Unser Titel

- und Untertitel -

Project Report

Group: Nikos Bosse and Felix Süttmann

Georg-August-University Göttingen
Statistisches Praktikum

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Faculty of Business and Economics
Georg-August-University Göttingen
https://www.uni-goettingen.de/

Title:
Project Title

Theme:
Scientific Theme

Project Period:
Summer Semester 2018

Project Group:
6

Participant(s):
Nikos Bosse
Felix Süttmann

Supervisor(s):
Paul Wiemann
Peter Pütz
Stanislaus Stadlmann

Copies: 1

Page Numbers: 37

Date of Completion:
September 9, 2019

Abstract:

XXXXXX.

Contents

1. Introduction

1.1 Our Goal / Scope of this report - Nikos / Felix

General Idea:
1. Predict Stocks using • Time Series • Machine Learning and Sentiments from Text 2. Use Predictions to Implement Trading Strategies 3. Evaluate and Compare to Other Trading Strategies

2. The Data

The following chapter gives an overview over the data on which we based our analysis. The firms we selected will be introduced as well as the Ravenpack Sentiment Data and the Analyst Reports. Lastly the stock data will be explored in detail appropriate transformations of the data will be motivated.

2.1 Selection of Firms to be Analyzed - Felix / Nikos ?

The stock data comprises 10 selected companies from the NASDAQ stock index. The stocks were determined as those are the stocks we have Ravenpack data and analyst reports about.
Figure / Table: Unsere Unternehmen –> Identifier_df

2.2 Ravenpack Sentiment Data - Felix

Next to the financial time series data, sentiment data from a financial data provider RavenPack was acquired. According to their website RavenPack is a leading provider of big data analytics for financial services (Rav). Our specific data is from RavenPack News Analytics, a service providing actionable event and sentiment information from TIME to TIME. The sentiment information has intra day precision. As such events are recorded at their time of appearance and not just on a daily basis. The variables consist of an indicator for the estimated positive or negative sentiment score called ESS ranging from 0 to 100 where 50 is neutral. This goes along with an estimated novelty score called ENS (0-100) describing how 'new' the news are and a RELEVANCE (0-100) variable that shows how closely the information is related to the underlying news, this can be converged to the number of news per day. Additionally two variables contain a 90 day rolling summary of events. One displays the percentage of positive events over a 90 day rolling window (AES) and AEV is the sum of events over the past 90 days. All variables have substantial numbers of missing values ...

Stock Data

Date	Open	High	Low	Close	Adj Close	Volume	ticker
2012-01-03	83.76	84.44	83.36	83.49	68.41	3380100	MMM
2012-01-04	83.13	84.26	83.11	84.18	68.98	3007400	MMM
2012-01-05	83.53	83.87	82.70	83.80	68.67	3116400	MMM
...	...	...	...	...	...	...	...
2015-11-27	29.11	29.21	29.03	29.19	25.97	34469600	GE
2015-11-30	29.16	29.28	28.79	28.79	25.61	82905200	GE
2015-12-01	28.84	29.09	28.72	29.01	25.80	56414600	GE
...	...	...	...	...	...	...	...
2017-12-27	108.42	108.55	107.46	107.64	105.31	5624000	DIS
2017-12-28	108.00	108.05	107.06	107.77	105.43	3477700	DIS
2017-12-29	108.05	108.34	107.51	107.51	105.18	4538400	DIS

Table 2.1

2.3 Analyst Reports - Felix

As the second external data source for this analysis we used analyst reports provided by Thomson One. originally the aim was to use general financial news but we could not obtain such data. These reports are part of the "company research" data and contain general research documents like SWAT analysis or sector reports. Typical contents are earning per share prognosis or Buy-Hold-Sell-Recommendations. This data is either represented in a structured format or flowing text, mostly both. Analyst reports are written by analysts mostly working for brokers like "Deutsche Bank Research" and are commonly specialized on a specific industry. Each report is linked to a specific stock and was cleaned to obtain only the flowing text. All illustrations and tables where thereby disregarded. This was done using a PDF-tool which converts PDF to .xlsx based on specific rules like the relationship of words to numbers or special characters.

1. Missing data
2. clustering of the data

2.4 Stock Data - Nikos

2.4.1 Overview over the Stock Data - Nikos

The stock data were downloaded from Yahoo Financial Data Base. Table 2.1 provides a small overview over the raw data. Figure 2.1 shows the Closing Prices of the selected assets. The closing prices have been provided adjusted for dividends by Yahoo.
Time series analysis is easiest with data that are at least weakly stationary. Weak stationarity implies that the mean of the time series is constant over time and that the covariance between two observations y_t and y_{t+h} depends only on h, not on t (see [Shumway and Stoffer 2011]). From figure 2.1 it can be clearly seen that most of the stocks exhibit a strong trend. Also the variance of most stocks increases steadily with time over the observed period. This increase in variance is illustrated in figure 2.2 where the standard deviation of the time series is shown. The data are therefore clearly not stationary. We can also formally test for stationarity using the augmented Dickey-Fuller test [REFERENCE]. This test, applied to all time series in no case is able to reject the null hypothesis of a unit root (implying non-stationarity) at any reasonable confidence level.

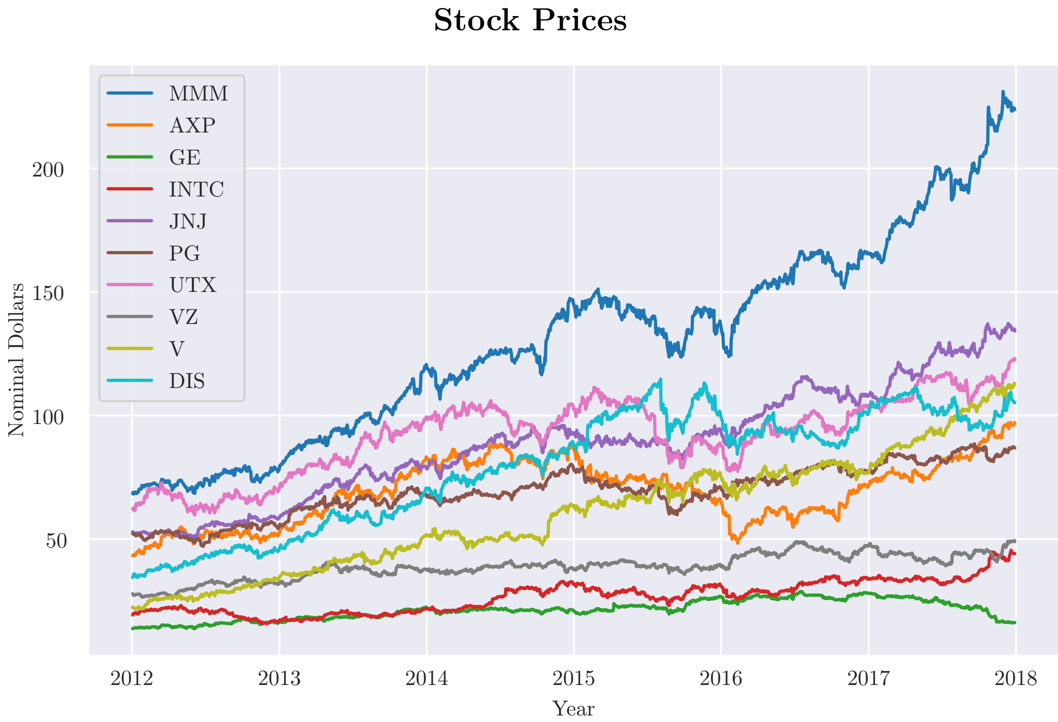

Figure 2.1: Time series of the adjusted closing prices of all 10 stocks looked at in this paper

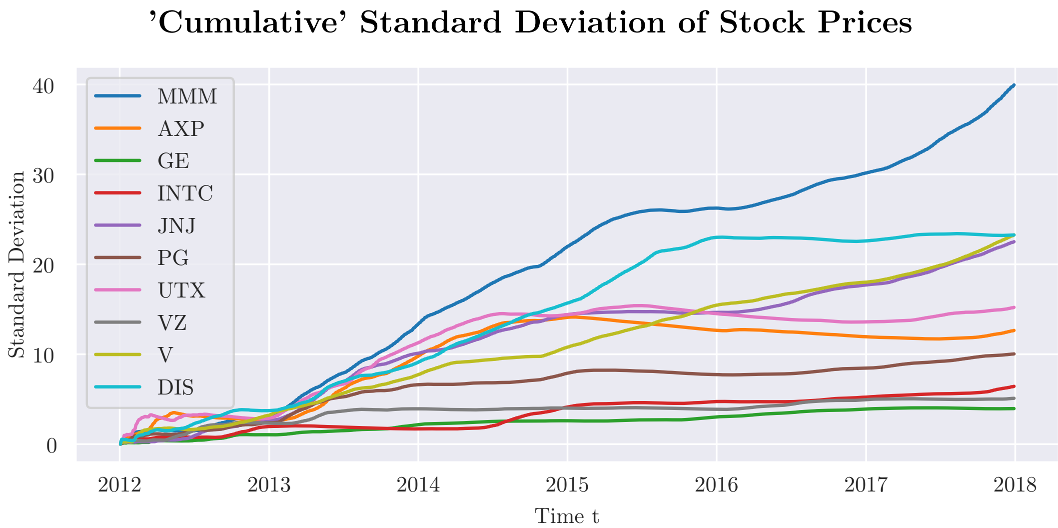

Figure 2.2: Standard deviation for the time series of stock prices. The value of the graph at point t is calculated as the standard deviation of all recorded values of the respective stocks up to that point t.

2.4.2 Data Transformation

In order to obtain weakly stationary time series the data needs to be transformed. There are different ways to proceed that are often equivalent or very similar to each other. Economists usually work with either returns or log-returns, albeit the nomenclature may be a bit confusing. (Daily) returns can be calculated as

$$r_t^{(1)} = \frac{r_t}{r_{t-1}} \qquad \text{or as} \qquad r_t^{(2)} = \frac{r_t - r_{t-1}}{r_{t-1}} = r_t^{(1)} - 1$$

Usually $r_t^{(2)}$ is used and is called either returns or log-returns, even though no logging takes place. For increased conceptual clarity, in this report $r_t^{(1)}$ will be called returns and $log(r_t^{(1)})$ will be called log-returns, while $r_t^{(2)}$ will not explicitly be used at all. Log-returns are computationally convenient and numerically stable. For very small values they are also very close to the returns $r_t^{(2)}$ often used in economic literature as $\log(r_t^{(1)}) \approx r_t^{(1)} - 1 = r_t^{(2)}$ for values of $r_t^{(1)}$ close to 1. Using returns or log-returns instead of stock prices can make the time series stationary. Figure 2.3 illustrates that the trends in the time series have vanished after looking at log-returns. The data visually now looks like white noise.

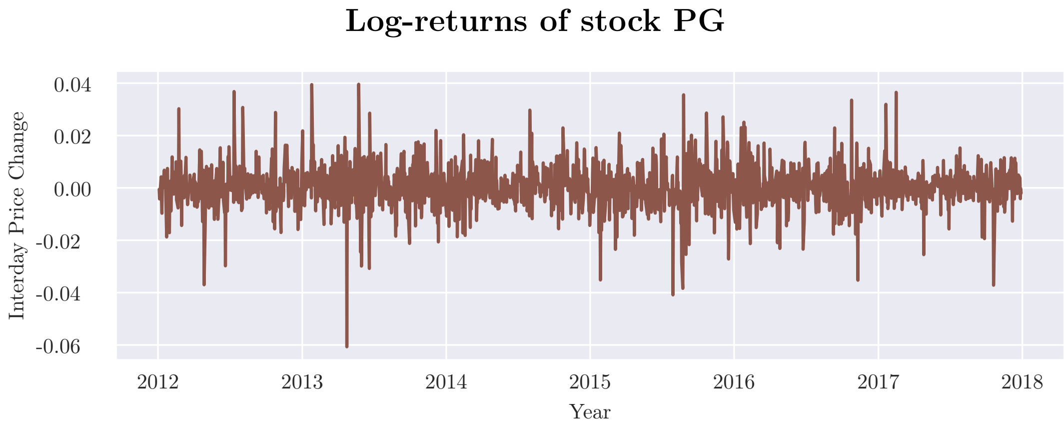

Figure 2.3: Log-returns (or equivalently, first differences of the log of adjusted closing prices) of the PG. Visually the data looks similar to white noise. The entire data can be seen in the appendix in figure A.3

