

Jumps in high frequency data

Masters Thesis submitted to

Prof. Dr. Ostap Okhrin

Prof. Dr. Brenda Lopez Cabrera

Ladislaus von Bortkiewicz Chair of Statistics

C.A.S.E.- Centre for Applied Statistics and Economics

Humboldt-Universität zu Berlin



by

Martin Schelisch

(527912)

in partial fulfillment of the requirements

for the degree of

Master of Science in Statistics

Berlin, 10 June 2011

Acknowledgment

I would heartily like to express my gratitude to Prof. Dr. Ostap Okhrin, whose help, advice and supervision from the preliminary to the concluding level of this study enabled me to develop a deep understanding of the subject. I would also like to offer my thanks to Prof. Dr. Brenda Lopez Cabrera and Prof. Dr. Wolfgang Härdle for their support. In addition, I would like to thank the *Collaborative Research Center 649: Economic Risk* and the *Research Data Center* for providing the data sets and especially Tomas Polak, whose help in understanding TAQ data sets was vital to me.

I would like to extend my thanks to all of those who supported me in any respect during the completion of the project, in particular Leslie Udvarhelyi and the whole staff of the Ladislaus von Bortkiewicz Chair of Statistics.

Last but not the least, I would like to express my deep thanks to Julia Ritter as well as my grandfather, Erhardt Schelisch, for listening and supporting me emotional throughout the biggest part of my life. Without their understanding and encouragement it would have not been possible for me to accomplish this work.

Martin Schelisch

Abstract

Due to high frequency data researchers can observe jumps in the price process which raises the question if these price changes can really be the result of only a pure diffusion process. In this work two different algorithms for identifying gradual and mathematical jumps are described. These algorithm can not only detect whether or not a jump has occurred on a specific day, but they can also determine the number of jumps on that day, the sign of the jump and produce an estimate of the jump size. The first algorithm, *ALGO1*, is based on the classic theory developed by Barndorff-Nielsen and Shephard (2004), using realised volatility and the bipower variation. The second algorithm, *ALGO2*, is based on the recent work of Kloessner (2010), who developed a new theory based on intradaily lows and highs.

Both algorithms were applied to trades and quotes data (TAQ) from the New York Stock Exchange (NYSE) for the period January 2008 till July 2009, and therefore covering the financial crisis in autumn 2008. *ALGO1* mainly reports days with one or two jumps, whereas *ALGO2* finds significantly more jumps (frequently four to six) per day. Jumps occur mainly at the beginning of the trading day, uncovering the typical L-shape for US data. Both algorithms detect that about 30 – 45% of the logreturn variation is attributed to jumps on average. In order to verify these results a simulation study was realised. Using the model presented by Heston (1993), 4,000 days were simulated with different amounts of mathematical and gradual jumps. A confusion matrix revealed that the test statistic of *ALGO2* is unable to hold the suspected confidence level in this simulation setup. It was also shown that, using this setup, z_{QPLM} is unable to detect gradual jumps whereas the test statistic used in *ALGO2* is truly able to detect this kind of jump.

Nevertheless, the results highlight the importance and the impact of jumps on daily estimates of volatility.

Keywords: jumps, volatility, high frequency data, market microstructure noise, assets

Zusammenfassung

Man nimmt an, dass der Preis von Finanzanlagen einem stetigen Diffusionsprozess folgt. Durch die gestiegene Verfügbarkeit von hochfrequenten Datensätzen, welche Preisbewegungen im Sekundenbereich beinhalten, werden jedoch immer wieder Preissprünge entdeckt, die durch einen solchen stetigen Prozess jedoch nicht zu erklären sind. In dieser Arbeit werden zwei Algorithmen dargestellt, welche Tage mit Sprüngen, die Anzahl der Sprünge sowie deren Richtung identifizieren und eine Schätzung der Sprunghöhe erstellen. Der erste Algorithmus, *ALGO1*, basiert auf der von Barndorff-Nielsen and Shephard (2004) entwickelten Theorie unter Verwendung von hochfrequenten Datensätzen, wohingegen der zweite Algorithmus, *ALGO2*, erst kürzlich veröffentlichte Ergebnisse von Kloessner (2010) nutzt und bereits mit so genannten OHLC-Datensätzen auskommt.

Beide Algorithmen wurden auf einen TAQ-Datensatz des NYSE angewandt, welcher den Preisverlauf von drei Aktien im Zeitraum Januar 2008 bis Juli 2009 abdeckt, und damit die Auswirkungen der Finanzkrise im Herbst 2008 verdeutlicht. Es wurde deutlich, dass *ALGO2* im Gegensatz zu *ALGO1* graduelle Sprünge entdecken kann, Sprünge hauptsächlich am Anfang eines Handelstages entstehen und ca. 30 – 45% der Variabilität des logarithmischen Aktienpreises ausmachen.

Diese Ergebnisse wurde mittels einer Simulationsstudie verifiziert, welche auf dem bekannten Heston (1993) Model basiert. Diese zeigt deutlich, dass *ALGO1* graduelle Sprünge nicht entdecken und die Teststatistik von *ALGO2* das Konfidenzniveau nicht halten kann.

Dennoch betonen die Resultate den starken Einfluss von Sprüngen auf die Preisentwicklung.

Schlagwörter: Sprünge, Volatilität, hochfrequente Datensätze, market microstructure noise, Anlagen

Contents

1	Introduction	1
2	Modelling intraday asset prices	5
2.1	The classic approach	6
2.2	Using intraday highs and lows	8
3	Identifying jumps	11
3.1	Estimating integrated quarticity	12
3.2	Test statistics to detect days with a jump	12
3.3	Determine jump size and intensity	15
4	Applied data cleaning and management procedure	19
5	Empirical Analysis	25
5.1	Preliminary analysis	26
5.2	Jump intensity	30
5.3	Jump size	33
6	Simulation study	37
6.1	Data-generating price process with jumps	37
6.2	Simulation details	38
6.3	Results	39
7	Conclusion	43

List of Figures

2.1	Japanese Candlestickplot for the asset of Citigroup on 2 January 2008.	9
3.1	Flowchart of <i>ALGO1</i> and <i>ALGO2</i>	17
5.1	Volatility Signature Plot.	25
5.2	Logprices and log returns.	27
5.3	Volatility estimates.	28
5.4	Values of the test statistics z_{QPLM} and $\max(TJ_p, TJ_n)$	29
5.5	Price series for days with a high amount of detected jumps	32
5.6	Recorded jump times reported by <i>ALGO2</i>	33
5.7	Jumpfree Volatility (<i>ALGO1</i>)	35
5.8	Jumpfree Volatility (<i>ALGO2</i>)	36

List of Tables

4.1	Excerpt of the data set	19
4.2	TAQ Quote data description	20
4.3	Applied cleaning and data management steps.	21
4.4	Excerpt of the generated homogeneous time series.	22
4.5	Number of observations after each cleaning step.	23
4.6	Number of observations after each cleaning step (continued).	24
5.1	Number of days with jumps	30
5.2	Jump Intensity	31
5.3	Robustness of jump intensity	31
5.4	Descriptive statistics for positive and negative jump sizes	34
6.1	Experimental design for the simulation study.	39
6.2	Confusion matrix for simulated days	40
6.3	Detection rate for simulated days with more than one jump	41

1 Introduction

Continuous time diffusion processes play a vital role in modern financial modelling as they open up an elegant way to solve problems such as the hedging and pricing of derivative products. Advances in computer technology, data recording and storage have made data sets increasingly accessible to researchers, these contain observations on financial variables at very fine time intervals, i.e. time stamped transaction-by-transaction or tick-by-tick. These data sets are referred to in the common literature as (ultra) high frequency data. Due to this improvement researchers can observe price changes at very slight time scales, these sometimes seem to be too sharp over too small time intervals, and raise the question if these price changes can really be the result of only a pure diffusion process. Consequently, in the last few years, the modelling, measuring and detecting discontinuities in asset returns, so-called jumps, has been quite a prominent topic in econometrics, as testified by the following (not exhaustive) list of studies: Andersen et al. (2007a), Andersen et al. (2010), Ait-Sahalia and Jacod (2009), Barndorff-Nielsen and Shephard (2004), Barndorff-Nielsen and Shephard (2006), Huang and Tauchen (2005), Kloessner (2009), Lee and Ploberger (2009) among many others. Often, the slightly philosophical question of "*what makes a jump a jump?*" was not addressed. However it is a vital question as it determines and specifies the goal more precisely. Despite the typical understanding of a jump, described as a (nearly instantaneously) rapid price movement from one level to another and in the following referred to as *mathematical jumps*, Barndorff-Nielsen et al. (2008b) came up with another kind of price movements: the *gradual jump*. This kind of price movement is characterised by a price adjustment from one level to another within several minutes, taking a lot of intermediate values before reaching its new equilibrium. Consequently, both types of jumps were investigated as both are vital for forecasting future price movements.

In this work two different algorithms for identifying gradual and mathematical jumps are described. These algorithm can not only detect whether or not a jump has occurred on a specific day, but they can also determine the number of jumps on that day, the sign of the jump and produce an estimate of the jump size.

The first algorithm was recently developed by Andersen et al. (2007b) and Ane and Metais (2010) and contains, in this upgraded version, further improvements to ensure a more precise measurement of the jump size as well as a procedure which is able to detect split up gradual jumps. It is based on the well known fact that the quadratic variation

can be divided into a continuous component and a jump component. Barndorff-Nielsen and Shephard (2004), Barndorff-Nielsen and Shephard (2006) developed strong econometric devices, building on a powerful asymptotic theory, to detect the presence of jumps in a non-parametric setting. The key element is the difference between two estimators for the volatility of the price process: the realised volatility, first introduced in Andersen et al. (2001), which includes the contribution of jumps, if any, as it approximates the quadratic variation for higher sampling frequencies, and the bipower variation, developed by Barndorff-Nielsen and Shephard (2004), which only approximates the continuous component of the quadratic variation. It is therefore, at least theoretically, possible to identify jumps by a significant difference between these two estimators.

The second algorithm is based on the recent work of Kloessner (2010), who developed a new theory and estimators for different quantities like the quadratic variation or the sum of squared jumps based on intradaily lows and highs. According to Kloessner (2010), his approach is the only one available up until now which can detect gradual jumps, due to the use of more data than common methods.

Nevertheless, the identification of (small) jump sizes is hindered by market microstructure noise at high sampling frequencies, which often dominates classical estimates for realised variance. It is therefore important to account for this noise in the estimation procedure, as discussed by Zhang et al. (2005), Hasbrouck (2004), Hansen and Lunde (2006), Barndorff-Nielsen et al. (2004), Bandi and Russel (2006) among many others. As the choice of an adequate sampling frequency is also vital for the detection of gradual jumps, sampling frequencies higher than five minutes were avoided in this work. This seems to be a good trade-off between loss of information and circumventing both problems mentioned.

Despite this hurdle, both algorithms were applied to trades and quotes data (TAQ) from the New York Stock Exchange (NYSE) for the period January 2008 till July 2009, and therefore covering the financial crisis in autumn 2008. To validate the results and to obtain evidence for the quality of the algorithm, a simulation study using the well known Heston model was realised.

