_week[6:639,1],quantile_data_sim[4,6:639,index.number-1]*
data_week[5,index.number],lwd=2,lty=2,col='blue') #75% quantile
lines(data_week[6:639,1],quantile_data_sim[1,6:639,index.number-1]*
data_week[5,index.number],lwd=2,lty=3,col='blue') #5% quantile
lines(data_week[6:639,1],quantile_data_sim[5,6:639,index.number-1]*
data_week[5,index.number],lwd=2,lty=3,col='blue') #95% quantile
lines(data_week[6:639,1],data_week[6:639,index.number],lwd=2,
col='black')
lines(c(Date[21],Date[21]),c(0,1.25*max(quantile_data_sim[5,6:639,
index.number-1]*data_week[5,index.number])),lwd=1,col='gray')
lines(c(Date[435],Date[435]),c(0,max(quantile_data_sim[5,6:639,
index.number-1]*data_week[5,index.number])),lwd=1,col='gray')
```

```
lines(c(Date[782],Date[782]),c(0,max(quantile_data_sim[5,6:639,
index.number-1]*data_week[5,index.number])),lwd=1,col='gray')
lines(c(Date[2391],Date[2391]),c(0,max(quantile_data_sim[5,6:639,
index.number-1]*data_week[5,index.number])),lwd=1,col='gray')
lines(c(Date[2651],Date[2651]),c(0,max(quantile_data_sim[5,6:639,
index.number-1]*data_week[5,index.number])),lwd=1,col='gray')
lines(c(Date[3195],Date[3195]),c(0,max(quantile_data_sim[5,6:639,
index.number-1]*data_week[5,index.number])),lwd=1,col='gray')
#legend('top',legend=c('Data','50 % quantile','5 % and 95 %
 quantile','25 % and 75 % quantile'),bty='n',col=c('black','blue',
'blue','blue'),lty=c(1,1,3,2),cex=0.8,lwd=c(1,1,1,1))
}
dev.off()

##################################################################

#Simulation
n_sim=1000
set.seed(400)
simulation=array(0,c(700,10,n_sim))

for (j in 1:10) {
simulation[1,j,]=data_week[659,j+1]
}

for (k  in 1:n_sim) {

week=1
type=1 #1=down , 2=up

while (week < 261) {
if (type==1) {
too.short.period = 1
#while (too.short.period ==1){
#too.short.period = 0
regime=sample(c(2,4),size =1)
type=2
if (regime==2) {
regime_length=round(rgamma(1,
                              length(period2[,1])))
model=gp2
pca=pca_period2
} else {
regime_length=round(rgamma(1,
```

```
                              length(period4[,1])))
model=gp4
pca=pca_period4
}
#if (week==1) { #Adjust length of the
     first period
regime_length_test=regime_length-
                              (length(logr_data_week[,1])-640)
#if (regime_length_test<0){
#too.short.period = 1
#}
#}
             #}
} else {
regime=sample(c(1,3,5),size =1)
type=1
if (regime==1) {
regime_length=round(rgamma(1,
                        length(period1[,1])))
model=gp1
pca=pca_period1
} else  if (regime==3){
regime_length=round(rgamma(1,
                        length(period3[,1])))
model=gp3
pca=pca_period3
} else {
regime_length=round(rgamma(1,
                        length(period5[,1])))
model=gp5
pca=pca_period5
}
}
data_sim_pca=array(numeric(n_pc*regime_length),
      c(regime_length,n_pc))
for (i in 1:n_pc){
data_sim_pca[,i]=t(garch.sim(model[[i]]$coef,
      n=regime_length))
}
data_sim_logr=restore(data_sim_pca,pca=pca,n=n_pc) #restore
data_sim_logr=array(data_sim_logr,c(regime_length,10))
data_sim=exp(apply(data_sim_logr,2,cumsum)) #data org. space
for (i in 1:10){
simulation[(week+1):(week+regime_length),i,k]=
```

```
      matrix(data_sim[,i],ncol = 1)*simulation[week,i,k]
}
week=week+regime_length
}
}

mean_data_sim=apply(simulation[2:261,,1:n_sim],c(1,2),mean)
quantile_data_sim=apply(simulation[2:261,,1:n_sim],c(1,2),
quantile,probs = c(0.05,0.25,0.5,0.75,0.95))
quantile_data_sim_test=apply(simulation[2:261,,1:350],c(1,2),
quantile,probs = c(0.05,0.25,0.5,0.75,0.95))
##
#Draw down
DD=array(0,c(n_sim,261,10))
MDD=DD
MDD_simulation_number=array(0,c(10,1))
for (k in 1:10){ # 1: number of indices
for (j in 1:n_sim){
peak = 0
for (i in 2:261){
if (simulation[i,k,j] > peak) {
     peak = simulation[i,k,j]
} else {
DD[j,i,k] = 100.0 * (peak - simulation[i,k,j])
                  / peak #Relative DD in %
# DD[j,i,k] = peak - simulation[i,k,j] #Abs. DD
}
if (DD[j,i,k] > MDD[j,i,k]){
       MDD[j,i:261,k] = DD[j,i,k] #set Maximum DD
}
}
}
MDD_simulation_number[k,]=which(max(MDD[,261,k])==MDD[,261,k])
                                           #Scenario with MDD
}

#################
#Plot of Scenario index 1:3
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/GARCH_scenarios_1_3.pdf",width=8,height=12)
par(mfrow=c(3,1),mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0),cex.axis=1.6,
cex.lab=1.6,cex.main=1.8,lwd=2)
for (index_number in 1:3){
screen( index_number )
```