Motivating the Transformation by Looking at Autocorrelation

That this is indeed a suitable transformation can be showed by looking at a different route of transforming the data: taking the log of the stock prices and then using the first difference of the logged time series. Logging the time series transforms an exponential trend in the time series into a linear one and also serves to stabilize the variance. However, the trend does not vanish and after the log-transformation, the log value of a stock at time t is still mostly determined by its log-value at time t - 1. To remove this effect, the time series needs to be differenced. To put this into a clearer perspective we can look at the autocorrelation and partial autocorrelation function of the series.
The autocorrelation at lag j is the correlation between an observation at time t with the observation at time t - j. For a stationary series, the autocorrelation does not depend on t, but only on the number of periods that lie between one observation y_t and another y_{t+h}. The autocorrelation function (ACF) can then be expressed as

$$\mathrm{ACF}(h) = corr(y_t, y_{t+h}) \tag{2.1}$$

Partial autocorrelation between an observation y_t and another observation y_{t+1} is the correlation between y_t and y_{t+h} that is not already explained by a linear dependence on the observations in between y_t and y_{t+h}. Formally the partial autocorrelation function (PACF) can be defined as

$$\mathrm{PACF}(h) = corr(y_t - \hat{y}_t, y_{t+h} - \hat{y}_{t+h}) \tag{2.2}$$

where $\hat{y}_{t+h} = \beta_1 y_{t+h-1} + \beta_2 y_{t+h-2} + ... + \beta_{h-1} y_{t+1}$
and $\hat{y}_t = \beta_1 y_{t+1} + \beta_2 y_{t+2} + ... + \beta_{h-1} y_{t+h-1}$

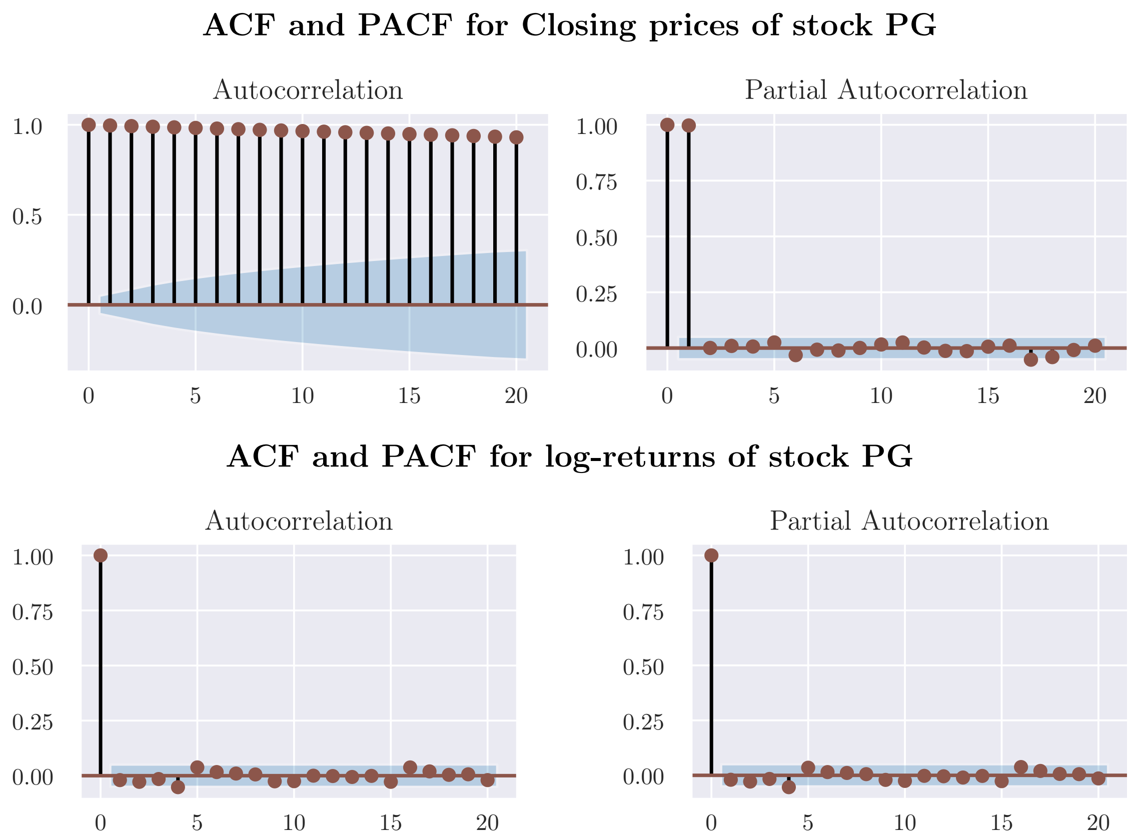

Figure 2.4: Autocorrelation function (ACF) and partial autocorrelation function (PACF) for the Adjusted Closing Prices (top) and log-returns (bottom) of PG. (For convenience, only one stock is shown. ACF and PACF for other stocks can be seen in the appendix in figure A.2 and A.4)

are the linear combinations $\{y_{t+1}, ..., y_{t+h-1}\}$ that minimize the mean squared error of a regression of y_{t+h}, and y_t respectively, on $\{y_{t+1}, ..., y_{t+h-1}\}$. Both $y_t - \hat{y}_t$ and $y_{t+h} - \hat{y}_{t+h}$ are uncorrelated with $\{y_{t+1}, ..., y_{t+h-1}\}$.
An ACF which very slowly decays to zero is an indicator that differencing may be appropriate (see Shumway and Stoffer 2011, p. 145) to make the series stationary. A large partial autocorrelation at lag 1, as shown in figure 2.4 also supports the conjecture that the dependence of the current on the previous value almost exclusively depends on the first lag and can be eliminated through differencing. After differencing we arrive again at the log-returns as $\log r_t^{(1)} = \log \frac{y_t}{y_{t-1}} = \log y_t - \log y_{t-1}$. We can see now in figure 2.4 that the the partial autocorrelation at lag 1 has vanished after differencing and that the autocorrelation has also dropped to insignificance. We can also see that we have not induced any negative autocorrelation, the data therefore is not overdifferenced. The means of our time series is very close to zero (as shown in table 2.2).

Means of the log-returns for all Stocks

	MMM	AXP	GE	INTC	JNJ	PG	UTX	VZ	V	DIS
mean	0.000786	0.000534	0.000110	0.000548	0.000616	0.000337	0.000450	0.000372	0.001068	0.000739

Table 2.2

Overall this pattern strongly suggests the time series are stationary now. We can also perform a formal test whether our data is stationary or not. The augmented Dickey-Fuller test (ADF) EXPLANATION TEST. P-values of the ADF for all log-returns are smaller than 10^{-12} even after correcting for multiple testing so we can safely assume the time series are stationary. The data is, however, not normally distributed. By looking at figure 2.5 we can clearly see that the distribution has fat tails: extreme events appear more often than would be expected if the data was normally distributed.

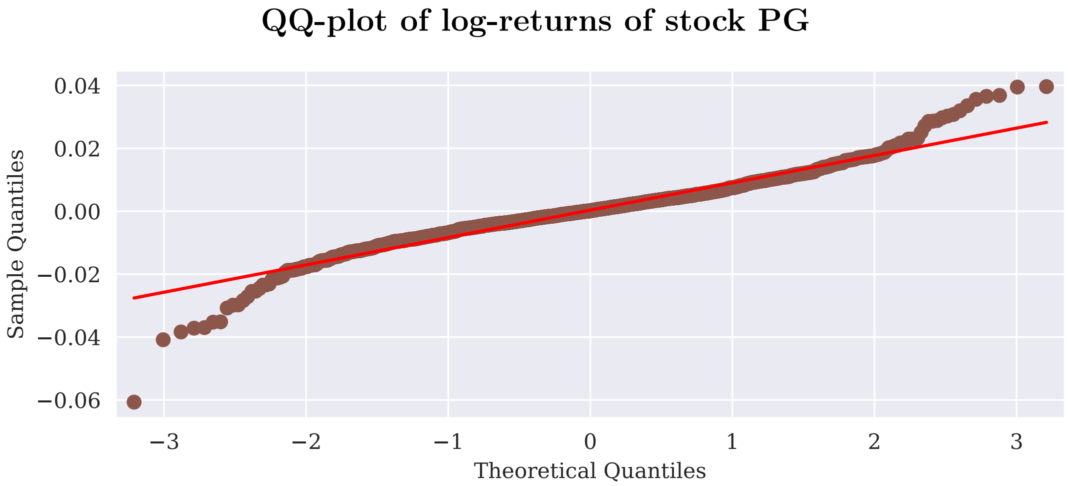

Figure 2.5: QQ-Plot for log-returns of stock PG. QQ-plots for the other stocks can be seen in the Appendix in figure A.5.

The time series are also not homoscedastic. While stationarity implies that the unconditional variance is constant over time, the variance of the time series fluctuates conditional on past observations [Beschreibung nachgucken, QUELLE!]. This conditional heteroskedasticity is quite common in financial data. (SOURCE). The pattern can be observed in figure 2.6. One can try to model this volatility by assuming a time series model for the conditional variance. This will be discussed in chapter 4.

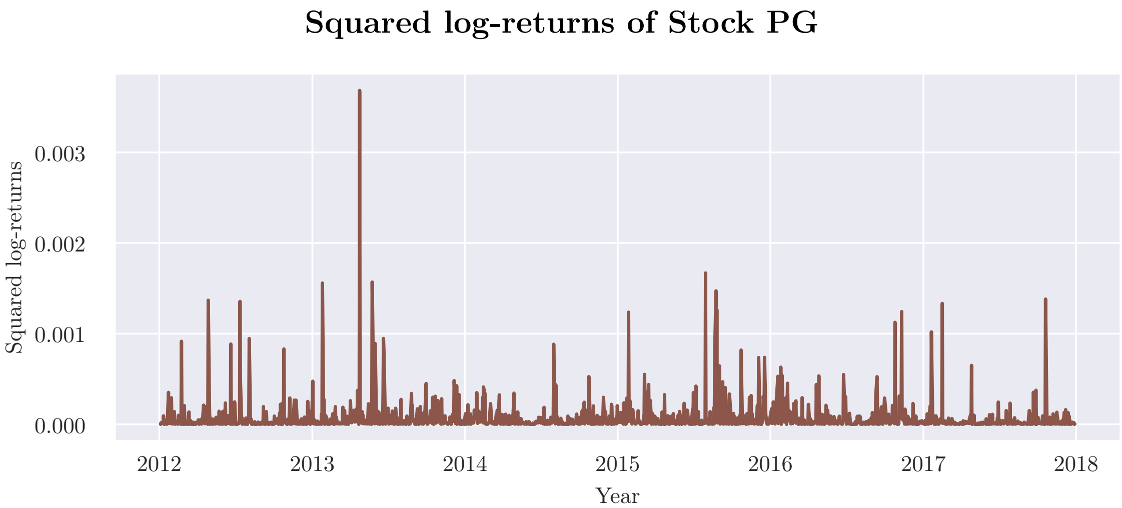

Figure 2.6: Plot of squared log returns of stock PG. This serves as an approximation of the variance of the log returns, as $Var(x) = E[(x - E(x))]^2$ and the mean of the log returns is close to zero. The pattern looks similar for all stocks, therefore only one is shown. In the Appendix in figure A.1 the ACF and PACF of the squared log-returns of some selected time series can be seen.

3. Predicting Stocks With Machine Learning - Felix

The movement of time series data for financial data is influenced by external effects. To leverage this we tried to add information from news sources to our trading strategies. For the hybrid prediction we estimate sentiment scores on news sources. The original aim was to use financial news data to predict stock price movement and volatility for trading strategies. To achieve this, large amounts of text data would need to be preprocessed and analyzed regarding their connections to specific stocks, their topic and sentiment. The news data would need to be as precise as possible, because [. . .] mention that an effect on the stocks an only be measured up to 20min after the news appear. Other sources say that. . . . As we were not able to acquire access to a reliable and precise news sources, we tried to implement our approach on the available analyst reports regarding the ten specific stocks. The problem with these reports is, that they are more an indicator of performance over the past month and a prediction about the future performance. As such they do not cover sudden events that would be present in the news. The reports also cluster around certain dates with long stretches of no or very few reports in between (ABBILDUNG). This makes it unlikely that they are valuable for trading strategies. The goal was to identify the connection of specific articles to listed companies and compute a sentiment score for the article. There are many ways to calculate sentiment scores from flowing text data. Common procedures would be to use a library of previously known positive and negative adjectives or 4-grams. The simplest way to utilize these would be to simply count their appearance, this is known as a bag-of-words method (ZITAT?). Other approaches extract parts of the text at the location of the specific adjectives and use Support Vector Machines or Naive Bayes Classifier to extract sentiment, see Westerski (2007) for further references. Many of these more advanced sentiment classification techniques are supervised, as such the need a labelled data set for initial training. The (NUMBER #) analyst reports are unstructured and not labelled making it unrealistic to use these methods. They also have a very specific format and language, therefore other pretrained models or other labelled training data sets could not be used. Another possibility to custom label the data would have been possible using intra day trading and news data. By looking at the movement or volatility of the period close after the news release approximate sentiment scores can then be computed (Robertson et al., 2007). As the obtained stock data is only inter day we could not apply this method.
To get around these restrictions unsupervised methods where chosen for the estimation of sentiment scores. A common approach for unsupervised
Analyst report data beschreiben...
To get reliable sentiment scores text data has to be preprocessed. The preprocessing was done using `R` (R Core Team, 2017). At first words where converted to lowercase and tokenized using the R package *tidytext* (Silge and Robinson, 2016). Next all the stop words where removed using the stop word library from the *tidytext* package, as well as a custom set. In the next step all links to websites, hyper-references, numbers and words with numbers are removed as well. The last step is lemmatizing the words using

the *textstem* package (Rinker, 2018). Lemmatizing words means reducing them to their inflectional forms. Commonly stemming is also applied, because words sometimes have derivationally related forms. This was not done to have more flexibility for the later applied text analysis. Additionally we could have also used the term frequency–inverse document frequency (tf-idf) matrix (ZITIEREN) for further reductions in the number of words. The issue here would have been that highly informative words for the stock sentiment could have been removed.