The outline of this work is as follows: In Chapter 2 the underlying continuous-time jump diffusion process is presented as well as various estimators and definitions necessary to catch, loosely speaking, price variations, commonly referred to as volatility. The following Chapter contains a collection of recently published test statistics to identify trading days with (detectable) jumps in the price process (Section 3.2) as well as a detailed description of both algorithms (Section 3.3). As all estimators rely on (subsets of) high frequency data sets, important issues concerning the data management and cleaning are discussed in Chapter 4. This chapter also contains information regarding

the applied cleaning procedure for the empirical part of this work, presented in Chapter 5, where the presented algorithms were applied to three different assets, traded on the New York Stock Exchange from January 2008 to July 2009.

2 Modelling intraday asset prices

'Arguably, no concept in financial mathematics is as loosely interpreted and as widely discussed as 'volatility'. [...] 'volatility' has many definitions, and is used to denote various measures of changeability.' Shiryaev (1999, p. 345)

Although this thirteen-year-old citation from Shiryaev still hits the nail on the head, volatility estimation was, and still is, a very prominent topic in modern financial econometrics. This is due to the fact that volatility is a key element in many financial applications, e.g. asset and derivatives pricing, risk management or portfolio selection. As recent empirical evidence suggests that jumps may have a non-trivial contribution to the overall daily price variation, see e.g. Andersen et al. (2007a), assume that the logarithmic price of a financial asset is given by the following continuous-time jump diffusion process:

$$dp_t = \mu_t dt + \sigma_t dW_t + k_t dq_t, \quad (2.1)$$

where μ_t is a continuous mean process with finite variance, the volatility process σ_t is assumed to be a non negative càdlàg process to allow for occasional jumps and W_t is a standard Wiener process. The jump process $k_t dq_t$ can be decomposed into its counting process q_t and the jump size process $k_t = p_t - p_{t-}$, with $p_{t-} = \lim_{s \uparrow t} p_s$.

Suppose m is the number of intraday observations; the i th *intraday return* is defined as the difference between two logarithmic prices

$$r_i^{(m)} = p_{\frac{i}{m}} - p_{\frac{i-1}{m}}, \quad i = 1, 2, \dots, m. \quad (2.2)$$

Importantly, $p_{\frac{i}{m}}$ and $p_{\frac{i-1}{m}}$ are not necessarily two subsequently observed logarithmic prices.

The amount of variation, accumulated over a past time interval (i.e. one day), called *quadratic variation (QV)*, *ex-post variation* or *total variation*,

$$QV = \lim_{m \rightarrow \infty} \sum_{i=1}^m r_i^{(m)}, \quad i = 1, 2, \dots, m, \quad (2.3)$$

is, in many econometric questions, of interest. *QV* overlaps often with a term called

integrated variance or integrated volatility (IV)

$$IV = \int_{t-1}^t \sigma^2(s) ds. \quad (2.4)$$

In the presence of jumps, i.e. $k_t \neq 0$, QV can be subdivided into

$$QV = IV + SSJ, \quad (2.5)$$

where the last quantity is the sum of squared jumps

$$SSJ = \sum_{0 \leq s \leq 1} (k_s)^2. \quad (2.6)$$

As IV contains information about the contribution of the continuous part of the log price process to volatility, it is often referred to *diffusive volatility*. On the other hand SSJ is often called *jump risk* or *volatility due to jumps* as it, analogous to IV , contains information about the discontinuous part of the log price process.

In the following the estimation of QV , IV and SSJ , over one period, i.e. one trading day, is described and for the ease of exposition the time subscript for different periods is dropped.

2.1 The classic approach

A natural estimator of QV , which became well known as *realised volatility* (RV), is defined by the sum of squared intraday returns:

$$RV^{(m)} = \sum_{i=1}^m r_i^{(m)2}. \quad (2.7)$$

Andersen et al. (2003) have shown that the realised volatility $RV^{(m)}$ converges uniformly in probability to the integrated variation as the sampling frequency of returns approaches infinity, i.e.

$$RV^{(m)} \xrightarrow{P} IV \text{ as } m \rightarrow \infty,$$

thus providing a consistent estimate of the integrated variance assuming that the underlying price follows equation (2.1) without jumps ($q(t) = q(t-1) \forall t$). Nevertheless, the question of precision arises. The asymptotic distribution of $RV^{(m)}$ was derived by Barndorff-Nielsen and Shephard (2002):

$$\frac{\sqrt{m}(RV^{(m)} - IV)}{\sqrt{2IQ}} \xrightarrow{\mathcal{L}} \mathbb{N}(0, 1), \quad (2.8)$$

with

$$IQ = \int_0^1 \sigma^4(s) ds \quad (2.9)$$

denoting the *integrated quarticity* (IQ). Unfortunately, the computation of the asymptotic distribution is infeasible, given that IQ is unknown (estimators for IQ will be discussed in Section 3). Furthermore, in the presence of jumps, the realised volatility $RV^{(m)}$ converges uniformly in probability to the *total* price variation, QV , as the sampling frequency of returns approaches infinity:

$$RV^{(m)} \xrightarrow{p} IV + SSJ \text{ as } m \rightarrow \infty. \quad (2.10)$$

This crucial fact can be faced by the breakthrough of Barndorff-Nielsen and Shephard (2004), which allows for a separation of both parts. To illustrate the main idea, define the *realised bipower variation* as

$$BPV^{(m)} = \mu_1^{-1} \frac{m}{m-1} \sum_{j=2}^m |r_j| |r_{j-1}|, \quad (2.11)$$

where μ_1 is a constant given by

$$\mu_k = \frac{2^{k/2}}{\sqrt{\pi}} \Gamma\left(\frac{k+1}{2}\right), \quad (2.12)$$

with Γ denoting the Gamma function. It was shown by Barndorff-Nielsen and Shephard (2004) that, even in the presence of jumps, $BPV^{(m)}$ converges to IV as the sampling frequency of returns approaches infinity, i.e.

$$BPV^{(m)} \xrightarrow{p} IV \text{ as } m \rightarrow \infty, \quad (2.13)$$

holds for the underlying continuous-time jump diffusion price process defined in equation (2.1). This result follows from the fact that only a finite number of terms in the sum of $BPV^{(m)}$ are affected by jumps while the remaining terms go to zero in probability. As the probability of jumps goes to zero as the sampling frequency of returns approaches infinity, those terms do not impact the limiting probability. Thus, only the effects of the continuous-time process were captured by the asymptotic convergence of $BPV^{(m)}$, even in the presence of jumps. One should note that the result is obtained without any additional assumptions regarding the counting process, the jump size distribution and the relationship between the jump process and the volatility component.

As noted by Barndorff-Nielsen and Shephard (2004), combining the two results in Equation (2.10) and (2.13) will yield that the contribution to the total price variation due to the discontinuous jump part in the underlying price process may be consistently

estimated as the sampling frequency of returns approaches infinity by

$$RV^{(m)} - BPV^{(m)} \xrightarrow{P} SSJ \text{ as } m \rightarrow \infty. \quad (2.14)$$

This asymptotical result is at the heart of nearly all published literature about jumps in high frequency data in the last years. Nevertheless, it is not the only way to determine *SSJ*.

2.2 Using intraday highs and lows

So far, all estimators discussed rely on the conventional subperiod's return, which suffers from an inherent subjectivity as the interval ends are chosen in a rather arbitrarily way. The estimators presented in this Chapter go one natural step further as they also use the subperiod's highest and lowest (log)return and therefore use more information from the data, which should lead to a significant improvement in estimation accuracy.

Below estimators for the previously discussed quantities *IV*, *IQ* and *SSJ* based on intraday highs and lows are shown as first presented in Kloessner (2010).

Up to now, every subinterval $\left[\frac{i-1}{m}, \frac{i}{m}\right]$ was only characterised by its opening and close price, $p_{\frac{i-1}{m}}$ and $p_{\frac{i}{m}}$, respectively. Following the ideas of Kloessner (2009), one additionally determines the highest and lowest price in every given subinterval

$$(p^*)_{i,m} = \sup_{\frac{i-1}{m} \leq t \leq \frac{i}{m}} p_t \text{ and } (p_*)_{i,m} = \inf_{\frac{i-1}{m} \leq t \leq \frac{i}{m}} p_t. \quad (2.15)$$

A candlestick chart visualises not only these four values in an appropriate way (see Figure 2.1), it is also a good starting point for illustrating the idea behind the estimators. Define the candlestick's body length $b_{i,m}$, the upper wick's length $uw_{i,m}$ and lower wick's length $lw_{i,m}$ as

$$b_{i,m} = |p_{\frac{i}{m}} - p_{\frac{i-1}{m}}| \quad (2.16)$$

$$uw_{i,m} = (p^*)_{i,m} - \max(p_{\frac{i}{m}}, p_{\frac{i-1}{m}}) \quad (2.17)$$

$$lw_{i,m} = \min(p_{\frac{i}{m}}, p_{\frac{i-1}{m}}) - (p_*)_{i,m}. \quad (2.18)$$

Without loss of generality the subscripts for the corresponding interval and sampling frequency were dropped in the following theory. Estimates for one trading day are given by the sum over all subperiod estimates.

Kloessner (2010) has shown that candlesticks with a huge body but tiny wicks are quite unlikely in the absence of jumps as the sizes of the candlestick's body and wicks are determined by different functions of the terminal, maximal and minimal value of a Wiener process. On the other hand, large wicks (compared to the candlestick's body)

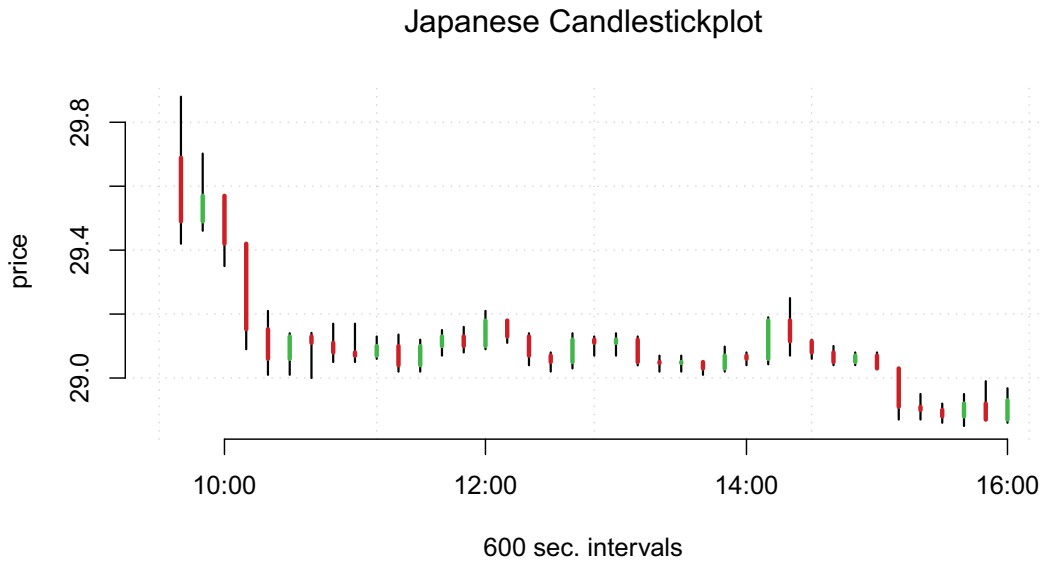


Figure 2.1: Japanese Candlestickplot for the asset of Citigroup on 2 January 2008. The thick red (green) lines, also called bodies, represent a loss (growth) in the stock price during that time interval whereas the upper (lower) wicks length represents the highest (lowest) price in that time interval. [JHFDJapCandlestick](#)

are an indication of diffusive behaviour in the absence of jumps. It is therefore possible to construct estimators based on these values.