```
par(mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0))
plot(simulation[2:261,index_number,1],type='n',ylim=c(0,1.1*max(
quantile_data_sim[5,,index_number])),xlab='Time [Year]',
ylab='Index value',main=names(data[index_number+1]),xaxt='n')
#plot(simulation[2:261,index_number,1],type='n',ylim=c(min(
simulation[2:261,index_number,
MDD_simulation_number[index_number,]]),max(
simulation[2:261,index_number,MDD_simulation_number[index_number,]
])),xlab='Time [Year]',ylab='Index value',main=names(
data[index_number+1]),xaxt='n')
axis(1,(21+c(0:4)*52),labels=rep('',5))
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
for (j in 51:70){
lines(simulation[2:261,index_number,j],col='gray',lwd=1)
}
lines(simulation[2:261,index_number,MDD_simulation_number[
index_number,]],col='red',lwd=1)
#lines(mean_data_sim[,index_number],lwd=2,col='blue') # mean
lines(quantile_data_sim[3,,index_number],lwd=2)  #50% quantile
lines(quantile_data_sim[2,,index_number],lwd=2,lty=2) #25% quantile
lines(quantile_data_sim[4,,index_number],lwd=2,lty=2) #75% quantile
lines(quantile_data_sim[1,,index_number],lwd=2,lty=3) #5% quantile
lines(quantile_data_sim[5,,index_number],lwd=2,lty=3) #95% quantile
legend('topleft',legend=c('MDD','20 scenarios','50% quantile',
'5% and 95% quantile','25% and 75% quantile'),col=c('red','gray',
'black','black','black'),bty='n',cex=1.4,lty=c(1,1,1,3,2),lwd=
c(2,2,2))
}
dev.off()

#Plot of Scenario index 4:6
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/GARCH_scenarios_4_6.pdf",width=8,height=12)
par(mfrow=c(3,1),mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0),cex.axis=1.6,
cex.lab=1.6,cex.main=1.8,lwd=2)
for (index_number in 4:6){
screen( index_number )
par(mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0))
plot(simulation[2:261,index_number,1],type='n',ylim=c(0,1.2*max(
quantile_data_sim[5,,index_number])),xlab='Time [Year]',
ylab='Index value',main=names(data[index_number+1]),xaxt='n')
#plot(simulation[2:261,index_number,1],type='n',ylim=c(min(
simulation[2:261,index_number,MDD_simulation_number[index_number,]
]),max(simulation[2:261,index_number,
```

```
MDD_simulation_number[index_number,]])),xlab='Time [Year]',
ylab='Index value',main=names(data[index_number+1]),xaxt='n')
axis(1,(21+c(0:4)*52),labels=rep('',5))
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
for (j in 51:70){
lines(simulation[2:261,index_number,j],col='gray',lwd=1)
}
lines(simulation[2:261,index_number,MDD_simulation_number[
index_number,]],col='red',lwd=1)
#lines(mean_data_sim[,index_number],lwd=2,col='blue') # mean
lines(quantile_data_sim[3,,index_number],lwd=2)  #50% quantile
lines(quantile_data_sim[2,,index_number],lwd=2,lty=2) #25% quantile
lines(quantile_data_sim[4,,index_number],lwd=2,lty=2) #75% quantile
lines(quantile_data_sim[1,,index_number],lwd=2,lty=3) # 5% quantile
lines(quantile_data_sim[5,,index_number],lwd=2,lty=3) #95% quantile
legend('topleft',legend=c('MDD','20 scenarios','50% quantile',
'5% and 95% quantile','25% and 75% quantile'),col=c('red','gray',
'black','black','black'),bty='n',cex=1.4,lty=c(1,1,1,3,2),
lwd=c(2,2,2))

}
dev.off()

#Plot of Scenario index 7:9
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/GARCH_scenarios_7_9.pdf",width=8,height=12)
par(mfrow=c(3,1),mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0),cex.axis=1.6,
cex.lab=1.6,cex.main=1.8,lwd=2)
for (index_number in 7:8){
screen( index_number )
par(mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0))
plot(simulation[2:261,index_number,1],type='n',ylim=c(0,1.1*
max(quantile_data_sim[5,,index_number])),xlab='Time [Year]',
ylab='Index value',main=names(data[index_number+1]),xaxt='n')
#plot(simulation[2:261,index_number,1],type='n',ylim=c(min(
simulation[2:261,index_number,MDD_simulation_number[index_number,]
]),max(simulation[2:261,index_number,MDD_simulation_number[
index_number,]])),xlab='Time [Year]',ylab='Index value',
main=names(data[index_number+1]),xaxt='n')
axis(1,(21+c(0:4)*52),labels=rep('',5))
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
for (j in 51:70){
lines(simulation[2:261,index_number,j],col='gray',lwd=1)
}
```

```
lines(simulation[2:261,index_number,
MDD_simulation_number[index_number,]],col='red',lwd=1)
#lines(mean_data_sim[,index_number],lwd=2,col='blue') # mean
lines(quantile_data_sim[3,,index_number],lwd=2)  #50% quantile
lines(quantile_data_sim[2,,index_number],lwd=2,lty=2) #25% quantile
lines(quantile_data_sim[4,,index_number],lwd=2,lty=2) #75% quantile
lines(quantile_data_sim[1,,index_number],lwd=2,lty=3) # 5% quantile
lines(quantile_data_sim[5,,index_number],lwd=2,lty=3) #95% quantile
legend('topleft',legend=c('MDD','20 scenarios','50% quantile',
'5% and 95% quantile','25% and 75% quantile'),col=c('red','gray',
'black','black','black'),bty='n',cex=1.4,lty=c(1,1,1,3,2),
lwd=c(2,2,2))
}
index_number=9
plot(simulation[2:261,index_number,1],type='n',ylim=c(0.95*
min(simulation[2:261,index_number,]),1.05*max(quantile_data_sim[5,
,index_number])),xlab='Time [Year]',ylab='Index value',
main=names(data[index_number+1]),xaxt='n')
axis(1,(21+c(0:4)*52),labels=rep('',5))
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
for (j in 51:70){
lines(simulation[2:261,index_number,j],col='gray',lwd=1)
}
lines(simulation[2:261,index_number,MDD_simulation_number[
index_number,]],col='red',lwd=1)
#lines(mean_data_sim[,index_number],lwd=2,col='blue') #    mean
lines(quantile_data_sim[3,,index_number],lwd=2)  #50% quantile
lines(quantile_data_sim[2,,index_number],lwd=2,lty=2) #25% quantile
lines(quantile_data_sim[4,,index_number],lwd=2,lty=2) #75% quantile
lines(quantile_data_sim[1,,index_number],lwd=2,lty=3) # 5% quantile
lines(quantile_data_sim[5,,index_number],lwd=2,lty=3) #95% quantile
legend('topleft',legend=c('MDD','20 scenarios','50% quantile',
'5% and 95% quantile','25% and 75% quantile'),col=c('red','gray',
'black','black','black'),bty='n',cex=1.4,lty=c(1,1,1,3,2),
lwd=c(2,2,2))
dev.off()