3.0.1 Feature engineering of sentiment scores

3.0.2 Sentiment Library

	Positiv	Negativ
1	acclaim	abandonment
2	accomplishment	abdication
3	advantage	abolish
4	assure	abrogation
5	attractiveness	abuse
6	good	accuse
7	breakthrough	acquittal
8	collaborator	adversary

3.0.3 Joint sentiment topic model

1. Weil wir nur wenige stocks haben kann es sein das das JST company specifix topics bildet

2. Schätzen von neuen texten möglich basierend auf den posterior wort probabilities

3.0.4 XG-Boost on sentiment data

- XGB kann auch einfach nur verwendet werden um zu erkennen welche features wichtig sind

XGBoost
Here I expalain what XGBoost is an how it works

Classification results
Take Mean of different predictions?
XGBOOST, LightGBM whatever. Theorie, Datenaufbereitung, Ergebnisse

3.0.5 XGB-Boost

4. Predictions Using Time Series - Nikos

4.1 Idea/Process and Evaluation - Nikos

irgendwas in der Richtung: wir benutzen Time Series Modelle, machen Predictions und gucken uns am Ende dann den Mean Squared Error an. Sinnvollerweise immer die ersten

	topic1sent1	topic2sent1	topic3sent1	topic28sent1	topic29sent1	topic30sent1
1	disney	report	visa	canaccord	piper	stanley
2	walt	market	volume	research	jaffray	morgan
3	network	disclosure	growth	genuity	report	research
4	park	capital	payment	investment	stock	instrument
5	medium	ebit	revenue	analyst	analyst	corp
6	company	stock	transaction	financial	rating	investment
7	cable	topeka	europe	limit	cover	information
8	espn	investment	debit	issuer	topeka	limit
9	entertainment	security	process	person	relative	datum
10	consumer	recommendation	incentive	author	sell	client
11	studio	information	credit	relevant	price	stock
12	resort	page	cross	affiliate	time	industry
13	revenue	locate	border	company	note	investor
14	film	issue	network	relate	disclosure	bank
15	sport	option	fee	discuss	company	international
16	disneys	company	international	client	distribution	month
17	content	margin	client	information	person	provide
18	product	financial	purchase	designate	coverage	coverage
19	broadcast	appendix	currency	trade	represent	management
20	advertise	research	card	distribute	locate	financial
21	star	base	global	month	market	sell
22	interactive	product	price	broke	horizon	disclosure
23	growth	instrument	expense	material	page	rate
24	time	share	datum	dealer	rate	universe
25	program	analysis	mastercard	australia	revenue	wwwmorganstanl
26	segment	rate	rate	provide	include	regulate
27	channel	contain	operate	express	fundamental	relevant
28	include	cover	impact	risk	target	trade
29	war	client	constant	disclosure	median	msci
30	theme	express	world	corp	opinion	weight

10 Perioden verwerfen, um den MSE vergleichbar zu machen zwischen allen Gruppen, auch denen, bei denen die ersten paar Perioden nicht definiert sind. In-Sample vs. Out of Sample Prediction?

4.2 Theoretical Overview - Nikos

Time series predictions in this paper will be made using Random Walks, autoregressive (AR), moving average (MA) and generalized autoregressive conditional heteroscedasticity (GARCH) models. The following will first give a short theoretical overview. Then different models will be applied to a training data set of two chosen stocks. Then predictions will be made for the other stocks using the above mentioned techniques.

4.2.1 Random Walks - Nikos du

Random walks serve as the baseline against which every prediction can be compared. Assuming a random walk as the underlying process implies that we know nothing about the future and can do no better than assuming tomorrow's stock price will on average be the same as today. Formally, a random walk follows

$$y_t = y_{t-1} + w_t \tag{4.1}$$

where y_t is the value of the time series at time t and w_t is a random realisation of a stationary white noise process with mean 0 and variance σ^2. We can expand equation 4.1 by allowing for a constant trend, a drift. A random walk with drift can be represented as

$$y_t = \delta + y_{t-1} + w_t \tag{4.2}$$

where δ is a drift parameter. Predictions for period t + 1 are therefore exactly the value at time t. As we have already eliminated the trend by transforming the data to log-returns we will not use the drift representation here. If we did our analysis on the original stock values then a drift would be appropriate. Note that the random walk (with or without drift) is not a stationary process.

4.2.2 Autoregressive Models - Nikos

An autoregressive process of order p (AR(p)) implies that the current value of a time series can be described as a combination of the previous p values plus a random shock. As those previous values intern depend on previous values, the current value is indirectly influenced by its entire past. Formally, an AR(p) process follows

$$y_t = \psi_1 y_{t-1} + \psi_2 y_{t-2} + ... + \psi_p y_{t-p} + w_t \tag{4.3}$$

where y_t is stationary, $\psi_1, ..., \psi_p$ are constants and w_t is white noise. The mean of y_t is assumed to be zero. If the mean is μ instead of zero, equation 4.4 can be rewritten as

$$y_t - \mu = \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + ... + \phi_p(y_{t-p} - \mu) + w_t \tag{4.4}$$

This can also be expressed as

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + w_t \tag{4.5}$$

with $\alpha = \mu(1 - \phi_1 - ... - \phi_p)$.

4.2.3 Moving Average Models - Nikos

A moving average process of order q implies that the current value of a time series consists of the average of the previous q observations plus a random shock. As the mean of the time series μ is constant this average can also be simply expressed as an average of the past random shocks $\{w_{t-1}, ...w_{t-q}\}$. In constrast to the AR(p) process, the shocks affect the future directly (and not only indirectly through past values) and only affect the next q values. Formally, the MA(q) process can be expressed as

$$y_t = \mu + w_t + \theta w_{t-1} + ... + \theta w_{t-q} \tag{4.6}$$

where w_t represents white noise and $\theta_1, ..., \theta_q$ are parameters and q is the number of lags in the moving average.

4.2.4 Autoregressive Moving Average Models - Nikos

Autoregressive Moving Average Models of order p and q (ARMA(p,q)) form a combination of the above described AR(p) and MA(q) models. Formally, an ARMA(p,q) process follows

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q} \tag{4.7}$$

if the mean of y_t is μ, then the above results in

$$y_t = \alpha + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q} \tag{4.8}$$

with $\alpha = \mu(1 - \phi_1 - ... - \phi_p)$.

4.2.5 GARCH Models - Nikos

The GARCH(p,q) model is specified as follows:

$$r_t = \sigma_t \epsilon_t \tag{4.9}$$

where ϵ_t is Gaussian white noise with $\epsilon_t \sim \mathcal{N}(0,1)$ and

$$\sigma_t^2 = \alpha_0 + \underbrace{\alpha_1 r_{t-1}^2 + ... + \alpha_p r_{t-p}^2}_{\text{autoregressive part}} + \underbrace{\beta_1 \sigma_{t-1}^2 + ... + \beta_q \sigma_{t-q}^2}_{\text{moving average part}} \tag{4.10}$$

In equation 4.9 the returns r_t are therefore modelled as white noise with mean zero and variance variance σ_t^2. When compared to white Gaussian noise with constant variance this can produce a leptokurtic (fat-tailed) distribution similar to what we observed in the QQ-Plots in figure 2.6. Equation 4.9 is called the mean model of the GARCH(p,q) process. This mean model can also be altered as needed. The GARCH model can then be specified in the following way:

$$r_t = x_t + y_t \tag{4.11}$$

where x_t can be any constant mean, regression or time series process and y_t is a GARCH process that satisfies equations 4.9 and 4.10. In a similar way, the distribution of ϵ_t can be altered. In praxis, researchers often assume a t-distribution instead of a standard normal distribution.

A further expansion is the GJR-GARCH model (SOURCE). The GJR-GARCH model includes a separate term for past negative shocks. GJR-GARCH(1,1,1) is specified as follows

$$\sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \gamma_1 \epsilon_{t-1}^2 I(\epsilon_{t-1}^2 < 0) + \beta_1 \sigma_{t-1}^2 \tag{4.12}$$

where $I(\epsilon_{t-1}^2 < 0)$ is an indicator function that is one if ϵ_{t-1}^2 is negative and zero else. This means that negative shocks may have a different impact on future volatility than positive shocks, e.g. a sudden drop in a stock will cause the stock to be disproportionally more volatile in the near future.

4.3 Full Example Analysis with Training Data - Nikos

We start with a subselection of the available stocks and do a full analysis. Our aim is to get a feeling of what might work for in terms of analysis and predictions and to avoid overfitting by working on the very same data we are trying to predict. In the following we will analyze the stocks of Visa (V) and Intel (INTC).

4.3.1 Data Exploration - Nikos

Eventuell noch plot mit den log-returns / der gesamten Zeitreihe für V und INTC.

Looking at Autocorrelation - Nikos

Figure 4.1 shows the ACF and PACF for the log-returns of INTC and V. From looking at the plots we can presume that for V, AR and MA models of order one or two might be a reasonable try. For INTC it looks like there is very little information included as none of the lower order lags bear any significance.

We start by applying a formal test for autocorrelation to the time series, the Box/Pierce and Ljung/Box tests. EXPLANATION / Source. They have slightly different properties regarding their handling of very large and very small numbers of observations, but for both the null hypothesis is that there is no autocorrelation in the series. Figure 4.2 shows the p-values for the first 40 lags. The test suggests that for V there may be some significant autocorrelations that could be used for modeling. Compared to the plot of ACF and PACF however, the test seems overly optimistic. For INTC, the test confirms non-significance for the first few lags. While higher order lags may be significant, modeling a time series process with that many coefficients is almost certain to overfit the data.

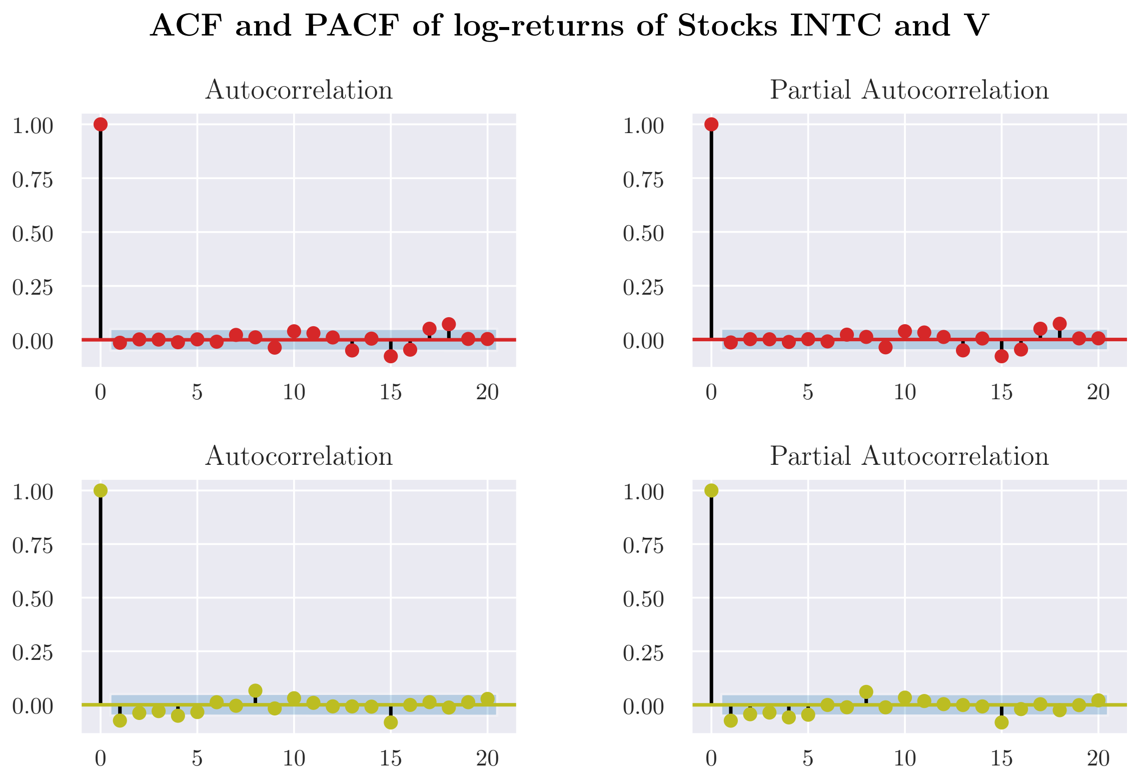

Figure 4.1

4.3.2 Applying ARMA models

We start by fitting ARMA models to the time series. This turned out to be prone to numerical instability. While the BIC could always be calculated, for some of the models standard errors could not be computed as the algorithm was not able to invert the Hessian matrix. The underlying problem is exacerbated when dealing with GARCH models, as all residuals are squared. We eventually managed to alleviate the problem by multiplying log-returns by 100 (corresponding to an approximate percentage interpretation). Table