The author presents two consistent estimators for IV , which have the smallest variance under all linear combinations of the estimators given in Equations (2.16), (2.15) and (2.16) in the absence of jumps, namely $0.7244 \cdot m^{-2} \cdot \sigma_{\frac{i-1}{m}}^4$ and $0.2921 \cdot m^{-2} \cdot \sigma_{\frac{i-1}{m}}^4$, respectively. The estimators are defined as

$$IV_l^{(m)} = 1.3277uw^2 + 1.3227lw^2 + 2.4847uw \cdot lw, \quad (2.19)$$

$$IV_p^{(m)} = 0.4416(uw^2 + lw^2) + 1.3851lw \cdot uw + 1.1809(uw \cdot b + lw \cdot b). \quad (2.20)$$

The first estimator in Equation (2.19), $IV_l^{(m)}$, only uses the wicks' length in order to ensure robustness in the presence of jumps. The second estimator in Equation (2.20), $IV_p^{(m)}$, incorporates products of body and the wicks' length to reduce the variance of the estimator in the absence of jumps.

Naturally, the squared body length, b^2 , should be used to estimate jumps in a given interval, but one also has to account for the fact that the squared body length will approximately estimate IV in the absence of jumps. Kloessner (2010) presented the idea to decrease b^2 in a manner that the resulting estimator will have a vanishing mean

and a small variance in the absence of jumps, while primarily measuring the squared jump height if a jump occurred in the interval. The resulting estimators for SSJ are

$$SSJ_t^{(m)} = b^2 - 1.4383uw^2 - 1.4383lw^2 - 2.0605uw \cdot lw, \quad (2.21)$$

$$SSJ_p^{(m)} = 0.6576(uw^2 + lw^2) + 0.5552lw \cdot uw - 2.8089b \cdot (uw + lw) + b^2. \quad (2.22)$$

Following the argumentation above, the estimators for the sum of squared positive jumps, $SSpJ$, are given by

$$SSpJ_t^{(m)} = \begin{cases} b^2 - 3.2047uw^2 - 3.2047lw^2 - 3.0301uw \cdot lw, & r > 0, \\ 1.7633uw^2 + 1.7663lw^2 + 0.9697uw \cdot lw, & r < 0, \end{cases} \quad (2.23)$$

$$SSpJ_p^{(m)} = \begin{cases} b^2 + 0.7706(uw^2 + lw^2) + 0.7394lw \cdot uw \\ \quad - 3.1847(uw \cdot b + lw \cdot b), & r > 0, \\ -0.1130(uw^2 + lw^2) - 0.1842lw \cdot uw \\ \quad + 0.3758(uw \cdot b + lw \cdot b), & r < 0, \end{cases} \quad (2.24)$$

whereas the estimators for the sum of squared negative jumps, $SSnJ$, are defined as

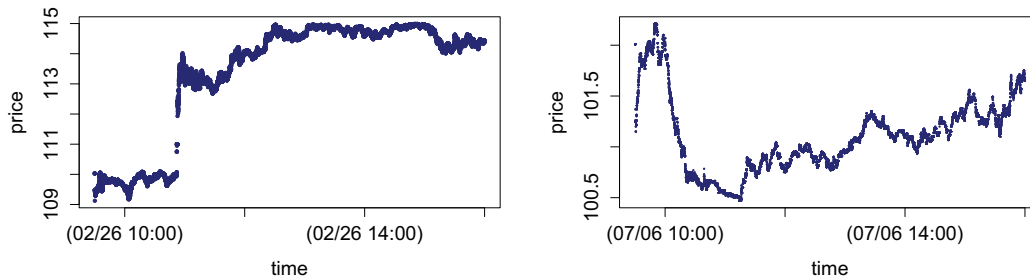
$$SSnJ_t^{(m)} = \begin{cases} 1.7633uw^2 + 1.7663lw^2 + 0.9697uw \cdot lw, & r > 0, \\ b^2 - 3.2047uw^2 - 3.2047lw^2 - 3.0301uw \cdot lw, & r < 0, \end{cases} \quad (2.25)$$

$$SSnJ_p^{(m)} = \begin{cases} -0.1130(uw^2 + lw^2) - 0.1842lw \cdot uw \\ \quad + 0.3758(uw \cdot b + lw \cdot b), & r > 0, \\ b^2 + 0.7706(uw^2 + lw^2) + 0.7394lw \cdot uw \\ \quad - 3.1847(uw \cdot b + lw \cdot b), & r < 0. \end{cases} \quad (2.26)$$

Thus, the theory developed by Kloessner (2010) provides estimators for all three important quantities, IV , QV , and SSJ , and additionally allows for a separation of SSJ into its positive and negative part.

3 Identifying jumps

Now that the basic ideas have been presented, Chapter 3 deals with the issue of how to identify jumps, their respective jump sizes and jump intensity.



(a) **Mathematical jump:** Nearly instantaneous rapid price movement from one level to another.
 (b) **Gradual jump:** Price adjustment from one level to another within several minutes, taking intermediate values before reaching their new equilibrium.

Jumps are generally understood as discontinuities in the price path, representing the reaction of market participants to some important information becoming known. Assume at time t important economic news becomes known which causes the logprice to jump by an amount of k_t . If the observed logprices p_t adjust to the new level instantaneous, this is called a *mathematical jump*:

$$p_t = p_t^* + k_t, \tag{3.1}$$

where p_t^* denotes the logprice at time t without the incorporation of the news.

However, if the reaction to the news takes some time δ , the new logprice $p_{t+\delta}$ will result only at time $t + \delta$,

$$p_{t+\delta} = p_t + k_t, \tag{3.2}$$

while between times t and $t + \delta$, the observed logprices somehow adjust to their new level. This type of jump was first mentioned by Barndorff-Nielsen et al. (2008b) and is since then known as a *gradual jump*.

Consequently, both types of jumps were investigated as both are vital for forecasting future price movements.

3.1 Estimating integrated quarticity

In order to establish nonparametric test statistics to identify jumps one needs to estimate the unknown integrated quarticity, IQ . One of the most common estimators for estimating IQ in the presence of jumps is the *realised quadpower quarticity*, QP , designed by Barndorff-Nielsen and Shephard (2004)

$$QP^{(m)} = m\mu_1^{-4} \frac{m}{m-3} \sum_{i=4}^m \prod_{j=0}^3 |r_{j-i}|, \quad (3.3)$$

where μ_1 is defined as in Equation (2.12). Andersen et al. (2007a) proposed to estimate IQ using *realised tripower quarticity*, TP , defined as

$$TP^{(m)} = m\mu_{4/3}^{-3} \frac{m}{m-2} \sum_{i=3}^m \prod_{j=0}^2 |r_{j-i}|^{4/3}, \quad (3.4)$$

where $\mu_{4/3}$ is defined as in Equation (2.12). Both estimators are based on realised multipower variations and belong therefore to the classic approach.

Another estimator, developed by Kloessner (2009), is based on intradaily highs and lows to promise a more precise estimation of IQ . Its compounds are estimates of the integrated quarticity based on (positive, negative) returns according to

$$IQp = \sum_{p_{\frac{i-1}{m}} < p_{\frac{1}{m}}} \frac{16}{3} \left[\left\{ (p^*)_{i,N} - p_{\frac{i}{m}} \right\}^4 + \left\{ p_{\frac{i-1}{m}} - (p^*)_{i,N} \right\}^4 \right], \quad (3.5)$$

$$IQn = \sum_{p_{\frac{i-1}{m}} > p_{\frac{1}{m}}} \frac{16}{3} \left[\left\{ (p^*)_{i,N} - p_{\frac{i-1}{m}} \right\}^4 + \left\{ p_{\frac{i}{m}} - (p^*)_{i,N} \right\}^4 \right], \quad (3.6)$$

$$IQz = \sum_{p_{\frac{i-1}{m}} = p_{\frac{1}{m}}} \frac{16}{3} \left[\left\{ (p^*)_{i,N} - p_{\frac{i-1}{m}} \right\}^4 + \left\{ p_{\frac{i}{m}} - (p^*)_{i,N} \right\}^4 \right], \quad (3.7)$$

whereas the final estimator is defined as

$$IQ_{OHLC} = \frac{1}{2}IQp + \frac{1}{2}IQn + IQz. \quad (3.8)$$

3.2 Test statistics to detect days with a jump

The results in Chapter 2 provide a basis for nonparametric statistics to identify jumps. Barndorff-Nielsen and Shephard (2004) and Barndorff-Nielsen and Shephard (2006) have shown that in the absence of jumps in the price process,

$$\frac{1}{m}^{-1/2} \frac{RV^{(m)} - BPV^{(m)}}{\{(\pi^2/2 + \pi - 1)IQ\}^{1/2}} \xrightarrow{p} N(0, 1) \text{ as } m \rightarrow \infty. \quad (3.9)$$

Hence, a test statistic based on Equation (3.9) is given by

$$\frac{\sqrt{m}RV^{(m)} - BPV^{(m)}}{\{(\pi^2/2 + \pi - 1) \cdot Q^*\}^{1/2}}, \quad (3.10)$$

with $Q^* \ni (TP^{(m)}, QP^{(m)})$.