#Plot of Scenario index 10
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/GARCH_scenarios_10.pdf",width=8,height=12)
par(mfrow=c(3,1),mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0),cex.axis=1.6,
cex.lab=1.6,cex.main=1.8,lwd=2)
index_number=10
par(mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0))
```

```
plot(simulation[2:261,index_number,1],type='n',ylim=c(0.95*min(
simulation[2:261,index_number,]),1.05*max(quantile_data_sim[
5,,index_number])),xlab='Time [Year]',ylab='Index value',
main=names(data[index_number+1]),xaxt='n')
axis(1,(21+c(0:4)*52),labels=rep('',5))
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
for (j in 51:70){
lines(simulation[2:261,index_number,j],col='gray',lwd=1)
}
lines(simulation[2:261,index_number,MDD_simulation_number[
index_number,]],col='red',lwd=1)
#lines(mean_data_sim[,index_number],lwd=2,col='blue') #    mean
lines(quantile_data_sim[3,,index_number],lwd=2)  #50% quantile
lines(quantile_data_sim[2,,index_number],lwd=2,lty=2) #25% quantile
lines(quantile_data_sim[4,,index_number],lwd=2,lty=2) #75% quantile
lines(quantile_data_sim[1,,index_number],lwd=2,lty=3) # 5% quantile
lines(quantile_data_sim[5,,index_number],lwd=2,lty=3) #95% quantile
legend('topleft',legend=c('MDD','20 scenarios','50% quantile',
'5% and 95% quantile','25% and 75% quantile'),col=c('red','gray',
'black','black','black'),bty='n',cex=1.4,lty=c(1,1,1,3,2),
lwd=c(2,2,2))
dev.off()
############################

#Plot of normalized mean scenario
MY.colors=c("black","red","green3","blue","cyan","magenta",
"yellow","gray","indianred2","brown","darkgreen")
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/Normalized_50_quantile_scenarios.pdf",width=11,height=7)
par(mar=c(4.3,4.3,2,1.5),mgp=c(3,1,0),cex.axis=1.6, cex.main=1.8)
plot(quantile_data_sim[3,,1]/quantile_data_sim[3,1,1],type='l',
ylab='Normalized index value',main='Normalized 50% quantile
scenarios',col=MY.colors[1],ylim=c(0.6,1.7),lwd=2,xaxt='n',
xlab='Time [Year]',cex.lab=1.6)
axis(1,(21+c(0:4)*52),labels=rep('',5))
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
for(index_number in 1:10){
lines(quantile_data_sim[3,,index_number]/quantile_data_sim[3,1,
index_number],lwd=2,col=MY.colors[index_number])
}
legend('topleft',legend=c(names(data[2:11])),title='Index',
lty=1,lwd=c(3,3),col=MY.colors,cex=1.5,seg.len=1,y.intersp=0.7)
dev.off()
```

```
#Plot of normalized Maximum drawdown scenario
MY.colors=c("black","red","green3","blue","cyan","magenta",
"yellow","gray","indianred2","brown","darkgreen")
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/Normalized_MDD_scenarios.pdf",width=11,height=7)
par(mar=c(4.3,4.3,2,1.5),mgp=c(3,1,0),cex.axis=1.6, cex.main=1.8,
cex.lab=1.6)
plot(simulation[2:261,1,MDD_simulation_number[1,]]/simulation[2,1,
MDD_simulation_number[1,]],ylab='Normalized index value',
xlab='Time [Year]',type='l',main='Normalized MDD scenarios',
col=MY.colors[1],ylim=c(0,2.2),lwd=2,xaxt='n')
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
for(index_number in 1:10){
lines(simulation[2:261,index_number,MDD_simulation_number[
index_number,]]/simulation[2,index_number,MDD_simulation_number[
index_number,]],lwd=2,col=MY.colors[index_number])
}
legend('topright',legend=c(names(data[2:11])),title='Index',
lty=1,lwd=c(3,3),col=MY.colors,cex=1.5,seg.len=1,y.intersp=0.7)
dev.off()

#End values
end_values=array(0,c(n_sim,10))
for (index_number in 1:10){
end_values[,index_number]=simulation[261,index_number,]
}

#Histogram of endvalues
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/HIST_end_values.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(3.3,3.5,2,0.6),mgp=c(2,0.8,0),cex.axis=1.6,
 cex.main=1.8,cex.axis=1.2, cex.lab =1.6,lwd=2)
textplot('Histogram of\nsimulation\nendvalues',valign='top')
for (index_number in 1:10){
options(scipen=5)
Mean=mean(end_values[,index_number])
B=quantile(end_values[,index_number], probs = c(0.05,0.95,0.50))
C=max(end_values[,index_number])-min(end_values[,index_number])
hist(end_values[,index_number],breaks=C*30/(B[2]-B[1]),prob=TRUE,
xlim=c(B[1],B[2]),main='',xlab='Endvalue',yaxt='n')
options(scipen=-2)
yticks = round(range(hist(end_values[,index_number],breaks=C*30/
(B[2]-B[1]),plot=F)$density),digits=5)
axis(2, at=yticks, labels=c(yticks))
```

```
A=density(end_values[,index_number],n=10000)
lines(A,lwd=2,col='blue')
lines(c(B[3],B[3]),c(1/7*max(A$y),0),col='red',lwd=5)
lines(c(B[2],B[2]),c(1/7*max(A$y),0),col='green',lwd=5)
lines(c(B[1],B[1]),c(1/7*max(A$y),0),col='green',lwd=5)
mtext(names(data)[index_number+1],3,line=0.2)
}
par(mar=c(0,0,0,0))
textplot('')
legend('top',legend=c('Density','5% & 95% quantile',
'50% quantile'),col=c('blue','green','red'),bty='n',pt.cex=1.5,
lwd=c(2,2,2),cex=1.6)
dev.off()

options(scipen=5)

#Relative changes in 5 year simulation
relative_changes=array(0,c(n_sim,10))
for (index_number in 1:10){
relative_changes[,index_number]=(simulation[261,index_number,]-
data_week[659,index_number+1])/data_week[659,index_number+1]
}