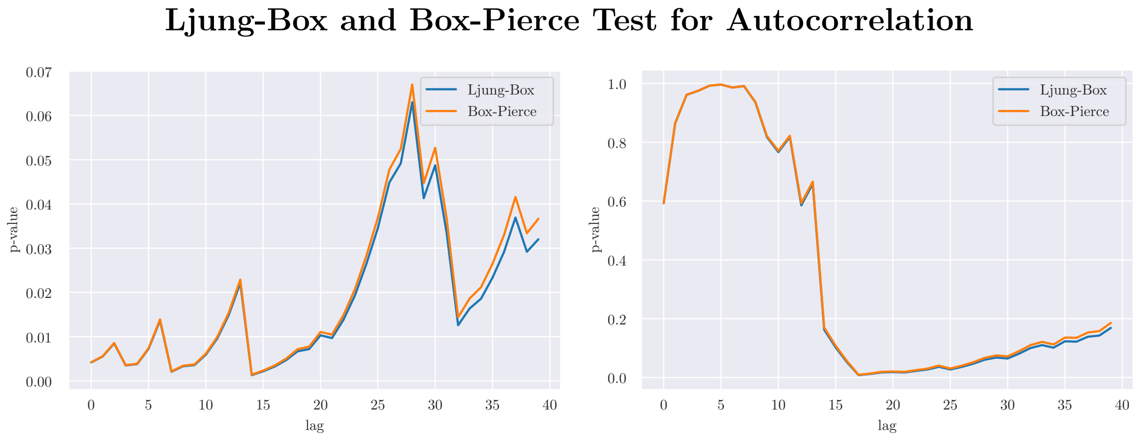

Figure 4.2

4.1 shows the BIC that were obtained from fitting different ARMA(p,q) models to the series of log-returns. To allow for comparison later on, models were fitted once with the original log-returns and once with the log-returns multiplied by 100.
For V the best model according to BIC is ARMA(1,1). However the difference to ARMA(0,0) seems as good as negligible. In order to avoid the peril of reading too much into random chance, it may therefore be more prudent to stick with a constant mean model (ARMA(0,0)). Recall that commons sense suggests that the past should not contain exploitable information about the future, which, however is an implication of assuming an ARMA(1,1) model. We will fit this model nevertheless, along with the constant mean model. For INTC the lowest BIC is reached by ARMA(0,0) which corresponds to our observation that there is no significant lower-level autocorrelation.
Table 4.3 shows the results for V. The change in AIC is larger than the change in BIC as the latter more heavily penalizes the number of parameters included in the estimation. Notably, even though the difference in BIC is quite small, the AR(1) and MA(1) are estimated very significantly and are high in relative magnitude. However, note that their effects go in opposing directions in similar magnitude and may as well cancel each other out. This interpretation may be somewhat plausible as the individual effects of the AR(1) and MA(1) terms are about an order of magnitude smaller (albeit still significant) and go in the same direction when estimating ARMA(1,0) and ARMA(0,1) models separately (-0.0736 for the AR(1) and -0.0809 for the MA(1) term in separate models). Table 4.4 shows the result of the ARMA(0,0) model to the log-returns of INTC. Not even the mean is significant, which is not too surprising, as Intel hardly gained in value in the observed period from 2012 to 2017 (See again figure 2.1. When exploratively fitting an ARMA(1,0), ARMA(0,1) or ARMA(1,1) model, as expected none of the coefficients reach significance (p-values all $>$ 0.5).

[ARMAX models]

We proceed in our analysis by adding our own generated sentiments as well as the ravenpack sentiments as external regressors. To make matters more complicated the stock market observations don't perfectly match the external data data points. We have about a third as many analyst reports (and therefore sentiment scores) as stock market observations. In order to obtain an estimable model we decided to discard all observations on days where no analyst reports existed. This also implies that we need to confine our analysis to a constant mean model (ARMA(0,0) as autoregressive and moving-average

BIC for different combinations of ARMA(p,q) for V

(p,q)	0	1	2
0	-8870.52	-8872.19	-8868.22
1	-8871.40	**-8872.93**	-8865.64
2	-8866.97	-8865.65	-8858.30
3	-8861.59	-8859.10	-8851.28
4	-8859.42	-8854.27	-8854.01

(p,q)	0	1	2
0	5027.88	5026.20	5030.17
1	5026.99	**5025.46**	5032.75
2	5031.42	5032.75	5039.26
3	5036.81	5039.29	5037.07
4	5038.98	5044.13	5044.39

BIC for different combinations of ARMA(p,q) for INTC

(p,q)	0	1	2
0	**-8692.98**	-8685.94	-8678.63
1	-8685.94	-8678.62	-8671.31
2	-8678.63	-8671.30	-8672.74
3	-8671.31	-8664.08	-8664.26
4	-8664.18	-8656.92	-8652.19

(p,q)	0	1	2
0	**5205.42**	5212.45	5219.77
1	5212.45	5219.77	5227.09
2	5219.77	5227.09	5226.21
3	5227.09	5234.32	5238.91
4	5234.21	5241.47	5246.20

Table 4.1: BIC presented for different combinations of ARMA(p,q) fit to the log-returns of V (top) and INTC (bottom). On the right side, those returns were multiplied by 100 in order to allow for comparison with the GARCH models later on.

components do not make sense when many observations in the time series are missing. The time series model hence reduces to a regression with an intercept and the sentiment data as independent variables. While analyst reports were rather scarce, the Ravenpack data usually had multiple entries per day with only few days missing. We therefore needed to aggregate sentiments to obtain one single observation per day.

ARMAX Sentiments Analyst Reports

The fact that we need to omit two-thirds of all data points changes the BIC considerably. To be able to compare the models according to BIC we have refit the baseline ARMA(0,0) model to the reduced data and obtained a BIC of -3001.07 for V and -4533.63 for INTC (the difference owing to the different number of observations). Table 4.5 shows the results of the ARMAX model for V, table 4.6 shows the results for INTC. Even though sent_mean is not too far from significance for V (and is even closer when only including sent_mean, with a p-value of 0.067), all models fare worse in terms of BIC than the baseline.

ARMAX Sentiments Ravenpack

In the Ravenpack data for V, 21 observations out of 1509 were missing. We decided it might still be interesting to continue having a look at ARMA(1,1) even though the estimation is now mildly distorted by the fact that some values are missing. We computed a new baseline BIC as we did in the previous analysis. The BIC for an ARMA(0,0) model with the 21 omitted values is now -8734.35 and for an ARMA(1,1) model on the same data -8735.23. For INTC there were no missing observations, so the baseline BIC stayed -8692.98. Table 4.7 shows the result for the ARMAX(0,0) model fit with Ravenpack sentiments to V, table 4.8 the results for INTC. We have tried different combinations of regressors to include in the model but only show the full model to avoid redundancy. The results for ARMAX(1,1) for V are not shown but look very similar. In all cases the external information worsened BIC and did not improve the fit. For the following

Results for an ARMA(1,1) process fit to the log-returns of V

Dep. Variable:	log_returns	**No. Observations:**	1509
Model:	ARMA(1, 1)	**Log Likelihood**	4451.108
Method:	css-mle	**S.D. of innovations**	0.013
Date:	Tue, 03 Sep 2019	**AIC**	-8894.215
Time:	13:15:26	**BIC**	-8872.938
Sample:	0	**HQIC**	-8886.291

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0011	0.000	4.194	0.000	0.001	0.002
ar.L1.log_returns	0.6156	0.116	5.307	0.000	0.388	0.843
ma.L1.log_returns	-0.6997	0.105	-6.681	0.000	-0.905	-0.494

	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	1.6243	+0.0000j	1.6243	0.0000
MA.1	1.4293	+0.0000j	1.4293	0.0000

Table 4.2

Results for an ARMA(0,0) process fit to the log-returns of V

Dep. Variable:	log_returns	**No. Observations:**	1509
Model:	ARMA(0, 0)	**Log Likelihood**	4442.581
Method:	css	**S.D. of innovations**	0.013
Date:	Wed, 04 Sep 2019	**AIC**	-8881.162
Time:	10:03:08	**BIC**	-8870.524
Sample:	0	**HQIC**	-8877.200

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0011	0.000	3.253	0.001	0.000	0.002

Table 4.3

GARCH analysis we will therefore not include external information.

GARCH models

Financial data often exhibit conditional heteroskedasticity. We therefore want to see whether GARCH models are able to improve the fit. Figure 4.3 shows the squared log-returns of V and INTC as well as their ACF and PACF. For both the first lag visually seems to be significant.
We formally test for the presence of a GARCH effect, i.e. autocorrelation in the squared residuals of the time series using the Lagrange Multiplier Test proposed by Engle (SOURCE). For both tests, the null hypothesis of 'No ARCH effect' cannot be rejected with p-values of 0.1160 (V) and 0.2693 (INTC). We nevertheless proceed to fit a GARCH model to the data. There is extensive literature on the optimal order for a GARCH model (e.g. does anything beat GARCH(1,1), does anything not beat GARCH(1,1), does anyone need

Results for an ARMA(0,0) process fit to the log-returns of INTC

Dep. Variable:	log_returns	**No. Observations:**	1509
Model:	ARMA(0, 0)	**Log Likelihood**	4353.809
Method:	css	**S.D. of innovations**	0.014
Date:	Wed, 04 Sep 2019	**AIC**	-8703.618
Time:	10:03:08	**BIC**	-8692.979
Sample:	0	**HQIC**	-8699.656

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0005	0.000	1.574	0.116	-0.000	0.001

Table 4.4

ARMAX(0,0) with analyst report sentiments fit to the log-returns of V

Dep. Variable:	log_returns	**No. Observations:**	535
Model:	ARMA(0, 0)	**Log Likelihood**	1508.522
Method:	css	**S.D. of innovations**	0.014
Date:	Wed, 04 Sep 2019	**AIC**	-3009.044
Time:	17:06:21	**BIC**	-2991.915
Sample:	0	**HQIC**	-3002.342

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0045	0.002	2.217	0.027	0.001	0.009
sent1_mean	-0.0047	0.003	-1.602	0.110	-0.010	0.001
sentBoW_mean	-0.0007	0.003	-0.241	0.810	-0.006	0.005

Table 4.5

GARCH(1,1)?).

For simplicity, we will stick to a GARCH(1,1) model with a constant mean model (the mean is modeled by an ARMA(0,0) process). While we have tried other models such as for example GARCH(3,1) they did not perform significantly better than GARCH(1,1). To avoid issues with numerical instability all log-returns are multiplied by 100 and the new baseline BIC become 5025.46 for V (ARMA(1,1)) and 5205.42 for INTC (ARMA(0,0)) as shown on the right hand side of 4.1. Table 4.9 shows the result for a GARCH(1,1) fit to V, table 4.10 shows the result for INTC. In both cases the fit improves and the BIC drops - not too much, but noticeably (from 5025.46 to 4948.68 (V) and from 5205.42 to 5169.75 (INTC)).

We now relax the assumption of normally distributed error terms ϵ_t (see 4.9) and instead assume a Student t-distribution with fatter tails. This time the BIC drops considerably (from 4948.68 to 4794.91 (V) and from 5169.75 to 4986.41 (INTC)). The results are shown in table 4.11 and table 4.12.

Another expansion is the inclusion of asymmetric shocks (see **??**). This assumes that sudden drops in a stock's value lead to higher volatility than an increase of the same

ARMAX(0,0) with analyst report sentiments fit to the log-returns of INTC

Dep. Variable:	log_returns	**No. Observations:**	821
Model:	ARMA(0, 0)	**Log Likelihood**	2274.402
Method:	css	**S.D. of innovations**	0.015
Date:	Wed, 04 Sep 2019	**AIC**	-4540.805
Time:	17:06:22	**BIC**	-4521.963
Sample:	0	**HQIC**	-4533.575

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0031	0.002	1.766	0.078	-0.000	0.006
sent1_mean	-0.0035	0.003	-1.215	0.225	-0.009	0.002
sentBoW_mean	0.0024	0.002	0.993	0.321	-0.002	0.007

Table 4.6

amount would. Results for the GJR-GARCH model are shown in tables 4.13 and 4.14. For V, the gamma-coefficient is positive and significant, indicating an asymmetric effect of shocks on volatility exists. The BIC drops accordingly by a small bit from 4794.91 to 4784.75. For INTC, the gamma coefficient is not significant and BIC even slightly increase from 4986.41 to 4993.23.