A number of variations of these test statistic, all of which asymptotically standard normal distributed, were proposed by Barndorff-Nielsen and Shephard (2004), Barndorff-Nielsen and Shephard (2006). A logarithmic form of the statistic is given by,

$$z_{TPL} = \frac{\log(RV^{(m)}) - \log(BPV^{(m)})}{\sqrt{(\pi^2/2 + \pi - 1) \cdot \frac{1}{m} \cdot \frac{TP^{(m)}}{\{BPV^{(m)}\}^2}}}, \quad (3.11)$$

and a similar version with an added maximum adjustment due to a Jensen's inequality argument

$$z_{TPLM} = \frac{\log(RV^{(m)}) - \log(BPV^{(m)})}{\sqrt{(\pi^2/2 + \pi - 1) \cdot \frac{1}{m} \cdot \max\left[1, \frac{TP^{(m)}}{\{BPV^{(m)}\}^2}\right]}}. \quad (3.12)$$

Analogous, one can use statistics based on $QP^{(m)}$ as defined in Equation (3.3)

$$z_{QPL} = \frac{\log(RV^{(m)}) - \log(BPV^{(m)})}{\sqrt{(\pi^2/2 + \pi - 1) \cdot \frac{1}{m} \cdot \frac{QP^{(m)}}{\{BPV^{(m)}\}^2}}}, \quad (3.13)$$

$$z_{QPML} = \frac{\log(RV^{(m)}) - \log(BPV^{(m)})}{\sqrt{(\pi^2/2 + \pi - 1) \cdot \frac{1}{m} \cdot \max\left[1, \frac{QP^{(m)}}{\{BPV^{(m)}\}^2}\right]}}. \quad (3.14)$$

Andersen et al. (2007a) and Huang and Tauchen (2005) suggest to using the ratio

$$RJ^{(m)} = \frac{RV^{(m)} - BPV^{(m)}}{RV^{(m)}}$$

instead of the logarithmic difference between $RV^{(m)}$ and $BPV^{(m)}$. As the difference between $RV^{(m)}$ and $BPV^{(m)}$ estimates the jump component and $RV^{(m)}$ estimates the total variation, this ratio is an estimator of the relative distribution of the jump com-

3 Identifying jumps

ponent to the total variation. The following statistics are based on this ratio

$$z_{TPR} = \frac{RJ^{(m)}}{\sqrt{(\pi^2/2 + \pi - 1) \cdot \frac{1}{m} \cdot \frac{TP^{(m)}}{\{BPV^{(m)}\}^2}}}, \quad (3.15)$$

$$z_{TPRM} = \frac{RJ^{(m)}}{\sqrt{(\pi^2/2 + \pi - 1) \cdot \frac{1}{m} \cdot \max \left[1, \frac{TP^{(m)}}{\{BPV^{(m)}\}^2} \right]}}, \quad (3.16)$$

$$z_{QPR} = \frac{RJ^{(m)}}{\sqrt{(\pi^2/2 + \pi - 1) \cdot \frac{1}{m} \cdot \frac{QP^{(m)}}{\{BPV^{(m)}\}^2}}}, \quad (3.17)$$

$$z_{QPRM} = \frac{RJ^{(m)}}{\sqrt{(\pi^2/2 + \pi - 1) \cdot \frac{1}{m} \cdot \max \left[1, \frac{QP^{(m)}}{\{BPV^{(m)}\}^2} \right]}}, \quad (3.18)$$

where Huang and Tauchen (2005) singled out the test statistic in Equation (3.16) as the one with the best properties with respect to size, followed by the test statistic z_{QPLM} in Equation (3.14).

These test statistics were later used to test the null hypothesis, H_0 , that there is no detectable jump in the logreturn process during a day, where the hypothesis is rejected for large values of statistics relative to the standard normal distribution. Since the statistics are based on the difference between two variances, where the difference is zero under the null hypothesis and greater than zero otherwise, the test is right-sided. The alternative hypothesis can only be, due to the fact that small jumps relative to the diffusion or noise processes are unlikely to be identifiable, to find detectable jumps.

Kloessner (2010) argued, that these test statistics have, especially for high-frequencies, only poor power against the alternative, as gradual jumps will produce small or even negative values for the numerators. An important consequence is that mathematical and gradual jumps tend to cancel out each other if they occur on the same day.

Another family of test statistics is based on the subperiod's highest and lowest return, again following the ideas of Kloessner (2010). Given a consistent estimate of IQ (see Section 3.1), the test statistics

$$TJ = \frac{\sqrt{m} \sum_{i=1}^m (SSJ_p)_i}{\sqrt{1.3014 \hat{I}Q}}, \quad (3.19)$$

$$TJ_p = \frac{\sqrt{m} \sum_{i=1}^m (SSpJ_t)_i}{\sqrt{0.8602 \hat{I}Q}}, \quad (3.20)$$

$$TJ_n = \frac{\sqrt{m} \sum_{i=1}^m (SSnJ_t)_i}{\sqrt{0.8602 \hat{I}Q}}, \quad (3.21)$$

with

$$(SSpJ_t)_i = \begin{cases} 0, & r < 0, \\ b^2 + 1.3982(uw^2 + lw^2) - 3.968b(lw + uw) \\ \quad + 2.0902(uw \cdot lw), & r > 0, \end{cases} \quad (3.22)$$

$$(SSnJ_t)_i = \begin{cases} b^2 + 1.3982(uw^2 + lw^2) - 3.968b(lw + uw) \\ \quad + 2.0902(uw \cdot lw), & r < 0, \\ 0, & r > 0, \end{cases} \quad (3.23)$$

are asymptotically standard normal in the absence of mathematical (positive, negative) jumps (H_0).

3.3 Determine jump size and intensity

Up until now the understanding of the jump process remains limited as it is still not possible to split up the daily overall contribution of jumps to QV into separate jumps. It is therefore important to estimate the number of jumps per day, determine the sign of the jump and the jump height. In order to accomplish this mission two algorithm are discussed in this Section.

The first algorithm, developed by Andersen et al. (2007b) and further improved by Ane and Metais (2010), is based on the classic approach. All days with a critical value above the value of a chosen test statistic are considered to have no jumps at a significance level of α ; the number of jumps for those days is set to zero. Otherwise at least one jump has occurred. The authors define the interval in which the jump has occurred in a natural way as the interval $j = [(j - 1)m^{-1}, jm^{-1}]$, with the highest squared return on that day

$$\{r_j^{(m)}\}^2 = \arg \max_{i=1, \dots, m} \{r_i^{(m)}\}^2. \quad (3.24)$$

Following Ane and Metais (2010) it can be assumed that the discontinuity in the j^{th} intraday return dominates the diffusive component and therefore the sign of $r_j^{(m)}$ is the same as the sign of the jump, i.e. for $r_j^{(m)} < 0$ they assume a negative jump and vice versa.

In addition and in order to determine the (based on the chosen sampling frequency) exact time of the jump, the starting time of the corresponding interval, $(j - 1)m^{-1}$, is noticed.

After identifying the first (and largest) jump of that day they look at additional jumps. As estimators for IV , like $BPV^{(m)}$ defined in Equation (2.11), are robust to the presence of jumps, these estimates will remain unchanged in the following. But the estimates for

3 Identifying jumps

QV have to be recalculated excluding the jump interval, i.e.

$$\widehat{QV_{i \neq j}^*} = \lim_{m \rightarrow \infty} \sum_{i=1}^m r_i^{(m)}, \quad i = 1, 2, \dots, j-1, j+1, \dots, m. \quad (3.25)$$

As this would lead to a downward bias Ane and Metais (2010) suggest replacing $\{r_j^{(m)}\}^2$ by the mean of all squared returns (up to now) unaffected by a jump

$$\{r_j^{(m)}\}^2 = \frac{1}{m-1} \sum_{i=1, i \neq j}^m \{r_i^{(m)}\}^2. \quad (3.26)$$

Subsequently, based on this new sequence of squared returns, the new value for the test statistic is computed. Again, if the critical value is above the value of the test statistic of this day, it is marked as a day with one jump, otherwise the procedure repeats itself. The algorithm used in this work follows the guidelines and ideas of Andersen et al. (2007b) and Ane and Metais (2010) up to this point.

The construction of an estimate of the jump size is not straight forward, as more than one jump can be recorded per day. Andersen et al. (2007b) suggest simply setting the jump size equal to the difference between $RV^{(m)}$ and $BPV^{(m)}$ as this quantity approximates the sum of all squared intraday jump sizes. In this work the jump size, k_j , is computed as a weighted difference between $RV^{(m)}$ and $BPV^{(m)}$

$$k_j = \frac{\{r_j^{(m)}\}^2}{\sum_{i=1}^m \{r_i^{(m)}\}^2} \cdot (RV^{(m)} - BPV^{(m)}). \quad (3.27)$$

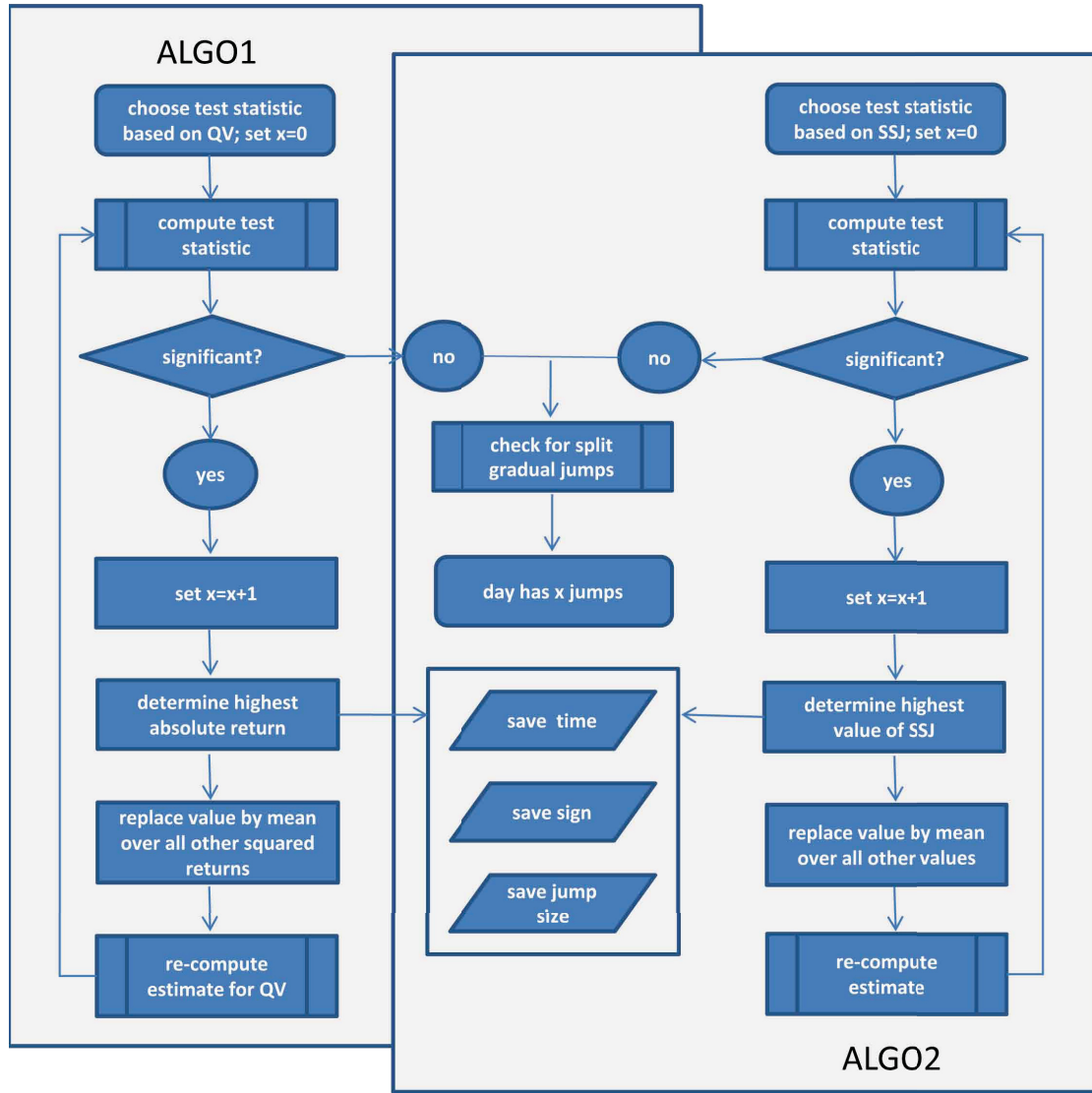
In addition, the sequence of times at which a jump occurs is investigated for consecutive times for every given day. If this is the case and if the recorded signs of the jumps at these times are the same, it can be assumed that the procedure has split up one gradual jump into several parts. To account for this possibility both entries are matched together keeping the starting time of the second entry and the sum of both jump sizes.