#SD on endvalues
sd(relative_changes)

#Histogram relative changes in scenarios
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/HIST_relative_end_values.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(3.3,3.5,1.5,0.5),mgp=c(2,0.8,0),
cex.axis=1.6, cex.main=1.8, cex.lab =1.6,lwd=2)
textplot('Histogram of rela-\ntive changes in\n5 year simulations'
,valign='top')
for (index_number in 1:10){
Mean=mean(relative_changes[,index_number])
B=quantile(relative_changes[,index_number], probs = c(0.05,0.95,
0.50))
C=max(relative_changes[,index_number])-min(relative_changes[,
index_number])
hist(relative_changes[,index_number],breaks=C*30/(B[2]-B[1]),
xlim=c(B[1],B[2]),prob=TRUE,main='',xlab='')
A=density(relative_changes[,index_number],n=10000)
lines(A,lwd=2,col='blue')
lines(c(B[3],B[3]),c(1/7*max(A$y),0),col='red',lwd=5)
```

```
lines(c(B[1],B[1]),c(1/7*max(A$y),0),col='green',lwd=5)
lines(c(B[2],B[2]),c(1/7*max(A$y),0),col='green',lwd=5)
mtext(names(data)[index_number+1],3,line=0.2)
}
par(mar=c(0,0,0,0))
textplot('')
legend('top',legend=c('Density','5% & 95% quantile','50% quantile'
),col=c('blue','green','red'),bty='n',pt.cex=1.5,
lwd=c(2,2,2),cex=1.6)
dev.off()

#Drawdown at time t
MaxDD=array(0,c(261,10))
quantile_DD_sim=array(0,c(5,261,10))
for (index_number in 1:10){
for (i in 1:261){
MaxDD[i,index_number]=max(DD[,1:i,index_number])
quantile_DD_sim[,i,index_number]=quantile(MDD[,i,index_number],
probs = c(0.05,0.25,0.5,0.75,0.95))
}
}

# MaxDD and drawdown(1:t)-plot
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/MaxDD_plot.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(3.3,3.5,1.5,0.5),mgp=c(2,0.8,0),
cex.axis=1.6, cex.main=1.8, cex.lab =1.6,lwd=2)
textplot('Drawdown from\nsimulation-start \nuntil time t')
for(index_number in 1:10){
plot(MaxDD[,index_number],type='l',lwd=2,main='',ylab='DD(t) [%]',
xlab='Time, t [Year]',col='red',xaxt='n')
axis(1,(21+c(0:4)*52),labels=rep('',5))
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
lines(quantile_DD_sim[3,,index_number],lwd=2,lty=1) #50% quantile
lines(quantile_DD_sim[2,,index_number],lwd=2,lty=2) #25% quantile
lines(quantile_DD_sim[4,,index_number],lwd=2,lty=2) #75% quantile
lines(quantile_DD_sim[1,,index_number],lwd=2,lty=3) #5% quantile
lines(quantile_DD_sim[5,,index_number],lwd=2,lty=3) #95% quantile
mtext(names(data)[index_number+1],3,line=0.2)
}
par(mar=c(0,0,0,0))
textplot('')
legend('top',legend=c('MDD(1:t)','50% quantile', '5% & 95%
quantile','25 % & 75 % quantile'),col=c('red','black','black',
```

```
'black'),bty='n',pt.cex=1.5,lwd=c(2,2,2,2),lty=c(1,1,3,2),cex=1.6)
dev.off()


###################################################################
#            #
# TESTING SCENARIOS                              #
#      #
###################################################################
relative_return=array(0,c(3,10))

length_of_data=(data[3291,1]-data[1,1])/365.242199  # in years

#Relative return data
for (index_number in 1:10){
relative_return[1,index_number]=(1+(data[3291,index_number+1]-
data[1,index_number+1])/data[1,index_number+1])^(1/as.numeric(
length_of_data))-1
relative_return[2,index_number]=(1+(quantile_data_sim[3,260,
index_number]-quantile_data_sim[3,1,index_number])/
quantile_data_sim[3,1,index_number])^(1/5)-1
relative_return[3,index_number]=(1+(quantile_data_sim[3,260,
index_number]-data_week[659,index_number+1])/data_week[659,
index_number+1])^(1/5)-1
}
plot(data[,6])
```

Appendix D

R-script for scenario generation using bootstrapping

```
##################################################################
## Filename: Scenario_generation_bootstrapping.R               ##
## Date: 20-01-2012                                            ##
## Author: Emil Ahlmann Østergaard s082632                     ##
## Description: R-script for chapter 8 in B.Sc. thesis:        ##
## "Scenario gerneration for financial market indices"         ##
##################################################################

setwd("C:/Users/Emil/Documents/Skole/Bachelorprojekt")
library('fGarch')
library('tseries')
library('TSA')
library('chron')
library('fields')
library('gplots')
library('SDMTools')

#Reading in data
```

```
data<-read.csv2('Data endelig version.csv',header=T)
data<-data[,1:12]
data$Date<-as.Date(data$Date,"%d-%m-%Y")
data$DK00S.N.Index[1:1161]<-NA
data$DK00S.N.Index<-as.numeric(data$DK00S.N.Index)
attach(data)

#Generating data for NDUEEGF
#end_month is where a month ends.
original.data.NDUEEGF=data$NDUEEGF
set.seed(200)
end_month=c(1,21,41,64,86,107,129,151,173,195,216,238,261,282,303,
326,346,369,391,412,435,456,478,500,521)
b=approx(x=end_month,y=data$NDUEEGF[end_month],n=521)
data$NDUEEGF[1:520]=b$y[1:520]+rnorm(520,mean=0,
sd=sd(diff(data$NDUEEGF[end_month-1]))/(sqrt(22)))

#Log return data
logr_data<-apply(log(data[,2:11]),2,diff)
logr_data<-rbind(c(rep(0,10)),logr_data)

#Weekly (friday) sample
data_week=data[seq(1,length(data[,1]),5),]

#Weekly (friday) log return sample
logr_data_week=apply(log(data_week[,2:11]),2,diff)
logr_data_week=rbind(c(rep(0,10)),logr_data_week)
logr_data_week_use=array(logr_data_week[6:659,],c(length(6:659),
10))