Forecasting and Forecast Precision

To assess the predictive performance we use the ARMA(0,0) model for V and INTC as well as the ARMA(1,1) model for V to generate predictions. To do so, we fit an ARMA model to the time series up to time y_t, predict the next value $\hat{y}_{t+1}$, then add the true value y_{t+1} to the time series and fit a model that lets us predict $\hat{y}_{t+2}$ and so forth. For the ARMA(0,0) model, this means predicting the next value on the basis of a cumulative mean of the past log-returns. Estimating an ARMA(1,1) consecutively for V turned out to be problematic because the result was not always a stationary process. Therefore, at some points the estimation of an ARMA(1,1) model was impossible. We therefore had to replace the prediction for those specific points with ARMA(0,0). To compare the results we compared the mean squares error (MSE) as well as a binary prediction accuracy that merely judges whether the direction of the prediction was correct. The predictions of the ARMA(0,0) model with constant mean (which corresponds to a random walk for the stock prices with drift) seem to be better then mere coin tosses. However they do not perform better than the naive strategy of always predicting an upwards movement of the stock (i.e. a positive return). Only the ARMA(1,1) model for V reaches a higher prediction accuracy. This might, however, also be due to chance. Table 4.15 shows the results of those predictions. The forecasts are shown in figure 4.4.

Forecasting Volatility

While GARCH models are not able to better predict future returns, they can help predict the volatility of future returns. Figure ?? gives an intuition to the usefulness of volatility forecasts. In the figure, the volatility of V predicted by an GJR-GARCH(1,1) process is plotted together with the absolute value of the log-returns of the stock. We can see that periods that visually look more volatile often coincide with a high predicted volatility.
This is useful for balancing portfolios, pricing options and XXX [THEORY!]. As the main

ARMAX(0,0) with Ravenpack sentiments fit to the log-returns of V

Dep. Variable:	log_returns	**No. Observations:**	1488
Model:	ARMA(0, 0)	**Log Likelihood**	4375.394
Method:	css	**S.D. of innovations**	0.013
Date:	Wed, 04 Sep 2019	**AIC**	-8736.788
Time:	17:23:40	**BIC**	-8699.652
Sample:	0	**HQIC**	-8722.948

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0124	0.012	1.061	0.289	-0.010	0.035
relevance	-0.0001	0.000	-1.037	0.300	-0.000	0.000
aes_min	-1.536e-05	2.42e-05	-0.634	0.526	-6.29e-05	3.22e-05
aes_max	1.781e-05	3.35e-05	0.531	0.595	-4.79e-05	8.35e-05
count	5.652e-06	1.41e-05	0.402	0.688	-2.19e-05	3.32e-05
aev_min	2.091e-07	2.53e-06	0.083	0.934	-4.76e-06	5.18e-06

Table 4.7

ARMAX(0,0) with Ravenpack sentiments fit to the log-returns of INTC

Dep. Variable:	log_returns	**No. Observations:**	1509
Model:	ARMA(0, 0)	**Log Likelihood**	4356.793
Method:	css	**S.D. of innovations**	0.013
Date:	Wed, 04 Sep 2019	**AIC**	-8699.585
Time:	17:23:41	**BIC**	-8662.351
Sample:	0	**HQIC**	-8685.718

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-0.0324	0.019	-1.747	0.081	-0.069	0.004
relevance	0.0003	0.000	1.670	0.095	-5.63e-05	0.001
aes_min	4.252e-05	2.68e-05	1.586	0.113	-1e-05	9.51e-05
aes_max	-4.529e-06	3.08e-05	-0.147	0.883	-6.49e-05	5.59e-05
count	-1.395e-06	6.02e-06	-0.232	0.817	-1.32e-05	1.04e-05
aev_min	-1.217e-06	1.83e-06	-0.667	0.505	-4.79e-06	2.36e-06

Table 4.8

focus of this paper is the prediction of returns, volatility forecasts will not be further discussed in detail.

4.3.3 Weighted Average of Predictions

a) of different time series models b) of time series and ML models

4.4 Time Series Predictions with the Remaining Stocks - Nikos

ARMA(0,0) ARMA(1,1)? Average predictions? Auto-ARMA

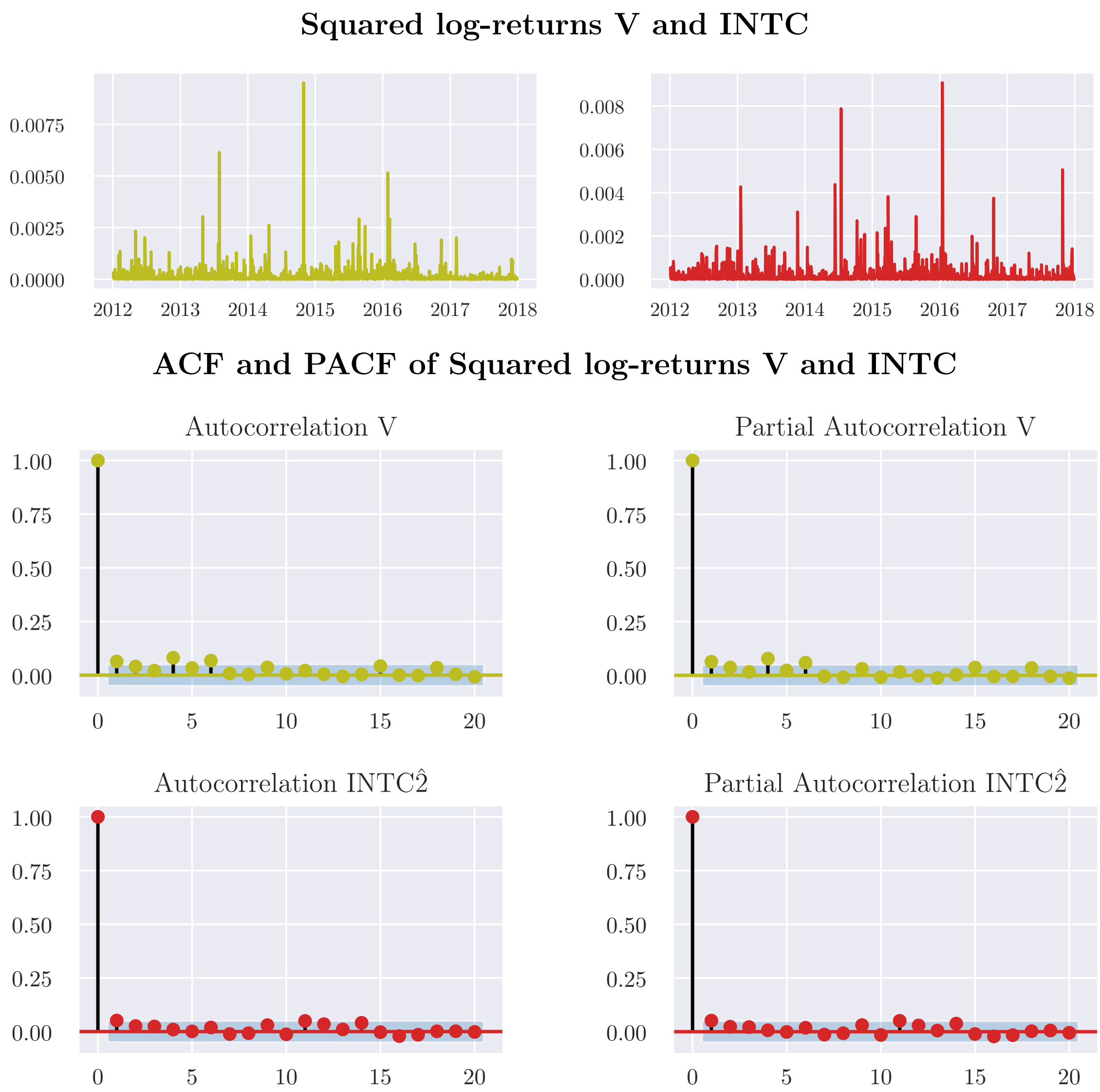

Figure 4.3: Squared log-returns and ACF and PACF for squared log-returns for V (top-left and middle) and INTC (top right and bottom).

Problem: We cannot look at the stocks beforehand.

5. Trading Strategies

5.1 Trading Strategies - Introduction and Theory

5.1.1 Idea and Process

We compare different Trading Strategies. With the Goal of Predicting Stocks this is interesting in and of itself. For the purpose of the project it also serves as a baseline to

Figure 4.4: Caption

Results for GARCH(1,1) with constant mean fit to the log-returns of V

Dep. Variable:	log_returns	**R-squared:**	-0.000
Mean Model:	Constant Mean	**Adj. R-squared:**	-0.000
Vol Model:	GARCH	**Log-Likelihood:**	-2459.70
Distribution:	Normal	**AIC:**	4927.40
Method:	Maximum Likelihood	**BIC:**	4948.68
		No. Observations:	1509
Date:	Wed, Sep 04 2019	**Df Residuals:**	1505

	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
mu	0.1199	2.971e-02	4.034	5.473e-05	[6.164e-02, 0.178]

	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
omega	0.0811	5.148e-02	1.576	0.115	[-1.979e-02, 0.182]
alpha[1]	0.0985	3.488e-02	2.823	4.758e-03	[3.010e-02, 0.167]
beta[1]	0.8585	5.388e-02	15.933	3.718e-57	[0.753, 0.964]

Covariance estimator: robust

Table 4.9

Results for GARCH(1,1) with constant mean fit to the log-returns of V

Dep. Variable:	log_returns	**R-squared:**	-0.000
Mean Model:	Constant Mean	**Adj. R-squared:**	-0.000
Vol Model:	GARCH	**Log-Likelihood:**	-2570.24
Distribution:	Normal	**AIC:**	5148.47
Method:	Maximum Likelihood	**BIC:**	5169.75
		No. Observations:	1509
Date:	Wed, Sep 04 2019	**Df Residuals:**	1505

	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
mu	0.0565	3.422e-02	1.650	9.897e-02	[-1.061e-02, 0.124]

	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
omega	0.9185	0.289	3.181	1.469e-03	[0.353, 1.484]
alpha[1]	0.2295	9.374e-02	2.448	1.436e-02	[4.576e-02, 0.413]
beta[1]	0.2977	0.168	1.772	7.642e-02	[-3.161e-02, 0.627]

Covariance estimator: robust

Table 4.10

compare our Trading Strategy against.

5.1.2 Mean Reversion Models and Momentum Trading - Theoretical Background

A vast body of scientific literature has tried to develop models that allow to understand and explain movements in stock markets. An even vaster community of traders has tried

Results for GARCH(1,1) with t-Distribution (V)

Dep. Variable:	log_returns	**R-squared:**	-0.000
Mean Model:	Constant Mean	**Adj. R-squared:**	-0.000
Vol Model:	GARCH	**Log-Likelihood:**	-2379.16
Distribution:	Standardized Student's t	**AIC:**	4768.31
Method:	Maximum Likelihood	**BIC:**	4794.91
		No. Observations:	1509
Date:	Wed, Sep 04 2019	**Df Residuals:**	1504

	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
mu	0.1150	2.551e-02	4.508	6.559e-06	[6.498e-02, 0.165]
	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
omega	0.1058	6.085e-02	1.739	8.196e-02	[-1.342e-02, 0.225]
alpha[1]	0.1297	4.143e-02	3.131	1.745e-03	[4.849e-02, 0.211]
beta[1]	0.8158	6.683e-02	12.208	2.823e-34	[0.685, 0.947]
	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
nu	4.9106	0.617	7.952	1.828e-15	[3.700, 6.121]

Covariance estimator: robust

Table 4.11

Results for GARCH(1,1) with t-Distribution (INTC)

Dep. Variable:	log_returns	**R-squared:**	-0.000
Mean Model:	Constant Mean	**Adj. R-squared:**	-0.000
Vol Model:	GARCH	**Log-Likelihood:**	-2474.91
Distribution:	Standardized Student's t	**AIC:**	4959.82
Method:	Maximum Likelihood	**BIC:**	4986.41
		No. Observations:	1509
Date:	Wed, Sep 04 2019	**Df Residuals:**	1504

	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
mu	0.0810	2.881e-02	2.810	4.950e-03	[2.449e-02, 0.137]
	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
omega	7.6664e-03	6.416e-03	1.195	0.232	[-4.909e-03,2.024e-02]
alpha[1]	0.0173	4.803e-03	3.604	3.137e-04	[7.896e-03,2.672e-02]
beta[1]	0.9792	5.563e-03	176.017	0.000	[0.968, 0.990]
	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
nu	4.1875	0.471	8.896	5.802e-19	[3.265, 5.110]

Covariance estimator: robust

Table 4.12

to implement these theories to do actual forecasting. While many of the theories are much

Results for GJR-GARCH(1,1) with t-Distribution (V)

Dep. Variable:	log_returns	**R-squared:**	-0.000
Mean Model:	Constant Mean	**Adj. R-squared:**	-0.000
Vol Model:	GJR-GARCH	**Log-Likelihood:**	-2370.42
Distribution:	Standardized Student's t	**AIC:**	4752.84
Method:	Maximum Likelihood	**BIC:**	4784.75
		No. Observations:	1509
Date:	Wed, Sep 04 2019	**Df Residuals:**	1503

	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
mu	0.0962	2.631e-02	3.657	2.549e-04	[4.465e-02, 0.148]
	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
omega	0.0857	5.488e-02	1.561	0.118	[-2.187e-02, 0.193]
alpha[1]	0.0347	2.612e-02	1.327	0.185	[-1.653e-02,8.585e-02]
gamma[1]	0.1630	5.265e-02	3.096	1.961e-03	[5.981e-02, 0.266]
beta[1]	0.8411	6.446e-02	13.050	6.383e-39	[0.715, 0.967]
	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
nu	5.1057	0.675	7.568	3.795e-14	[3.783, 6.428]

Covariance estimator: robust

Table 4.13

more complex, two basic ideas can be summarized as "Momentum Based Trading" and "Mean Reversion Based Trading". The former theory hypothesizes that stocks that do well now will likely continue to do so in the future while the latter states that especially good or past performance is an exception and that stocks will eventually return to their average performance. A third strategy, Pairs Trading, will be detailed later.