For further reference the algorithm described above is called *ALGO1*.

This improved algorithm is now adapted and applied to the estimators presented by Kloessner (2010). In contradiction to *ALGO1*, this algorithm is based on intraday lows and highs. Again, all days with a significant test statistic for identifying positive (negative) jumps, see Equations (3.19),(3.20) and (3.21), are considered to have at least one jump; the number of jumps for the other days is set to zero.

For significant days the highest value of $SSpJ$ ($SSnJ$) is determined as the value capturing the highest jump on that day

$$SSpJ_j = \arg \max_{i=1, \dots, m} SSpJ_i, \quad SSnJ_j = \arg \max_{i=1, \dots, m} SSnJ_i, \quad (3.28)$$


 Figure 3.1: Flowchart of *ALGO1* and *ALGO2*.

respectively. This is due to the fact that the test statistics based on intradaily highs and lows are based on the values of $SSpJ$ ($SSnJ$). In addition, the starting time $(j)m^{-1}$ at which $SSpJ_j$ ($SSnJ_j$) occurred is recorded as the jump time on that day.

Following the same argument as above, $SSpJ_j$ ($SSnJ_j$) is replaced by the mean over all $SSpJ_i$ ($SSnJ_i$), $i \neq j$ on that day. Based on this new sequence, the value for the test statistic is recomputed to determine additional significant jumps on that day. The jump height is set to the value of

$$k_j = SSpJ_j, \quad k_j = -SSnJ_i, \quad (3.29)$$

respectively. The jump size therefore represents - with the contribution of the jump to

3 Identifying jumps

the overall variability on that day - a weighted estimate of the realised volatility. As in *ALGO1*, a procedure to identify split gradual jumps is implemented.

Henceforth, the algorithm based on the classical approach, illustrated in Figure 3.1 is called *ALGO1*. The second algorithm is called *ALGO2*.

Both algorithms can, at least, be applied to data sets containing logreturns and intradaily highs and lows sampled at a reasonable frequency. Nevertheless, the results of this work are grounded on high frequency data sets, which are described in Chapter 4.

4 Applied data cleaning and management procedure

In the empirical part of this work, trades and quotes data (TAQ) from the New York Stock Exchange (NYSE), which contain both transaction and quote data, was analysed. The data set consists of recorded transactions for three stocks, namely *Alcoa Inc.*, *Citigroup* and *IBM*, in the period from January 2008 till July 2009, covering the financial crisis. The whole data set consists of 54 files each capturing a one month period for one stock. Due to the huge amount of data - requiring a total of approximately seven gigabyte of hard disk space - efficient programming and data handling is vital.

Table 4.1 provides a short excerpt of the data set, whereas the most important variables are listed in Table 4.2.

date	time	price	size	cond	corr	bid	bid_size	offer	offer_size
2008-01-02	34204	0.00	0		0	29.67	1	29.69	361
2008-01-02	34204	0.00	0		0	29.68	1	29.69	372
2008-01-02	34204	29.68	200	F	0	29.67	1	29.69	353
2008-01-02	34204	29.69	100	@	0	29.67	1	29.69	353
2008-01-02	34204	29.69	200	@	0	29.67	1	29.69	353
2008-01-02	34204	29.69	300	F	0	29.67	1	29.69	353
2008-01-02	34205	0.00	0		0	29.67	1	29.69	351
2008-01-02	34205	0.00	0		0	29.67	1	29.69	352
2008-01-02	34205	29.69	100	@	0	29.67	1	29.69	351
2008-01-02	34205	29.69	100	@	0	29.67	1	29.69	351
2008-01-02	34205	29.69	300	@	0	29.67	1	29.69	351
2008-01-02	34206	0.00	0		0	29.66	7	29.69	355

Table 4.1: Excerpt of the data set for Citigroup on 2 January 2008.

By looking closely at Table 4.1, several characteristics immediately point to the programming difficulties associated with the data.

First of all, the data set contains lots of redundant information which should be deleted in order to decrease the computation time. This means entries with $CORR \neq 0$ (S3) and abnormal *sale condition* (S4) were deleted. In addition, all entries before 9:30am and after 4:00pm, the time at which the exchange is closed, are deleted (S1). This

Variable	Description
date	Date of the transaction
time	Trade time in cumulative number of seconds since midnight
price	Actual trade price per share
size	Number of shares traded
cond	Sale condition; "@" indicating a regular sale (no condition)
corr	Correction indicator
bid	Bid price
bid_size	Bid size in units of trade
offer	Offer price
offer_size	Offer size in units of trade

Table 4.2: TAQ Quote data description (according to the NYSE TAQ manual).

eliminates outliers and strong fluctuations, typically observed at the start and end of a trading day (Barndorff-Nielsen et al. (2008a)).

In this analysis trade values, rather than midquotes, were used because they are naturally closer to the *true* price of an asset. Therefore all entries with a bid, ask or transaction price equal to zero were deleted (S2).

As shown by Brownless and Gallo (2006), proper data cleaning is one of the most important steps in computing realised volatility from high frequency data. Therefore all entries with prices above the *ask* plus the bid-ask spread as well as entries with prices below the *bid* minus the bid-ask spread were deleted to tame the trade data using quotes (S6).

As most of the estimators discussed in Chapter 2 treat all observations equally, a few outliers can heavily influence the results. So one important goal is to eliminate outliers. To detect outliers the procedure described in Brownless and Gallo (2006) was used (BG). Formally, let $\{p_i\}_{i=1}^N$ be the ordered tick-by-tick series. Remove false observation i if

$$|p_i - \bar{p}_i(k)| > 3 \cdot s_i(k) + \gamma, \quad (4.1)$$

where $\bar{p}_i(k)$ and $s_i(k)$ denote the moving mean and standard deviation of k observations, respectively. The addition of γ secures non-zero variances by sequences of k equal prices. Of course, this method suffers from end effects as outliers in the first and last $k/2$ observations cannot be detected. In this analysis $k = 60$ and $\gamma = 0.03$. This corresponds to a moving mean over one minute of observations with a medium value for γ , as suggested by Brownless and Gallo (2006) for frequently traded assets.

Secondly, each observation has a label for date and time in cumulative milliseconds since midnight. To make every trade identifiable by a simple unique value, both columns have to be matched together.

Abbr.	Description
S1	Delete entries with a time stamp outside 9:30am and 4:00pm
S2	Delete entries with a bid, ask or transaction price equal to zero
BG	Delete entries which difference to a rolling mean exceeds a specific threshold (see equation (4.1))
S3	Delete entries with corrected trades
S4	Delete entries with abnormal <i>Sale Condition</i>
S5	If multiple transactions have the same time stamp, replace them with the average price weighted by the size of the trade
S6	Delete entries with prices above the <i>ask</i> plus the bid-ask spread as well as entries with prices below the <i>bid</i> minus the bid-ask spread

Table 4.3: Applied cleaning and data management steps.

Finally, it is not untypical for frequently traded stocks that multiple trades at various prices do occur at the same time point. In order to generate a homogeneous time series this issue has to be addressed. One way is to replace such transactions by the median price, which is robust against outliers. On the one hand this is a reasonable and quite fast computational procedure, on the other hand one can perform better in means of statistical accuracy, especially as most of the crucial outliers were removed in the previous step. Therefore the following, much more computationally intensive, procedure was applied, for all transactions with the same time stamp a weighted average with weights equal to the size of the transactions was computed (S5).

Table 4.5 gives a summary of the reduced observations in each cleaning and data management step. One can see that the reduction of observations due to S3 is negligible. BG seems to perform well as in the month before and after the impact of the financial crisis in October 2008 and only very few observations were deleted, as one would expect to have only very few outlier in this period. Therefore, this result is in line with the results of Brownless and Gallo (2006). Interestingly, BG detects a significant growth in the number of outliers in October 2008, which indicates that the price finding on the market during extrem events is somehow hindered, resulting in single extreme prices compared to the overall level of the price on that day.

In general, it can be seen that the number of observations in the raw data is much higher for the stock of Citigroup from August until October 2008, afterwards the stock was traded much less frequently whereas the trading frequency of the other stocks seem to be much more stable.

After all data cleaning and management steps were applied, each month consists of roughly about 150,000 observations, which corresponds to one trade every 16 seconds on average.

Time	Price
(01/02/08 09:30:00)	29.68875
(01/02/08 09:30:01)	29.68875
(01/02/08 09:30:02)	29.68875
(01/02/08 09:30:03)	29.68875
(01/02/08 09:30:04)	29.68875
(01/02/08 09:30:05)	29.69000
...	
(01/02/08 15:59:56)	28.94757
(01/02/08 15:59:57)	28.96000
(01/02/08 15:59:58)	28.96000
(01/02/08 15:59:59)	28.93032

Table 4.4: Excerpt of the generated homogeneous time series.

As the definition of $RV^{(m)}$ imposes no particular requirement on the way in which prices are sampled, as long as the corresponding returns are non overlapping and span the interval of interest, it is not surprising that a variety of different sampling schemes have been used in the literature. In this work the most widely used sampling scheme, called calendar time sampling (CTS), is used, in which the intervals between observations are equidistant in calendar time, i.e. the data is sampled every 5 or 10 minutes.

The construction of CTS is not straight forward as in practice intraday data are irregularly spaced. Hansen and Lunde (2006) showed that the *previous tick method*, in which the first observation in an interval (possibly containing several prices) is used, is a reasonable way to sample prices in calendar time.

Therefore it is necessary, from a programming point of view, to create a homogeneous time series where at every second during the trading day information is available. To ensure this, the *last-observation-carried-forward* procedure was applied, where at each second with no recorded transaction, the value of the previous transaction was used. In this way it is assured that it is possible to create various sub grids on the time series. An excerpt of the resulting homogeneous time series is given in Table 4.4.