#################
# Bootstrapping #
#################
SD_week=5 #SD on weekjump
min_max=(sqrt(SD_week^2*12+1)-1)/2 #min/max in centered uniform
                                   #dist. with SD=SD_week

n_sim=1000
set.seed(400)
simulation=array(0,c(261,10,n_sim))
data_sim=array(0,c(261,10,n_sim))
data_sim_logr=array(0,c(261,10,n_sim))
tarray=array(0,c(261,10,n_sim))
data_sim_logr[1,,1:n_sim]=logr_data_week_use[654,]
```

```
for (j in 1:n_sim){
for (i in 2:261){
t=1000
while (t>654 || t<1){
t=round(runif(1,min=-min_max+which(data_sim_logr[i-1,
1,j]==logr_data_week_use[,1]),max= min_max+which(
data_sim_logr[i-1,1,j]==logr_data_week_use[,1])))
}
tarray[i,,j]=t
data_sim_logr[i,,j]=logr_data_week_use[t,]
}
}

data_sim[,,]=exp(apply(data_sim_logr[,,],c(2,3),cumsum))#sim. data
for (index_number in 1:10){
simulation[,index_number,]=data_week[659,index_number+1]*
data_sim[,index_number,]
}

mean_data_sim=apply(simulation[2:261,,1:n_sim],c(1,2),mean)
quantile_data_sim=apply(simulation[2:261,,1:n_sim],c(1,2),quantile,
probs = c(0.05,0.25,0.5,0.75,0.95))

#Draw down
DD=array(0,c(n_sim,261,10))
MDD=DD
MDD_simulation_number=array(0,c(10,1))
for (k in 1:10){ # 1: number of indices
for (j in 1:n_sim){
peak = 0
for (i in 2:261){
if (simulation[i,k,j] > peak) {
     peak = simulation[i,k,j]
} else {
DD[j,i,k] = 100.0 * (peak - simulation[i,k,j]) /
 peak #Relative DD in %
# DD[j,i,k] = peak - simulation[i,k,j]#Absolute DD
}
if (DD[j,i,k] > MDD[j,i,k]){
       MDD[j,i:261,k] = DD[j,i,k] #set Maximum DD
}
}
}
MDD_simulation_number[k,]=which(max(MDD[,261,k])==MDD[,261,k])
```

```
#Scenario with MDD
}

#Plot af Scenarier index 1:3
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/GARCH_scenarios_1_3_BOOTSTRAP.pdf",width=8,height=12)
par(mfrow=c(3,1),mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0),cex.axis=1.6,
cex.lab=1.6,cex.main=1.8,lwd=2)
for (index_number in 1:3){
screen( index_number )
par(mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0))
plot(simulation[2:261,index_number,1],type='n',ylim=c(0,1.1*
max(quantile_data_sim[5,,index_number])),xlab='Time [Year]',
ylab='Index value',main=names(data[index_number+1]),xaxt='n')
#plot(simulation[2:261,index_number,1],type='n',
ylim=c(min(simulation[2:261,index_number,
MDD_simulation_number[index_number,]]),
max(simulation[2:261,index_number,
MDD_simulation_number[index_number,]])),
xlab='Time [Year]',ylab='Index value',
main=names(data[index_number+1]),xaxt='n')
axis(1,(21+c(0:4)*52),labels=rep('',5))
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
for (j in 1:20){
lines(simulation[2:261,index_number,j],col='gray',lwd=1)
}
lines(simulation[2:261,index_number,
MDD_simulation_number[index_number,]],col='red',lwd=1)
#lines(mean_data_sim[,index_number],lwd=2) #   mean
lines(quantile_data_sim[3,,index_number],lwd=2)       #50%quantile
lines(quantile_data_sim[2,,index_number],lwd=2,lty=2) #25%quantile
lines(quantile_data_sim[4,,index_number],lwd=2,lty=2) #75%quantile
lines(quantile_data_sim[1,,index_number],lwd=2,lty=3) #5%quantile
lines(quantile_data_sim[5,,index_number],lwd=2,lty=3) #95%quantile
legend('topleft',legend=c('MDD','20 scenarios','50% quantile',
'5% and 95% quantile','25 % and 75 % quantile'),col=c('red',
'gray','black','black','black'),bty='n',cex=1.4,lty=c(1,1,1,3,2),
lwd=c(2,2,2))
}
dev.off()

#Plot af Scenarier index 4:6
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/GARCH_scenarios_4_6_Bootstrap.pdf",width=8,height=12)
```

```
par(mfrow=c(3,1),mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0),cex.axis=1.6,
cex.lab=1.6,cex.main=1.8,lwd=2)
for (index_number in 4:6){
screen( index_number )
par(mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0))
plot(simulation[2:261,index_number,1],type='n',ylim=c(0,
1.2*max(quantile_data_sim[5,,index_number])),xlab='Time [Year]',
ylab='Index value',main=names(data[index_number+1]),xaxt='n')
#plot(simulation[2:261,index_number,1],type='n',
ylim=c(min(simulation[2:261,index_number,
MDD_simulation_number[index_number,]]),
max(simulation[2:261,index_number,
MDD_simulation_number[index_number,]])),
xlab='Time [Year]',ylab='Index value',
main=names(data[index_number+1]),xaxt='n')
axis(1,(21+c(0:4)*52),labels=rep('',5))
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
for (j in 51:70){
lines(simulation[2:261,index_number,j],col='gray',lwd=1)
}
lines(simulation[2:261,index_number,
MDD_simulation_number[index_number,]],col='red',lwd=1)
#lines(mean_data_sim[,index_number],lwd=2) #   mean
lines(quantile_data_sim[3,,index_number],lwd=2)       #50%quantile
lines(quantile_data_sim[2,,index_number],lwd=2,lty=2) #25%quantile
lines(quantile_data_sim[4,,index_number],lwd=2,lty=2) #75%quantile
lines(quantile_data_sim[1,,index_number],lwd=2,lty=3) #5%quantile
lines(quantile_data_sim[5,,index_number],lwd=2,lty=3) #95%quantile
legend('topleft',legend=c('MDD','20 scenarios','50% quantile',
'5% and 95% quantile','25 % and 75 % quantile'),col=c('red',
'gray','black','black','black'),bty='n',cex=1.4,lty=c(1,1,1,3,2),
lwd=c(2,2,2))
}
dev.off()