In 1985 [Bondt and Thaler] were one of the early scholars to analyze mean reversion behaviour when they examined the hypothesis that markets tend to overreact. They looked at monthly returns of assets listed on the New York Stock Exchange in between 1926 and 1982 and constructed portfolios of winners and losers that were updated every three years. Winners and losers were those stocks that had performed the best / worst over the previous years. Their idea was that "if stock prices systematically overshoot, then their reversal should be predictable from past return data alone, with no use of any accounting data such as earnings. [...] Extreme movements in stock prices will be followed by subsequent price movements in the opposite direction." Indeed they observed that the portfolio of losers outperformed the market significantly. While they focus on a very long time horizon of three years, Jagadeesh(1991) suggests that this behaviour could also be observed in the shorter term: "These papers show that contrarian strategies that select stocks based on their returns in the previous week or month generate significant abnormal returns." While mean reversion behaviour has sometimes been observed in practice, it is not trivial to derive it from economic theory. Many economic theories like the famous model from Fama and French describe the return of the stocks of a company as the result of market properties like the baseline market return rate and inherent properties of that company, like for example its book-to-market-ratio [Fama and French 1993]. If such a relationship

Results for GJR-GARCH(1,1) with t-Distribution (INTC)

Dep. Variable:	log_returns	**R-squared:**	-0.000
Mean Model:	Constant Mean	**Adj. R-squared:**	-0.000
Vol Model:	GJR-GARCH	**Log-Likelihood:**	-2474.66
Distribution:	Standardized Student's t	**AIC:**	4961.31
Method:	Maximum Likelihood	**BIC:**	4993.23
		No. Observations:	1509
Date:	Wed, Sep 04 2019	**Df Residuals:**	1503

	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
mu	0.0808	2.862e-02	2.822	4.777e-03	[2.466e-02, 0.137]
	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
omega	0.0152	4.217e-02	0.360	0.719	[-6.748e-02,9.783e-02]
alpha[1]	0.0146	9.131e-03	1.601	0.109	[-3.276e-03,3.252e-02]
gamma[1]	0.0133	5.218e-02	0.255	0.799	[-8.898e-02, 0.116]
beta[1]	0.9712	4.137e-02	23.476	7.225e-122	[0.890, 1.052]
	coef	**std err**	**t**	**P> |t|**	**95.0% Conf. Int.**
nu	4.1624	0.473	8.808	1.272e-18	[3.236, 5.089]

Covariance estimator: robust

Table 4.14

	MSE	Binary Accuracy	Naive Binary Accuracy
V - ARMA(0,0)	0.2457	54.1528 %	**54.2193** %
V - ARMA(1,1)	0.2473	**55.8139**	54.2193
INTC - ARMA(0,0)	0.2760	51.7608	**52.3588**

Table 4.15: Caption

exists than this in turn implies that the actual observed stock movements should be random fluctuations around some much more slowly changing true return rate. However, it is unclear whether this mean reversion should happen in the form of a pendulum that swings forth and back or more in the form of coin tosses reverting to their equilibrium by flooding past observations with new random ones. Something about multiplicative connections?
On the opposite side of the spectrum of trading strategies lies the idea of momentum. [Japateesh and CX] in 1993 were one of the first to describe "that strategies which buy stocks that have performed well in the past and sell stocks that have performed poorly in the past generate significant positive returns over 3- to 12-month holding periods." [Japateesh and CX] looked at portfolios and saw mean reversion over 12 to 48 months. Thaler and Bondt observe mean reversion over the time frame of 36 months. Many scholares agree that mean reversion and momentum behaviour are not necessarily contradictions, but that the time frame determines in which way a stock will behave [Balvers and Wu].
PAIRS TRADING

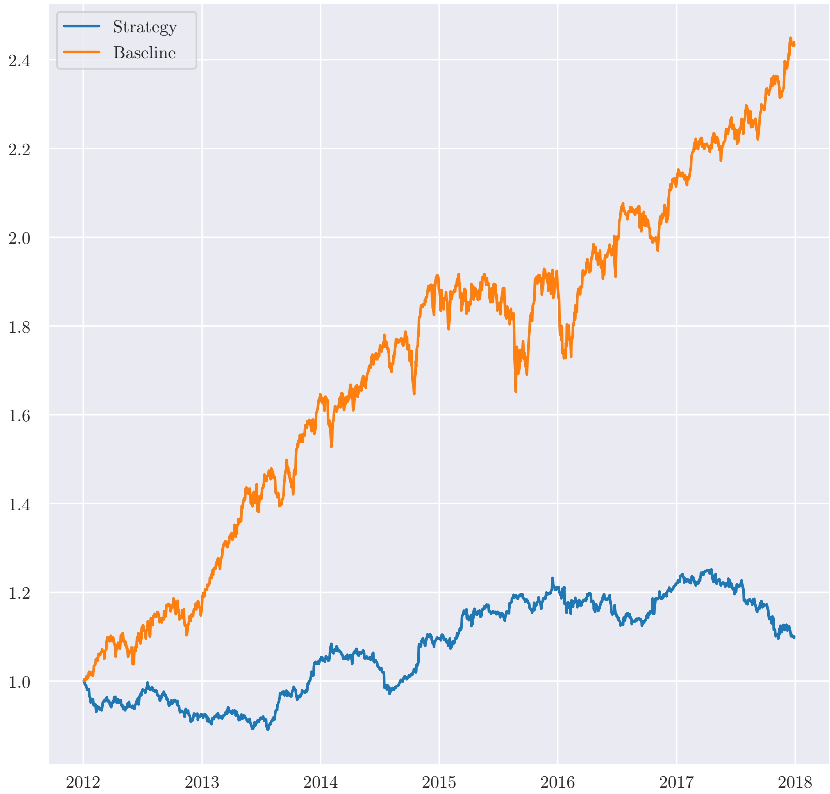

Figure 5.1: Caption

5.2 Trading Strategies - Implementation

5.2.1 Mean Reversion Portfolio

While their analysis is focused on monthly returns over a much longer time frame, the basic idea that was replicated in our trading strategy was the same. While they found excessive returns of that strategy from 1932 to 1977 our implementation was unforturnately much less successful.
Implementation algorithm: rank portfolio everyday along their returns from the last day. Then buy the two worst performers and sell the two best performers. Interestingly, reversing that does not produce better results. The time frame is obviously not right.

5.2.2 Mean Reversion - single stocks

Cumulative Mean - Stock prices

Idea: Plot: Cumulative Mean vs. Actual Time Series –> We see that the cumulative mean does not capture the time series well. Trend is always behind the current development, since we have a trend

Cumulative Mean - Returns

Comparison of 90d and 30d moving averages

Idea: If 30d averages is above 90d average, then sell. And vice versa

Mean Reversion Portfolio

Idea: Look at entire portfolio.

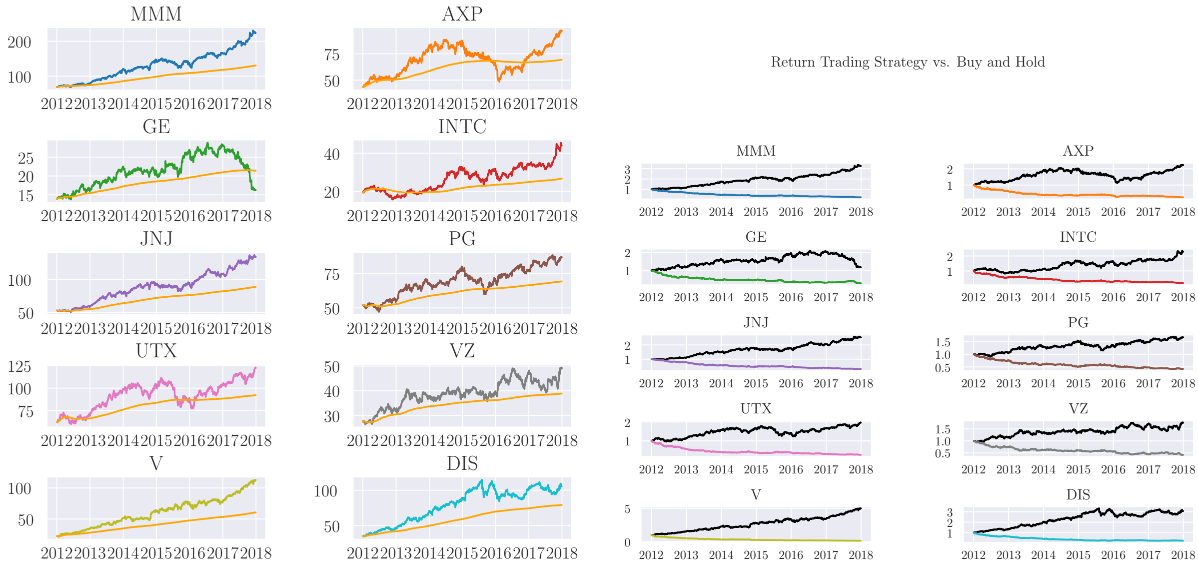

Figure 5.2: Caption

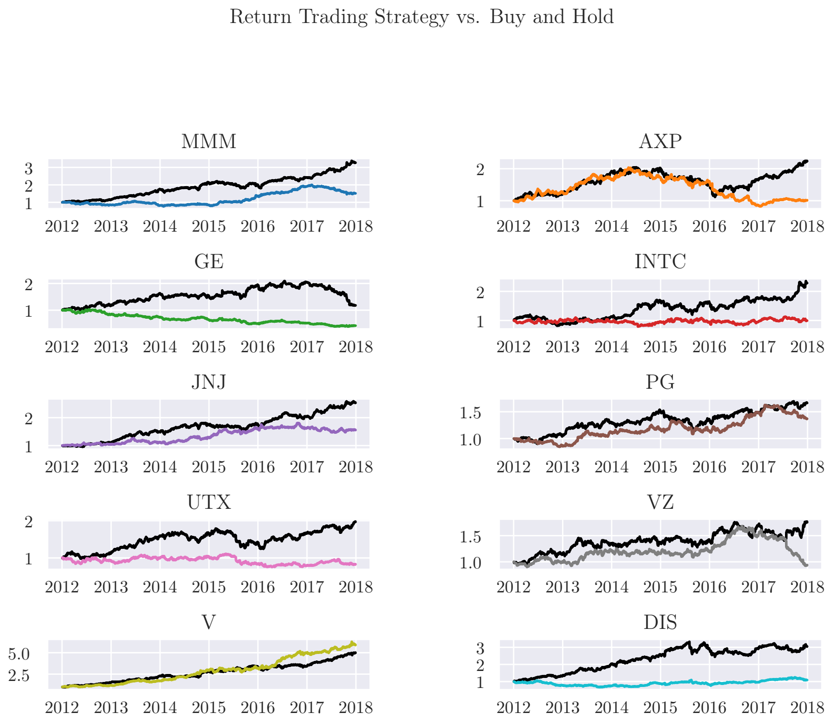

Figure 5.3: Mean Reversion based on past returns for single stocks

5.2.3 Momentum Based Trading

5.2.4 Pairs Trading

5.3 Trading Strategies Based on Our Predictions

5.4 Hybrid Trading Strategies

6. Conclusion

Bibliography

Ravenpack. https://www.ravenpack.com/. Accessed: 2019-08-30.

Adam Westerski. Sentiment analysis: Introduction and the state of the art overview. *Universidad Politecnica de Madrid, Spain*, pages 211–218, 2007.

Calum S Robertson, Shlomo Geva, and Rodney C Wolff. News aware volatility forecasting: Is the content of news important? In *Proceedings of the sixth Australasian conference on Data mining and analytics-Volume 70*, pages 161–170. Australian Computer Society, Inc., 2007.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

Julia Silge and David Robinson. tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, 1(3), 2016. doi: 10.21105/joss.00037. URL http://dx.doi.org/10.21105/joss.00037.

Tyler W. Rinker. *textstem: Tools for stemming and lemmatizing text*. Buffalo, New York, 2018. URL http://github.com/trinker/textstem. version 0.1.4.