Month	Stock	Number of obs. in		Reduced observations in step					
		raw data	clean data	S1	S2	S3	S6	BG	S5
Jan 08	AA	2,442,814	143,449	83	2,172,090	0	13,936	1	113,255
	C	7,532,198	260,001	103	6,806,038	0	45,772	0	420,284
	IBM	2,221,869	162,847	75	1,899,794	0	16,498	9	142,646
Feb 08	AA	1,713,258	108,582	71	1,516,581	0	7,706	0	80,318
	C	6,104,981	204,319	85	5,613,500	0	19,040	0	268,037
	IBM	1,938,002	123,651	71	1,702,108	0	10,943	1	101,228
Mar 08	AA	2,201,960	127,100	70	1,958,737	0	7,231	0	108,822
	C	8,019,886	232,620	127	7,384,469	0	20,926	0	381,744
	IBM	2,029,604	120,151	94	1,799,846	0	7,299	5	102,209
May 08	AA	1,628,323	105,433	95	1,447,271	0	3,148	0	72,376
	C	5,701,666	204,103	96	5,229,855	0	6,739	0	260,873
	IBM	1,590,750	110,784	77	1,398,355	0	3,750	3	77,781
Apr 08	AA	1,612,033	108,717	72	1,415,803	0	3,680	0	83,761
	C	5,019,603	165,120	85	4,656,321	0	4,034	0	194,043
	IBM	1,247,016	100,349	67	1,065,943	0	2,731	0	77,926
Jun 08	AA	1,384,235	110,505	70	1,180,785	0	3,229	0	89,646
	C	6,225,996	198,070	109	5,735,934	0	5,692	0	286,191
	IBM	1,394,567	104,095	80	1,193,260	0	3,603	2	93,527
Jul 08	AA	1,465,534	142,999	88	1,189,245	0	4,641	2	128,559
	C	6,152,361	229,350	133	5,544,935	0	14,017	0	363,926
	IBM	1,373,269	124,679	80	1,119,033	0	4,072	7	125,398
Aug 08	AA	767,672	83,910	73	627,558	0	1,721	0	54,410
	C	2,235,957	152,654	117	1,887,067	0	3,934	2	192,183
	IBM	925,384	87,050	74	768,149	0	1,662	5	68,444
Sept 08	AA	1,237,372	109,852	91	1,030,441	0	1,719	1	95,268
	C	2,963,766	215,448	113	2,319,458	1	10,955	44	417,747
	IBM	1,380,308	120,787	79	1,122,566	0	5,492	38	131,346
Oct 08	AA	1,642,509	131,164	109	1,346,844	0	2,058	6	162,328
	C	3,167,247	220,073	130	2,534,688	1	8,041	42	404,272
	IBM	2,711,645	165,422	104	2,305,111	0	7,765	145	233,098
Nov 08	AA	1,062,644	81,088	120	909,764	0	353	0	71,319
	C	3,507,130	184,840	136	2,974,830	0	3,586	0	343,738
	IBM	2,109,657	123,829	77	1,814,149	0	4,424	14	167,164
Dec 08	AA	1,092,733	95,815	293	928,606	0	520	0	67,499
	C	2,784,855	161,660	181	2,415,587	0	1,373	0	206,054
	IBM	1,962,992	117,977	104	1,706,919	0	2,441	9	135,542

Table 4.5: Number of observations after each cleaning step (see Table 4.3).

Month	Stock	Number of obs. in		Reduced observations in step					
		raw data	clean data	S1	S2	S3	S6	BG	S5
Jan 09	AA	1,367,546	103,882	180	1,165,406	1	672	0	97,405
	C	3,139,450	164,372	190	2,719,947	0	3,358	0	251,583
	IBM	1,805,597	120,137	137	1,540,691	1	3,809	3	140,819
Feb 09	AA	1,457,106	86,202	149	1,264,903	0	843	0	105,009
	C	2,413,151	111,797	145	2,142,008	0	753	0	158,448
	IBM	2,066,499	120,464	92	1,808,429	0	5,004	1	132,509
Mar 09	AA	2,095,203	116,937	231	1,793,099	0	3,047	1	181,888
	C	1,453,739	83,201	99	1,293,625	0	303	0	76,511
	IBM	2,717,439	149,763	117	2,394,157	0	4,724	2	168,676
Apr 09	AA	1,665,620	78,294	164	1,488,225	0	1,042	0	97,895
	C	1,453,739	83,201	99	1,293,625	0	303	0	76,511
	IBM	2,370,761	122,780	123	2,106,837	0	3,620	4	137,397
May 09	AA	1,411,984	73,287	114	1,254,316	0	730	0	83,537
	C	1,453,739	83,201	99	1,293,625	0	303	0	76,511
	IBM	1,984,954	97,493	63	1,791,068	0	2,916	2	93,412
Jun 09	AA	2,036,834	92,790	139	1,835,804	0	891	0	107,210
	C	1,115,701	58,972	100	998,942	0	65	0	57,622
	IBM	1,743,168	88,731	70	1,570,824	0	2,548	2	80,993
Jul 09	AA	2,131,953	81,147	152	1,965,562	0	972	0	84,120
	C	1,468,356	82,883	126	1,264,040	0	273	0	121,034
	IBM	1,841,431	84,191	80	1,675,317	0	2,248	0	79,595

Table 4.6: Number of observations after each cleaning step (continued).

5 Empirical Analysis

Having a cleaned and homogeneous time series to hand one can proceed with the empirical analysis of the data set.

In order to determine an appropriate sampling frequency, so-called "*Volatility Signature Plots*" were investigated. Andersen et al. (2000) introduced this kind of signature plot, which plots average realised variance (on one day) against the sampling frequency.

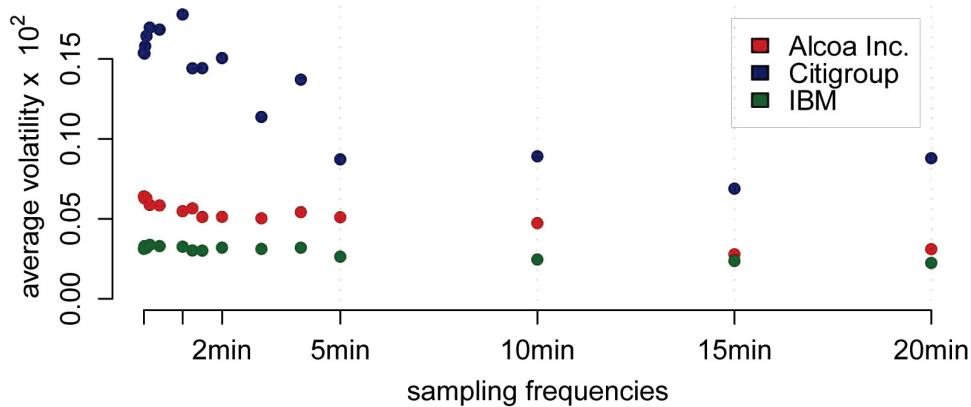



Figure 5.1: Volatility Signature Plot for the stocks of Alcoa Inc. (red), Citigroup (blue) and IBM (green).  JHFDVolaSig

Figure 5.1 shows exemplarily the volatility signature plot constructed for the three stocks on a typical day. The plots (not all shown here) support 5–10 minute sampling intervals. Indeed, for all three stocks the realised volatility estimates increase as the sampling frequency increases. This phenomenon is due to the existence of market microstructure noise in the data and would lead to spurious jumps in the data. As highlighted earlier, this effect prohibits the use of very high sampling frequencies. On the other hand, as all test statistics are based on asymptotical results, it is essential to sample as frequently as reasonable. The following calculations are therefore based on five minutes logreturns, corresponding to 78 sampling intervals during one trading day. Finally, the choice of five minutes logreturns is in line with most of the existing empirical literature.

5.1 Preliminary analysis

Naturally, one starts by investigating the recorded price series. Figure 5.2 on page 27 provides an overview of the stocks logprices and logreturns from January 2008 to July 2009. The logprices are plotted to ensure a better comparison of all three stocks as a change of one unit always corresponds to a change of 1% in the stockprice. A huge decrease in all three stock prices can be noticed starting in October 2008 and falling at the time of the financial crisis. The series of logreturns appears to be smooth until that moment and afterwards starts to get quite rough with high positive and negative peaks for all three stocks indicating rapid price movements. This effect can be seen for Alcoa Inc. as well as for Citigroup.

The logreturn series is characterised by the usual traits of financial data as they are: the presence of fatter tails than those of a Gaussian distribution and a negative asymmetry. Furthermore, the graphs confirm the presence of heteroscedasticity clustered in high and low volatility periods for all three stocks.

Interestingly, this specific pattern can also be found in the daily volatility estimates. Figure 5.3 displays estimates of $RV^{(78)}$ for all three stocks. It can be seen that the estimates are also clustered in high and low periods, nearly perfectly matching the periods of the logreturns for all three stocks.

As both algorithm rely heavily on the chosen test statistic, one should have a closer look at the ones available (see Section 3.2). It is natural to choose both test statistics in Equations (3.20) and (3.21) in order to determine positive and negative jumps in *ALGO2*, as these are the only ones that rely on intradaily highs and lows and allow for separating positive and negative jumps. In *ALGO1* the feasible logarithmic test statistic z_{QPLM} from Equation (3.12) is used. This test statistic performs well at for five minutes logreturns in the sense of detecting 'jumpy' days for all stocks. The other test statistics described in Section 3.2 perform less well as they either detect jumps on every day or on (nearly) no days. As this test statistic performed very well in the simulations of Huang and Tauchen (2005) it seems reasonable to use z_{QPLM} .

The results of all three test statistics are displayed in Figure 5.4. It can be seen that the resulting pattern for each stock is quite comparable and matches the one already revealed in the logreturns and volatility estimates. However, TJ_p and TJ_n take on values much higher than the corresponding values of z_{QPLM} , strongly indicating jump(s) on these days.

To conclude this preliminary analysis, Table 5.1 summarises the findings up to here. Both test statistics indicate, at a significance level of 5%, a comparable amount of days with jump(s), ranging from 39–57% of the days. Importantly, only 25% of the days were identified by both test statistics. In order to integrate these findings into the existing literature, one has to mention that these values indicate a quite high amount of days