#Plot af Scenarier index 7:9
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/GARCH_scenarios_7_9_Bootstrap.pdf",width=8,height=12)
par(mfrow=c(3,1),mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0),
cex.axis=1.6,cex.lab=1.6,cex.main=1.8,lwd=2)
for (index_number in 7:8){
screen( index_number )
par(mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0))
plot(simulation[2:261,index_number,1],type='n',ylim=c(0,
```

```
1.1*max(quantile_data_sim[5,,index_number])),xlab='Time [Year]',
ylab='Index value',main=names(data[index_number+1]),xaxt='n')
#plot(simulation[2:261,index_number,1],type='n',
ylim=c(min(simulation[2:261,index_number,
MDD_simulation_number[index_number,]]),
max(simulation[2:261,index_number,
MDD_simulation_number[index_number,]])),
xlab='Time [Year]',ylab='Index value',
main=names(data[index_number+1]),xaxt='n')
axis(1,(21+c(0:4)*52),labels=rep('',5))
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
for (j in 51:70){
lines(simulation[2:261,index_number,j],col='gray',lwd=1)
}
lines(simulation[2:261,index_number,
MDD_simulation_number[index_number,]],col='red',lwd=1)
#lines(mean_data_sim[,index_number],lwd=2) #    mean
lines(quantile_data_sim[3,,index_number],lwd=2)       #50% quantile
lines(quantile_data_sim[2,,index_number],lwd=2,lty=2) #25%quantile
lines(quantile_data_sim[4,,index_number],lwd=2,lty=2) #75%quantile
lines(quantile_data_sim[1,,index_number],lwd=2,lty=3) #5%quantile
lines(quantile_data_sim[5,,index_number],lwd=2,lty=3) #95%quantile
legend('topleft',legend=c('MDD','20 scenarios','50% quantile',
'5% and 95% quantile','25 % and 75 % quantile'),col=c('red',
'gray','black','black','black'),bty='n',cex=1.4,lty=c(1,1,1,3,2),
lwd=c(2,2,2))
}
index_number=9
plot(simulation[2:261,index_number,1],type='n',
ylim=c(0.95*min(simulation[2:261,index_number,]),
1.05*max(quantile_data_sim[5,,index_number])),xlab='Time [Year]',
ylab='Index value',main=names(data[index_number+1]),xaxt='n')
axis(1,(21+c(0:4)*52),labels=rep('',5))
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
for (j in 51:70){
lines(simulation[2:261,index_number,j],col='gray',lwd=1)
}
lines(simulation[2:261,index_number,
MDD_simulation_number[index_number,]],col='red',lwd=1)
#lines(mean_data_sim[,index_number],lwd=2)       #   mean
lines(quantile_data_sim[3,,index_number],lwd=2)       #50%quantile
lines(quantile_data_sim[2,,index_number],lwd=2,lty=2) #25%quantile
lines(quantile_data_sim[4,,index_number],lwd=2,lty=2) #75%quantile
lines(quantile_data_sim[1,,index_number],lwd=2,lty=3) #5%quantile
```

```
lines(quantile_data_sim[5,,index_number],lwd=2,lty=3) #95%quantile
legend('topleft',legend=c('MDD','20 scenarios','50% quantile',
'5% and 95% quantile','25 % and 75 % quantile'),col=c('red',
'gray','black','black','black'),bty='n',cex=1.4,lty=c(1,1,1,3,2),
lwd=c(2,2,2))
dev.off()

#Plot af Scenarier index 10
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/GARCH_scenarios_10_Bootstrap.pdf",width=8,height=12)
par(mfrow=c(3,1),mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0),cex.axis=1.6,
cex.lab=1.6,cex.main=1.8,lwd=2)
index_number=10
par(mar=c(3.1,3.1,2,0.5),mgp=c(2,0.8,0))
plot(simulation[2:261,index_number,1],type='n',
ylim=c(0.95*min(simulation[2:261,index_number,]),
1.05*max(quantile_data_sim[5,,index_number])),xlab='Time [Year]',
ylab='Index value',main=names(data[index_number+1]),xaxt='n')
axis(1,(21+c(0:4)*52),labels=rep('',5))
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
for (j in 51:70){
lines(simulation[2:261,index_number,j],col='gray',lwd=1)
}
lines(simulation[2:261,index_number,
MDD_simulation_number[index_number,]],col='red',lwd=1)
#lines(mean_data_sim[,index_number],lwd=2)       #     mean
lines(quantile_data_sim[3,,index_number],lwd=2)       #50%quantile
lines(quantile_data_sim[2,,index_number],lwd=2,lty=2) #25%quantile
lines(quantile_data_sim[4,,index_number],lwd=2,lty=2) #75%quantile
lines(quantile_data_sim[1,,index_number],lwd=2,lty=3) #5%quantile
lines(quantile_data_sim[5,,index_number],lwd=2,lty=3) #95%quantile
legend('topleft',legend=c('MDD','20 scenarios','50% quantile',
'5% and 95% quantile','25 % and 75 % quantile'),col=c('red',
'gray','black','black','black'),bty='n',cex=1.4,lty=c(1,1,1,3,2),
lwd=c(2,2,2))
dev.off()

#Plot of normalized 50% quantile scenario
MY.colors=c("black","red","green3","blue","cyan","magenta",
"yellow","gray","indianred2","brown","darkgreen")
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/Normalized_50_quantile_scenarios_Bootstrap.pdf",width=11,
height=7)
par(mar=c(4.3,4.3,2,1.5),mgp=c(3,1,0),cex.axis=1.6, cex.main=1.8)
```

```
plot(quantile_data_sim[3,,1]/quantile_data_sim[3,1,1],type='l',
ylab='Normalized index value',
main='Normalized 50% quantile scenarios',col=MY.colors[1],
ylim=c(0.7,1.7),lwd=2,xaxt='n',xlab='Time [Year]',cex.lab=1.6)
axis(1,(21+c(0:4)*52),labels=rep('',5))
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
for(index_number in 1:10){
lines(quantile_data_sim[3,,index_number]/quantile_data_sim[3,1,
index_number],lwd=2,col=MY.colors[index_number])
}
legend('topleft',legend=c(names(data[2:11])),title='Index',lty=1,
lwd=c(3,3),col=MY.colors,cex=1.5,seg.len=1,y.intersp=0.7)
dev.off()