A. Appendix A name

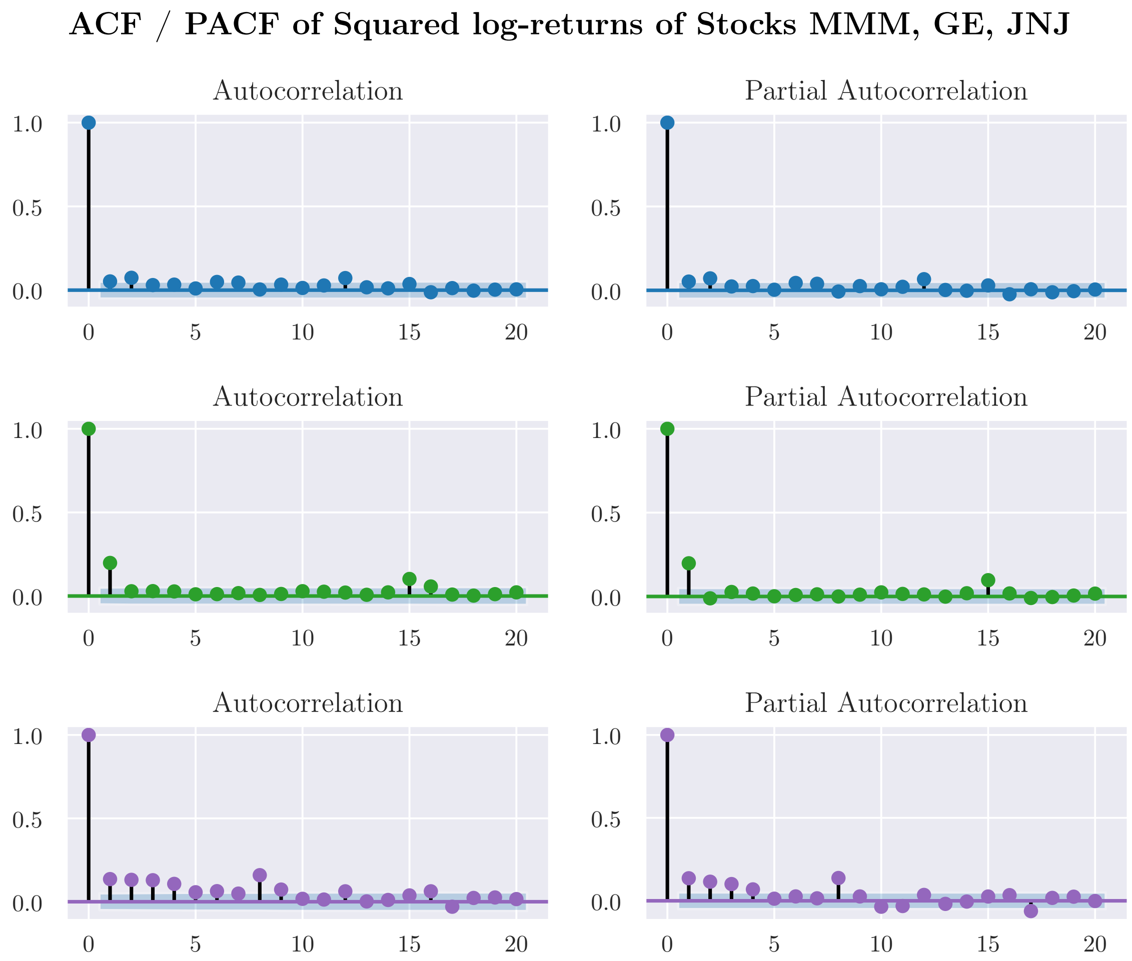

Figure A.1: ACF and PACF of squared residuals of stocks MMM, GE and JNJ. We see that some of the squared log-returns exhibit indeed autocorrelation, while others do less so. Strong autocorrelation implies that there is information about the future in the time series that can be modeled.

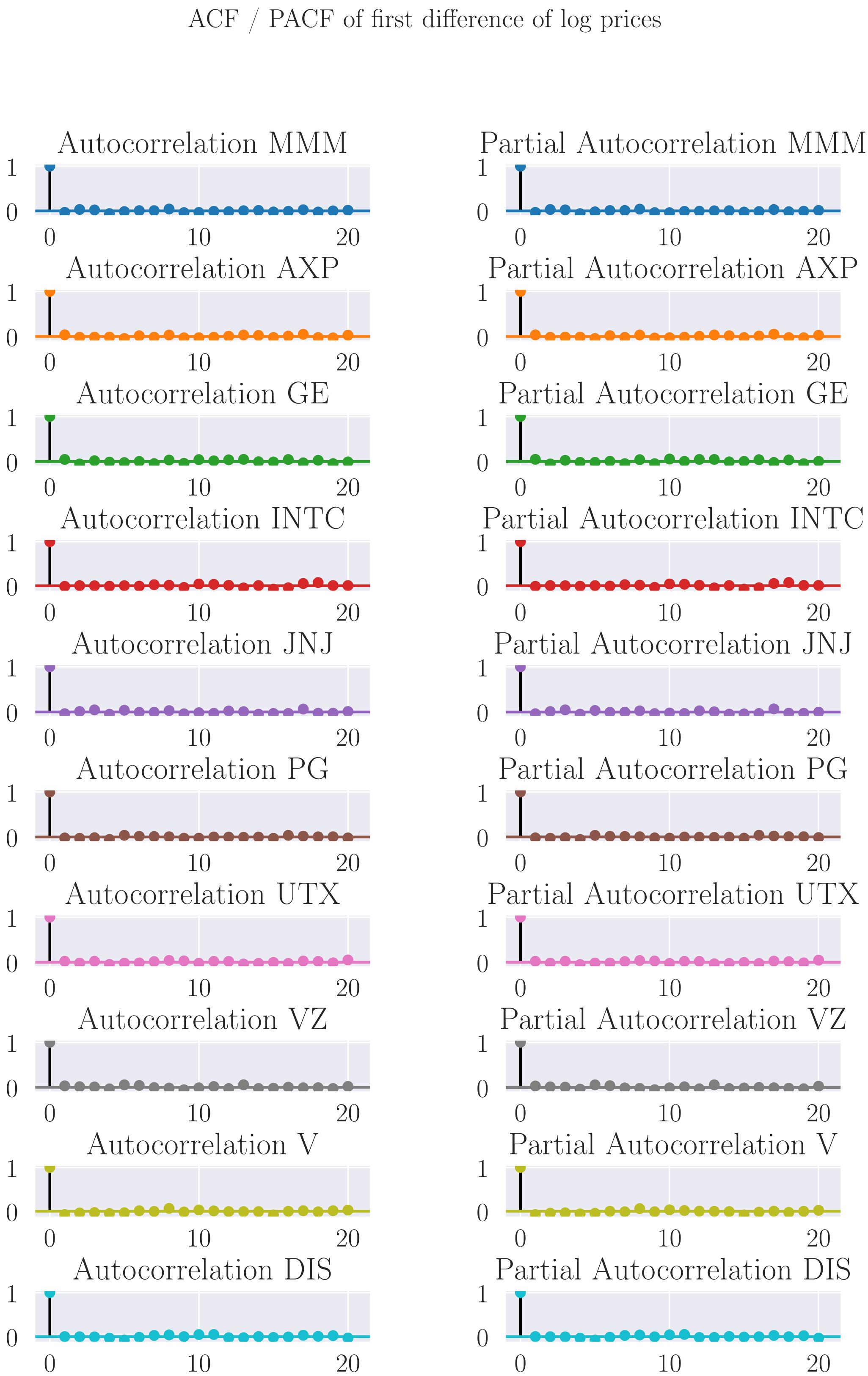

Figure A.4: Autocorrelation and partial autocorrelation for the first difference of log adjusted closing prices for all stocks

Results for an AR(1) process fit to the log-returns of V

Dep. Variable:	log_returns	**No. Observations:**	1509
Model:	ARMA(1, 0)	**Log Likelihood**	4446.683
Method:	css-mle	**S.D. of innovations**	0.013
Date:	Wed, 04 Sep 2019	**AIC**	-8887.366
Time:	13:41:26	**BIC**	-8871.409
Sample:	0	**HQIC**	-8881.423

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0011	0.000	3.503	0.000	0.000	0.002
ar.L1.log_returns	-0.0736	0.026	-2.868	0.004	-0.124	-0.023

	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	-13.5846	+0.0000j	13.5846	0.5000

Table A.1

Results for an MA(1) process fit to the log-returns of V

Dep. Variable:	log_returns	**No. Observations:**	1509
Model:	ARMA(0, 1)	**Log Likelihood**	4447.077
Method:	css-mle	**S.D. of innovations**	0.013
Date:	Wed, 04 Sep 2019	**AIC**	-8888.155
Time:	13:42:20	**BIC**	-8872.197
Sample:	0	**HQIC**	-8882.212

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0011	0.000	3.551	0.000	0.000	0.002
ma.L1.log_returns	-0.0809	0.027	-3.002	0.003	-0.134	-0.028

	Real	**Imaginary**	**Modulus**	**Frequency**
MA.1	12.3538	+0.0000j	12.3538	0.0000

Table A.2

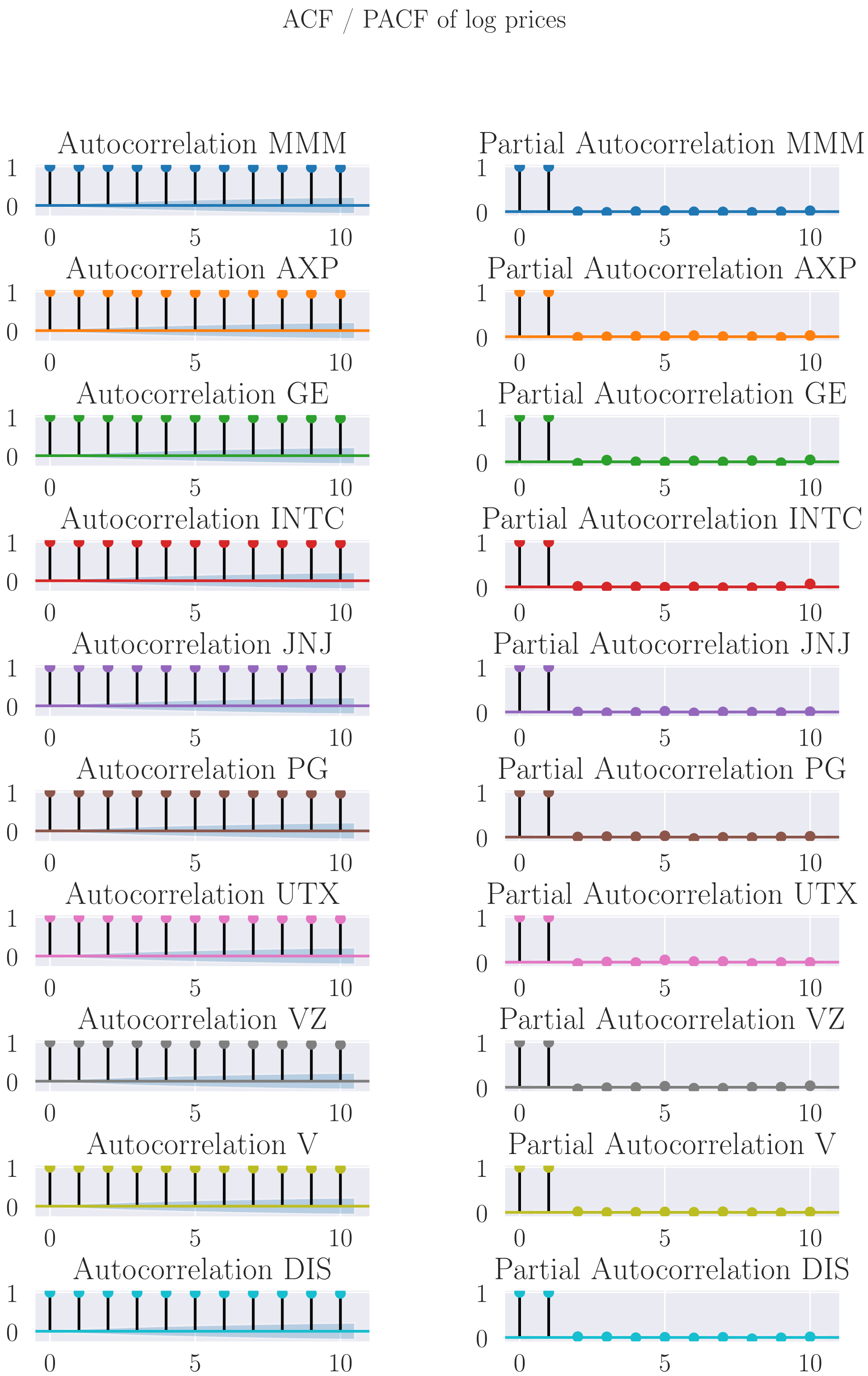

Figure A.2: Autocorrelation and partial autocorrelation for the log of the adjusted closing prices for all stocks