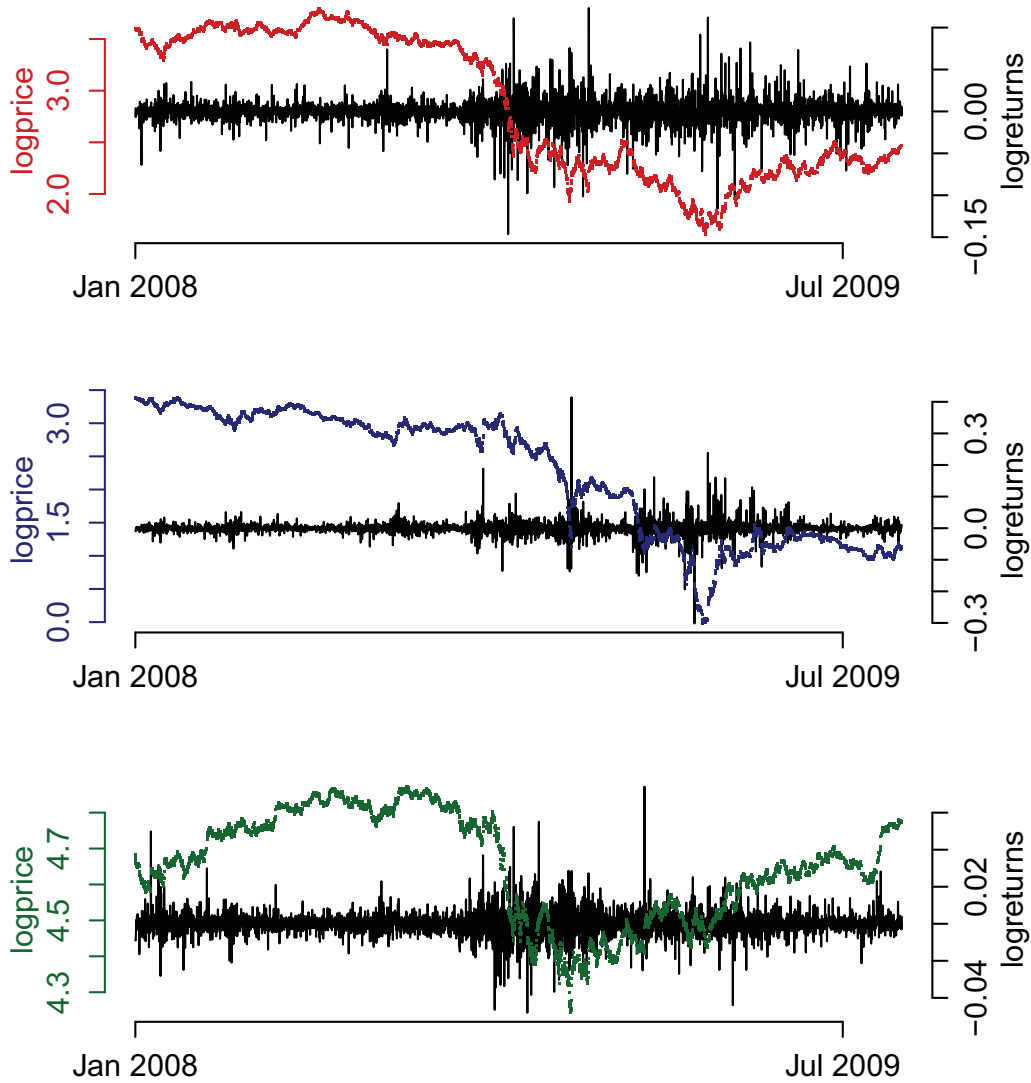


Figure 5.2: Logprices and log returns for the stocks of Alcoa Inc. (red), Citigroup (blue) and IBM (green) from January 2008 - July 2009 based on 30 minute intervals.
 JHFDLogPrRe

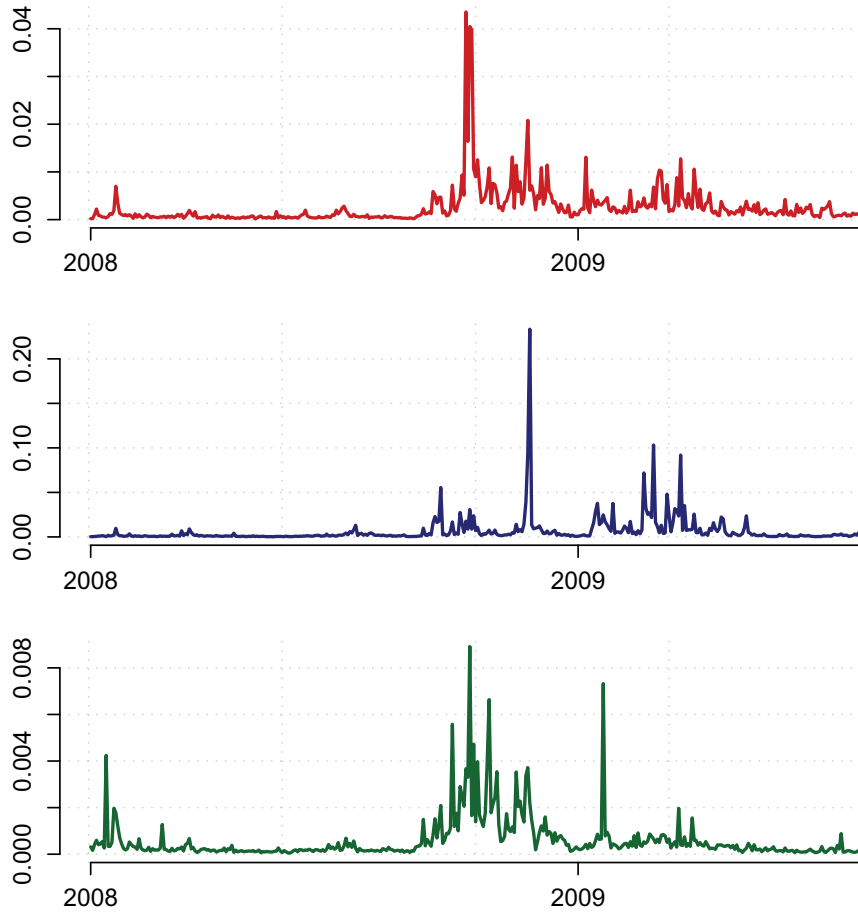


Figure 5.3: Volatility estimates for the stocks of Alcoa Inc. (red), Citigroup (blue) and IBM (green) from January 2008 – July 2009 based on 5 minute intervals.

with jump(s): Ane and Metais (2010) found values between 17–44% for European stock indices covering the seven-year period 2000 – 2007, Andersen et al. (2007b) reported values of around 16% for the 30 Dow Jones Industrial Average (DJIA) stocks for a five-year period spanning 1998 – 2002 and Huang and Tauchen (2005) with a ratio of about 26% for the S&P 500 Cash Index between 1997 – 2002, among others.

Nevertheless, one should expect a higher numbers of days with a significant jump test statistic in this work due to two reasons. Firstly, the test statistics TJ_p and TJ_n are the only test statistics which are able to detect gradual jumps (to the best of the author's knowledge), which means the number of gradual jumps is not included in the empirical results published previously. Secondly, this paper covers the dramatic changes of the financial crisis. Therefore, these results could be an effect of the financial crisis as the lowest ratio of days with jump(s) is recorded for IBM, which also has the lowest estimates of realised volatility and the smallest break in the logprice series.

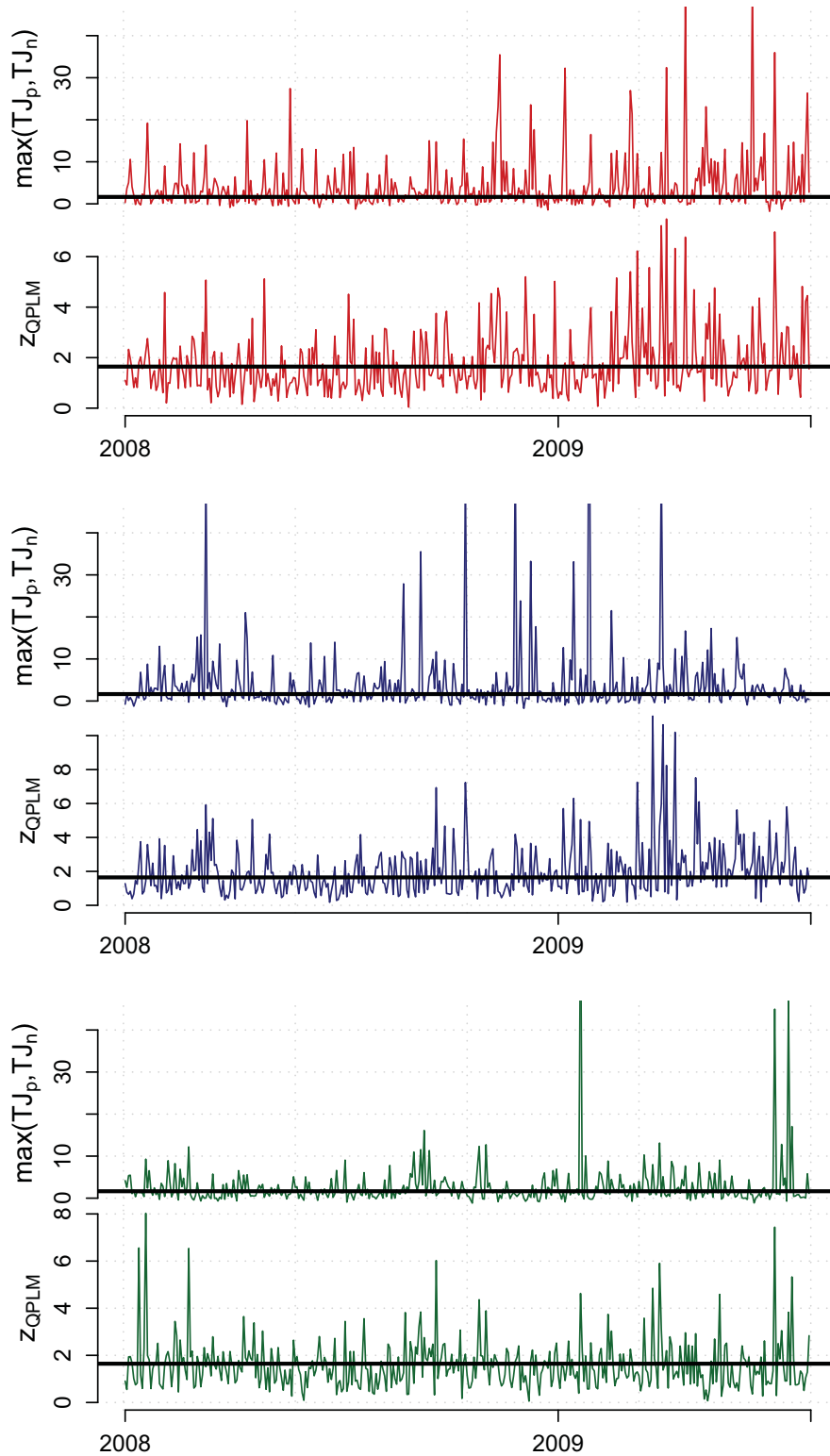



Figure 5.4: Values of the test statistics z_{QPLM} (top) and $\max(TJ_p, TJ_n)$ (bottom) for the stocks of Alcoa Inc. (red), Citigroup (blue) and IBM (green) from January 2008 – July 2009 based on 5 minute intervals.  JHFDGTeststat

Number of days with significant test statistic ...	AA	C	IBM
$\alpha = .95$			
$\max(TJ_p, TJ_n)$	228	203	175
z_{TPQM}	168	192	158
both test statistics are significant	112	124	97

Table 5.1: Number of days with jumps (n=399).

5.2 Jump intensity

Despite these first results, insight into the jump process is still limited. To overcome this, both algorithms discussed in Section 3.3 were applied to the data.

Having the exact jump count to hand, one can give an estimate of the jump intensity by dividing the number of jumps by the total number of days. Table 5.2 reports this value as well as the total number of positive (negative) jumps for all three stocks.

It can be seen that the number of jumps (and therefore the estimated jump intensity) reported by *ALGO2* is always higher than the value computed by *ALGO1*, ranging from 0.53 – 0.99 and 0.41 – 0.62, respectively. The total number of detected jumps for Alcoa Inc. is even twice as high as the value reported by *ALGO1*. These general results also hold for the total number of positive (negative) jumps in the data. A reason could be the better detection rate for gradual jumps of the underlying test statistic, or the fact that mathematical and gradual jumps cancel out each other for z_{QPLM} . Nevertheless, the results concerning the total number of jumps and the intensity of jumps computed by *ALGO1* are in line with the values recently reported by Ane and Metais (2010).