#Plot of normalized Maximum drawdown scenario
MY.colors=c("black","red","green3","blue","cyan","magenta",
"yellow","gray","indianred2","brown","darkgreen")
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/Normalized_MDD_scenarios_Bootstrap.pdf",width=11,height=7)
par(mar=c(4.3,4.3,2,1.5),mgp=c(3,1,0),cex.axis=1.6, cex.main=1.8,
cex.lab=1.6)
plot(simulation[2:261,1,MDD_simulation_number[1,]]/simulation[2,1,
MDD_simulation_number[1,]],ylab='Normalized index value',
xlab='Time [Year]',type='l',main='Normalized MDD scenarios',
col=MY.colors[1],ylim=c(0,3.1),lwd=2,xaxt='n')
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
for(index_number in 1:10){
lines(simulation[2:261,index_number,
MDD_simulation_number[index_number,]]/simulation[2,index_number,
MDD_simulation_number[index_number,]],
lwd=2,col=MY.colors[index_number])
}
legend('topleft',legend=c(names(data[2:11])),title='Index',
lty=1,lwd=c(3,3),col=MY.colors,cex=1.5,seg.len=1,y.intersp=0.7)
dev.off()

#End values
end_values=array(0,c(n_sim,10))
for (index_number in 1:10){
end_values[,index_number]=simulation[261,index_number,]
}

#Histogram of endvalues
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
```

```
figures/HIST_end_values_Bootstrap.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(3.3,3.5,2,0.6),mgp=c(2,0.8,0),cex.axis=1.6,
 cex.main=1.8,cex.axis=1.2, cex.lab =1.6,lwd=2)
textplot('Histogram of\nsimulation\nendvalues',valign='top')
for (index_number in 1:10){
options(scipen=5)
#Mean=mean(end_values[,index_number])
B=quantile(end_values[,index_number], probs = c(0.05,0.95,0.50))
C=max(end_values[,index_number])-min(end_values[,index_number])
hist(end_values[,index_number],breaks=C*30/(B[2]-B[1]),prob=TRUE,
xlim=c(B[1],B[2]),main='',xlab='Endvalue',yaxt='n',
ylim=c(0,1.1*range(hist(end_values[,index_number],
breaks=C*30/(B[2]-B[1]),plot=F)$density)[2]))
options(scipen=-2)
yticks = round(range(hist(end_values[,index_number],
breaks=C*30/(B[2]-B[1]),plot=F)$density),digits=5)
axis(2, at=yticks, labels=c(yticks))
A=density(end_values[,index_number],n=10000)
lines(A,lwd=2,col='blue')
lines(c(B[3],B[3]),c(1/7*max(A$y),0),col='red',lwd=5)
lines(c(B[1],B[1]),c(1/7*max(A$y),0),col='green',lwd=5)
lines(c(B[2],B[2]),c(1/7*max(A$y),0),col='green',lwd=5)
mtext(names(data)[index_number+1],3,line=0.2)
}
par(mar=c(0,0,0,0))
textplot('')
legend('top',legend=c('Density','5% & 95% quantile',
'50% quantile'),col=c('blue','green','red'),bty='n',pt.cex=1.5,
lwd=c(2,2,2),cex=1.6)
dev.off()

options(scipen=5)

#Relative changes in 5 year simulation
relative_changes=array(0,c(n_sim,10))
for (index_number in 1:10){
relative_changes[,index_number]=(simulation[261,index_number,]-
data_week[659,index_number+1])/data_week[659,index_number+1]
}

#SD on endvalues
sd(relative_changes)
```

```
#Histogram relative changes in scenarios
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/HIST_relative_end_values_Bootstrap.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(3.3,3.5,1.5,0.5),mgp=c(2,0.8,0),
cex.axis=1.6, cex.main=1.8, cex.lab =1.6,lwd=2)
textplot('Histogram of rela-\ntive changes in\n5 year simulations',
valign='top')
for (index_number in 1:10){
#Mean=mean(relative_changes[,index_number])
A[index_number]=quantile(relative_changes[,index_number],
probs = c(0.50))
C=max(relative_changes[,index_number])-
min(relative_changes[,index_number])
hist(relative_changes[,index_number],breaks=C*30/(B[2]-B[1]),
xlim=c(B[1],B[2]),prob=TRUE,main='',xlab='')
A=density(relative_changes[,index_number],n=10000)
lines(A,lwd=2,col='blue')
lines(c(B[3],B[3]),c(1/7*max(A$y),0),col='red',lwd=5)
lines(c(B[1],B[1]),c(1/7*max(A$y),0),col='green',lwd=5)
lines(c(B[2],B[2]),c(1/7*max(A$y),0),col='green',lwd=5)
mtext(names(data)[index_number+1],3,line=0.2)
}
par(mar=c(0,0,0,0))
textplot('')
legend('top',legend=c('Density','5% & 95% quantile',
'50% quantile'),col=c('blue','green','red'),bty='n',pt.cex=1.5,
lwd=c(2,2,2),cex=1.6)
dev.off()

##Drawdown at time t
MaxDD=array(0,c(261,10))
quantile_DD_sim=array(0,c(5,261,10))
for (index_number in 1:10){
for (i in 1:261){
MaxDD[i,index_number]=max(DD[,1:i,index_number])
quantile_DD_sim[,i,index_number]=quantile(MDD[,i,index_number],
probs = c(0.05,0.25,0.5,0.75,0.95))
}
}

# MaxDD and drawdown(1:t)-plot
pdf(file="C:/Users/Emil/Documents/Skole/Bachelorprojekt/Project/
figures/MaxDD_plot_Bootstrap.pdf",width=8,height=6)
par(mfrow=c(4,3),mar=c(3.3,3.5,1.5,0.5),mgp=c(2,0.8,0),
```