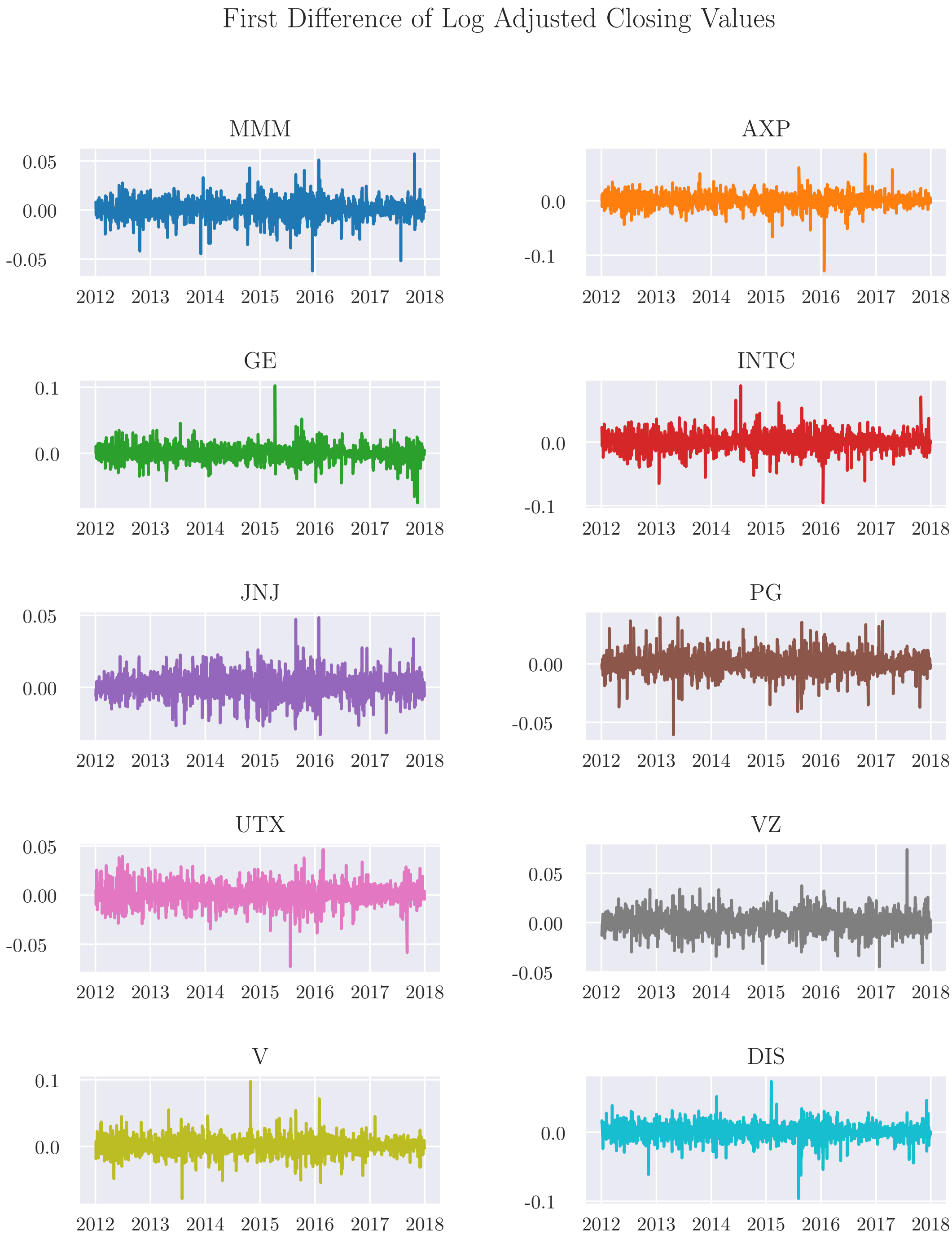

Figure A.3: First difference of log adjusted closing prices

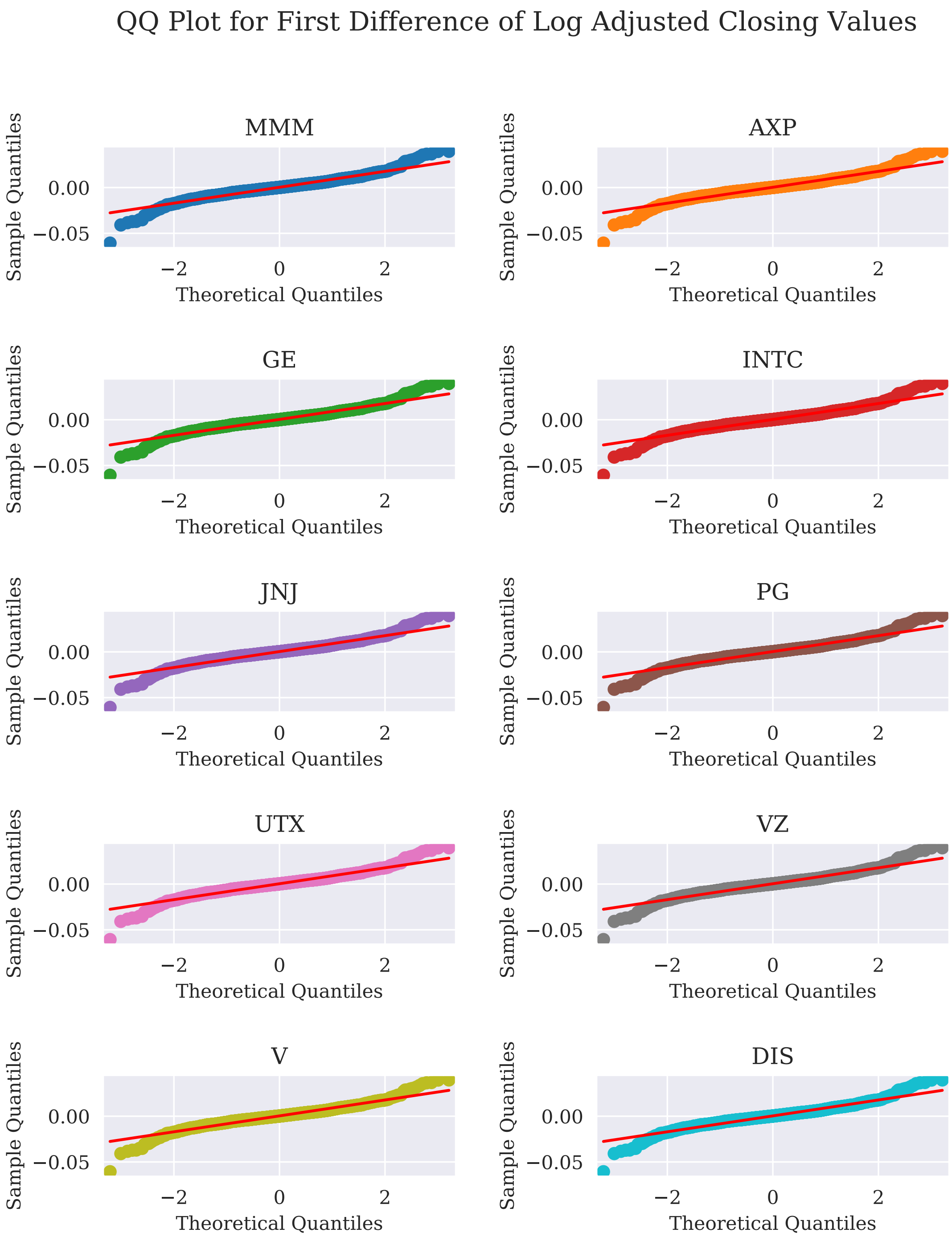

Figure A.5: QQ-Plots

Results for an ARMA(1,1) process fit to the log-returns of V

Dep. Variable:	log_returns	**No. Observations:**	1509
Model:	ARMA(1, 1)	**Log Likelihood**	4451.108
Method:	css-mle	**S.D. of innovations**	0.013
Date:	Tue, 03 Sep 2019	**AIC**	-8894.215
Time:	13:15:26	**BIC**	-8872.938
Sample:	0	**HQIC**	-8886.291

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0011	0.000	4.194	0.000	0.001	0.002
ar.L1.log_returns	0.6156	0.116	5.307	0.000	0.388	0.843
ma.L1.log_returns	-0.6997	0.105	-6.681	0.000	-0.905	-0.494

	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	1.6243	+0.0000j	1.6243	0.0000
MA.1	1.4293	+0.0000j	1.4293	0.0000

Table A.3

Results for an ARMA(2,1) process fit to the log-returns of V

Dep. Variable:	log_returns	**No. Observations:**	1509
Model:	ARMA(2, 1)	**Log Likelihood**	4451.124
Method:	css-mle	**S.D. of innovations**	0.013
Date:	Tue, 03 Sep 2019	**AIC**	-8892.248
Time:	13:15:28	**BIC**	-8865.652
Sample:	0	**HQIC**	-8882.343

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0011	0.000	4.189	0.000	0.001	0.002
ar.L1.log_returns	0.6033	0.135	4.467	0.000	0.339	0.868
ar.L2.log_returns	-0.0059	0.032	-0.182	0.855	-0.069	0.057
ma.L1.log_returns	-0.6850	0.133	-5.164	0.000	-0.945	-0.425

	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	1.6851	+0.0000j	1.6851	0.0000
AR.2	101.4345	+0.0000j	101.4345	0.0000
MA.1	1.4597	+0.0000j	1.4597	0.0000

Table A.4

Results for an AR(1) process fit to the log-returns of INTC

Dep. Variable:	log_returns	**No. Observations:**	1509
Model:	ARMA(1, 0)	**Log Likelihood**	4353.952
Method:	css-mle	**S.D. of innovations**	0.014
Date:	Wed, 04 Sep 2019	**AIC**	-8701.903
Time:	13:41:26	**BIC**	-8685.946
Sample:	0	**HQIC**	-8695.960

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0005	0.000	1.596	0.111	-0.000	0.001
ar.L1.log_returns	-0.0138	0.026	-0.534	0.593	-0.064	0.037

	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	-72.7156	+0.0000j	72.7156	0.5000

Table A.5

Results for an MA(1) process fit to the log-returns of INTC

Dep. Variable:	log_returns	**No. Observations:**	1509
Model:	ARMA(0, 1)	**Log Likelihood**	4353.951
Method:	css-mle	**S.D. of innovations**	0.014
Date:	Wed, 04 Sep 2019	**AIC**	-8701.902
Time:	13:42:20	**BIC**	-8685.945
Sample:	0	**HQIC**	-8695.959

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0005	0.000	1.596	0.111	-0.000	0.001
ma.L1.log_returns	-0.0137	0.026	-0.534	0.594	-0.064	0.037

	Real	**Imaginary**	**Modulus**	**Frequency**
MA.1	72.9645	+0.0000j	72.9645	0.0000

Table A.6

Results for an ARMA(1,1) process fit to the log-returns of INTC

Dep. Variable:	log_returns	**No. Observations:**	1509
Model:	ARMA(1, 1)	**Log Likelihood**	4353.953
Method:	css-mle	**S.D. of innovations**	0.014
Date:	Tue, 03 Sep 2019	**AIC**	-8699.906
Time:	13:15:27	**BIC**	-8678.629
Sample:	0	**HQIC**	-8691.982

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0005	0.000	1.594	0.111	-0.000	0.001
ar.L1.log_returns	-0.1300	1.384	-0.094	0.925	-2.842	2.582
ma.L1.log_returns	0.1163	1.387	0.084	0.933	-2.603	2.835

	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	-7.6907	+0.0000j	7.6907	0.5000
MA.1	-8.5960	+0.0000j	8.5960	0.5000

Table A.7

Results for an ARMA(2,1) process fit to the log-returns of INTC

Dep. Variable:	log_returns	**No. Observations:**	1509
Model:	ARMA(2, 1)	**Log Likelihood**	4353.952
Method:	css-mle	**S.D. of innovations**	0.014
Date:	Tue, 03 Sep 2019	**AIC**	-8697.904
Time:	13:15:28	**BIC**	-8671.308
Sample:	0	**HQIC**	-8687.999

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0005	0.000	1.595	0.111	-0.000	0.001
ar.L1.log_returns	0.3549	nan	nan	nan	nan	nan
ar.L2.log_returns	0.0055	nan	nan	nan	nan	nan
ma.L1.log_returns	-0.3686	nan	nan	nan	nan	nan

	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	2.7036	+0.0000j	2.7036	0.0000
AR.2	-66.7356	+0.0000j	66.7356	0.5000
MA.1	2.7127	+0.0000j	2.7127	0.0000

Table A.8

Statutory Declaration

We declare that we have authored this thesis independently, that we have not used other than the declared sources / resources, and that we have explicitly marked all material which has been quoted either literally or by content from the used sources.

Georg-August-University Göttingen, September 9, 2019

Nikos Bosse
<nikos.bosse@stud.uni-goettingen.de>

Felix Süttmann
<felix.suettmann@stud.uni-goettingen.de>