```
cex.axis=1.6, cex.main=1.8, cex.lab =1.6,lwd=2)
textplot('Drawdown from\nsimulation-start \nuntil time t')
for(index_number in 1:10){
plot(MaxDD[,index_number],type='l',lwd=2,main='',
ylab='DD(t) [%]',xlab='Time, t [Year]',col='red',xaxt='n')
axis(1,(21+c(0:4)*52),labels=rep('',5))
axis(1,(21+c(0:4)*52),labels=c('2012','2013','2014','2015','2016'))
lines(quantile_DD_sim[3,,index_number],lwd=2,lty=1) #50%quantile
lines(quantile_DD_sim[2,,index_number],lwd=2,lty=2) #25%quantile
lines(quantile_DD_sim[4,,index_number],lwd=2,lty=2) #75%quantile
lines(quantile_DD_sim[1,,index_number],lwd=2,lty=3) #5%quantile
lines(quantile_DD_sim[5,,index_number],lwd=2,lty=3) #95%quantile
mtext(names(data)[index_number+1],3,line=0.2)
}
par(mar=c(0,0,0,0))
textplot('')
legend('top',legend=c('MDD(1:t)','50% quantile',
 '5 % & 95 % quantile','25 % & 75 % quantile'),
col=c('red','black','black','black'),bty='n',pt.cex=1.5,lwd=c(2,2,2,2),
lty=c(1,1,3,2),cex=1.6)
dev.off()
```

Bibliography

[1] Knut Conradsen Bjarne Kjær Ersbøll. *An Introduction to Statistics, vol. 2*. 7 edition, 2007.

[2] Zvi Bodie, Alex Kane, and Alan J. Marcus. *Investments and Portfolio Management*. McGraw-Hill/Irwin, ninth revised edition, 2011. ISBN 978-007-128914-6.

[3] Richard B. Darlington. Is kurtosis really "peakedness?". *The American Statistician*, 24(2):19–22, April 1970. URL: http://www.jstor.org/pss/2681925.

[4] Brian S. Everitt and Hothorn Torsten. *A Handbook of Statistical Analyses Using R*. Chapman & Hall/CRC, firsth edition, 2006. ISBN 978-1584885399.

[5] Frank J. Fabozzi. *Fat-Tailed and Skewed Asset Return Distributions: Implications for Risk Management, Portfolio Selection, and Option Pricing*. John Wiley and Sons, 2005. ISBN 9780471718864.

[6] Robert H. Frank and Ben S. Bernanke. *Principles of Economics*. McGraw-Hill/Irwin, fourth edition, 2009. ISBN 978-0-07-128542-1.

[7] Jurgen Franke, Wolfgang Karl Hardle, and Christian Matthias Hafner. *Statistics of Financial Markets - An Introduction*. Springer, 2010. ISBN 978-3-642-16520-7.

[8] John C. Hull. *Options, Futures, and other derivatives*. Prentice Hall/Pearson, eighth edition, 2011. ISBN 978-0-13-216494-8.

[9] I.T. Jolliffe. *Principal Component Analysis*. Springer, 2 edition, 2002. ISBN 978-0387954424.

[10] J.P.Morgan, 11 January 2012. URL: http://documents.financialexpress.net/literature/2584110.pdf.

[11] Susan G. Stafford Kevin McGarigal, Sam Cushman. *Multivariate statistics for wildlife and ecology research*. Springer, 200. ISBN 978-0-387-98891-7.

[12] Philip Lawton and Todd Jankowski. *Investment Performance Measurement - Evaluating and Presenting Results*. John Wiley and Sons, 2009.

[13] Lexicon, 11 January 2012. URL: http://lexicon.ft.com/Term?term=EMBI.

[14] Henrik Madsen. *Time Series Analysis*. Chapman & Hall/CRC, 2007. ISBN 978-1-4200-5967-0.

[15] Henrik Madsen, Jan Nygaard Nielsen, and Mikkel Baadsgaard. *Stratistics in Finance*. 1998.

[16] MSCI, 28 December 2011. URL: http://www.msci.com/products/indices/licensing/msci_japan/.

[17] MSCI, 28 December 2011. URL: http://www.msci.com/products/indices/tools/index.html.

[18] MSCI, 28 December 2011. URL: http://www.msci.com/resources/pdfs/Base_Dates.pdf.

[19] MSCI, 28 December 2011. URL: http://uk.ishares.com/stream/pdf/false/?url=/publish/repository/documents/en/downloads/index_factsheet_msci_north_america.pdf&forceShowDialogBox=true.

[20] NIST, 11 January 2012. URL: http://itl.nist.gov/div898/handbook/prc/section2/prc213.htm.

[21] NASDAQ OMX Nordic, 28 December 2011. URL: https://indexes.nasdaqomx.com/docs/Methodology_OMXNORDIC.pdf.

[22] Peter Nystrup. *Scenario generation for financial market indices*. 2012.

[23] OECD, 5 January 2012. URL: http://www.oecd.org/document/29/0,3746,en_2649_34349_35725597_1_1_1_1,00.html.

[24] OECD, 5 January 2012. URL: http://www.oecd.org/dataoecd/22/47/44728410.pdf.

[25] OECD, 5 January 2012. URL: http://www.oecd.org/document/6/0,3746,en_2649_34349_35726918_1_1_1_1,00.html.

[26] Kourosh M. Rasmussen. Scenario generation methods for financial optimization models. Slideshow from lecture in DTU 42123 course Optimization in Finance, 2011.

[27] Kevin Sheppard. *Financial Econometrics Notes*, 2010.

[28] Lindsay I Smith. A tutorial on principal components analysis, 2002. URL: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf.

[29] Ruey S. Tsay. *Analysis of Financial Time Series*. Wiley & Sons, 2002. ISBN 0-471-41544-8.

[30] UCL, 11 January 2012. URL: http://www.stat.ucl.ac.be/ISdidactique/Rhelp/library/tseries/html/garch.html.

[31] Wikipedia, 13 January 2012. URL: http://en.wikipedia.org/wiki/Drawdown_(economics).

[32] Wikipedia, 12 January 2012. URL: http://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test.

[33] Wikipedia, 11 January 2012. URL: http://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test.

[34] Wikipedia, 11 January 2012. URL: http://en.wikipedia.org/wiki/KPSS_test.

[35] Wikipedia, 12 January 2012. URL: http://en.wikipedia.org/wiki/Ljung%E2%80%93Box_test.

[36] Wikipedia, 17 January 2012. URL: http://en.wikipedia.org/wiki/Uniform_distribution_(discrete).