Interestingly, both algorithms detect a comparable ratio of positive and negative jumps for every stock. So positive and negative jumps seem to happen equally often and do not show any remarkable differences in their intensity.

For further inspections the jump days were sorted by the detected number of jumps. Recall that both algorithms rely on the selected confidence level α and, consequently, so does the number of days exhibiting jumps. Therefore the robustness of the results of both algorithm with respect to the chosen confidence level is of importance. Hence, both algorithms were applied with different values of α ranging from 0.95 to 0.999. Mathematically, for lower values of α the overall number of detected jumps increases. Therefore a rise in the maximum number of jumps could be expected. The opposite holds for higher values of α . Table 5.3 summarises the results.

For all stocks (and across both algorithms), more than 25% of days with a significant test statistic exhibit only a single jump at a confidence level of $\alpha = 0.95$. If α increases,

	ALGO2			ALGO1		
	AA	C	IBM	AA	C	IBM
$\alpha = .95$						
Total number of						
... jumps	398	334	214	176	250	166
... positive jumps	200	169	120	89	124	78
... negative jumps	198	165	94	87	126	84
Intensity of						
... jumps	.99	.83	.53	.44	.62	.41
... positive jumps	.50	.42	.30	.22	.31	.19
... negative jumps	.49	.41	.23	.21	.30	.21

Table 5.2: Jump Intensity for assets of Alcoa Inc., Citigroup and IBM from January 2008 - July 2009 (n=399) based on $\alpha = .95$ for 5 minutes logreturns.

Number of days with	ALGO2			ALGO1		
	AA	C	IBM	AA	C	IBM
$\alpha = .95$						
... 1 jump	116	139	97	160	167	150
... 2 jumps	74	32	44	8	14	8
... 3 jumps	25	19	17	0	4	0
... > 3 jumps	13	13	17	0	7	0
Max. number of jumps per day	6	11	7	2	11	2
$\alpha = .99$						
... 1 jump	127	131	87	91	114	62
... 2 jumps	46	23	24	0	3	8
... 3 jumps	13	7	7	0	1	0
... > 3 jumps	4	5	10	0	2	0
Max. number of jumps per day	5	5	5	1	9	2
$\alpha = .999$						
... 1 jump	116	103	77	50	67	24
... 2 jumps	25	21	13	0	2	0
... 3 jumps	4	1	5	0	0	0
... > 3 jumps	2	1	1	0	1	0
Max. number of jumps per day	4	4	4	1	6	1

Table 5.3: Robustness of jump intensity for $\alpha = \{.95, .99, .999\}$ based on 5 minutes logreturns.

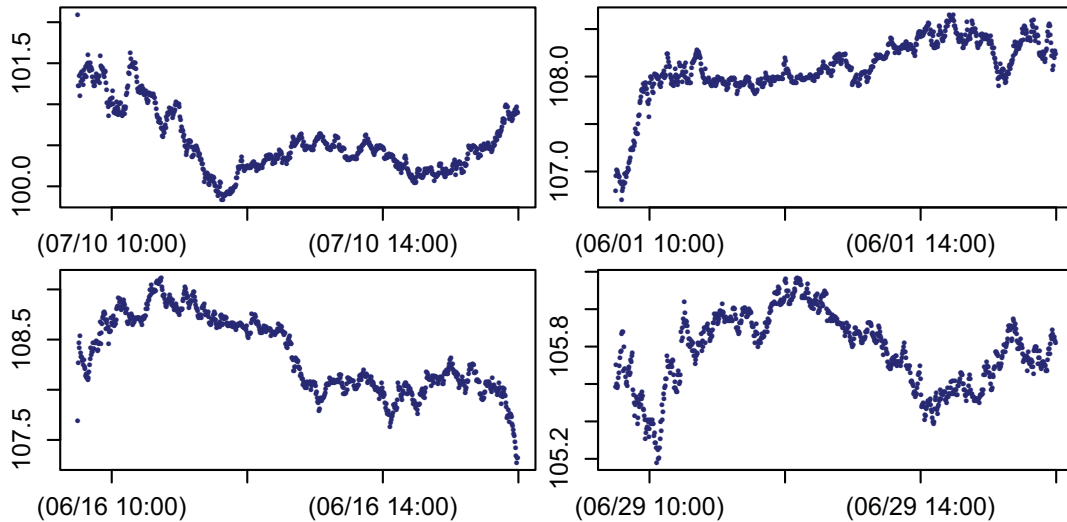


Figure 5.5: Price series for days with a high amount of detected jumps by *ALGO1* (bottom) and *ALGO2* (top) for IBM.

this value decreases slightly for *ALGO2* and heavily for *ALGO1*, ending by a mean value of 10% for *ALGO1* and 25% for *ALGO2*.

The number of detected jumps per day differs heavily between both algorithms. *ALGO1* mainly detects days with one or two jumps. Only for the stock of Citigroup, days with three or even more jumps were reported.

The maximum number of jumps per day reported by *ALGO2* is quite high, but somehow stable for increasing confidence levels. This is a surprising result as one often thinks of jumps as quite rare events, unlikely to happen more than once per day. Nevertheless, the results are in line with the findings of Ane and Metais (2010) who detected up to three jumps per day for European stock indices with a confidence level of (up to) 99.9%. The values for the maximum number of jumps per day, obtained via *ALGO1*, are also stable, but mostly do not exhibit a value of one or two jumps per day, except for the stock of IBM with a striking value of six. Notably, both algorithms report two days with eleven jumps for IBM at $\alpha = .95$, but further investigation shows that these values were not obtained for the same days. Figure 5.5 displays these four days. All price series obtain outliers at the beginning of the day as well as gradual jumps covering a time span of more than 15 minutes. As these gradual jumps do not increase constantly, the algorithms still split these jumps up into several non-consecutive parts.

Despite these facts, the results of both algorithms seem to be stable in the sense of detecting days with multiple jumps constantly for increasing confidence levels.

A valuable by-product of both algorithm is the intraday pattern in jump occurrences as both procedures recorded the intraday intervals in which discontinuities were observed. This pattern is visualised in Figure 5.6 by counting the recorded jump times in 30 minute

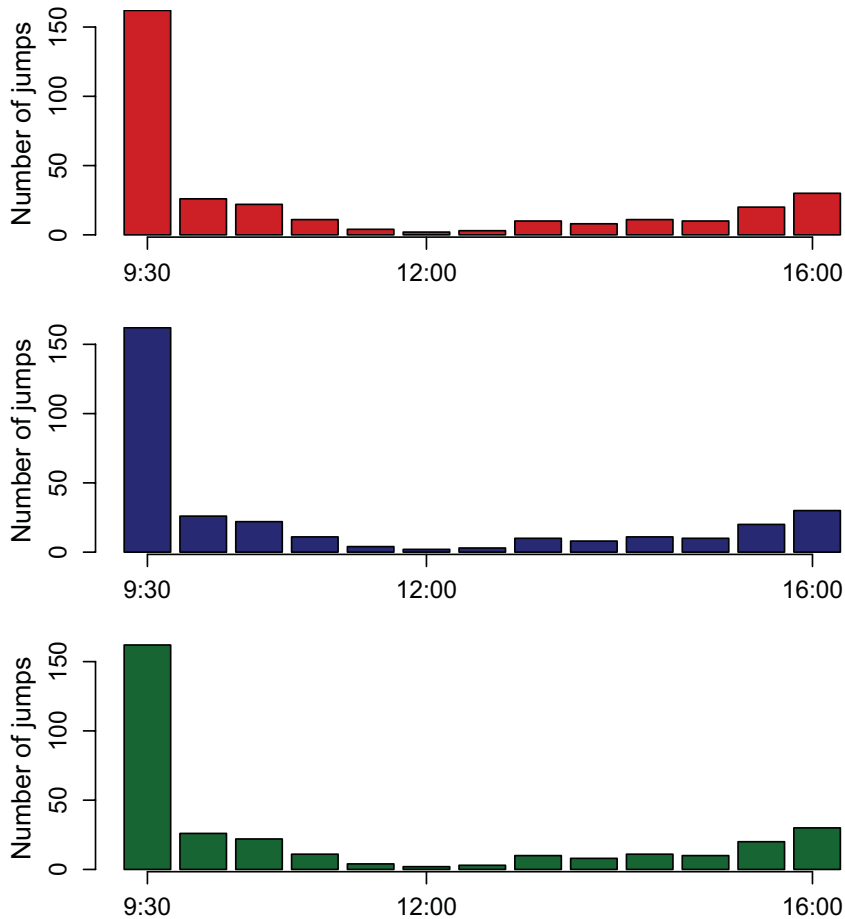


Figure 5.6: Recorded jump times in 30 minutes intervals reported by ALGO2 for stocks of Alcoa Inc., Citigroup and IBM.

intervals starting at 9:30 and ending at 16:00.

Conspicuously, the majority of jumps were recorded during the first 30 minutes of trading. The results of *ALGO1* revealed the same L-shaped pattern, which is also constant for varying confidence levels. This L-shape can be explained by "the pressure at a market opening after a long period of interrupted trading and the accompanying accumulated information" (Ane and Metais (2010)) and therefore is in line with the understanding of a jump as the incorporation of additional information into the price process.

5.3 Jump size

After the jump intensity has been studied, it is logical to turn attention to the second component of the discontinuity: the jump size.

Table 5.4 summarises the typical descriptive statistics for the estimates of positive and negative jump sizes, measured by *ALGO1* and *ALGO2*, respectively. Interestingly, the

	ALGO2			ALGO1		
	AA	C	IBM	AA	C	IBM
	<i>positive jump size</i>					
Mean	0.0006	0.0024	0.0001	0.0010	0.0037	0.0002
St.Dev.	0.0015	0.0135	0.0004	0.0029	0.0167	0.0007
Skewness	5.4635	11.2478	9.2298	7.2169	11.5494	6.5977
Kurtosis	35.9646	134.7586	97.9327	64.7832	151.0190	53.0401
	<i>negative jump size</i>					
Mean	-0.0005	-0.0016	-0.0001	-0.0010	-0.0021	-0.0001
St.Dev.	0.0013	0.0040	0.0003	0.0027	0.0060	0.0004
Skewness	-4.7428	-5.9218	-6.7956	-7.8963	-6.8540	-8.3598
Kurtosis	27.57724	46.0858	55.3789	79.5923	59.8513	87.4967

Table 5.4: Descriptive statistics for positive and negative jump sizes for the stocks of Alcoa Inc., Citigroup and IBM from January 2008 - July 2009 based on $\alpha = .95$ for 5 minutes logreturns.

results do not differ much between the two algorithms. In the end, both algorithms detect that about 30 – 45% of the logreturn variation on days with a significant jump test statistic is attributed to jumps on average. These values are slightly lower than the reported 44 – 57% by Ane and Metais (2010) and a bit higher than the 33% published by Andersen et al. (2007b). To focus the findings again, on days with a significant test statistic, about 40% of the logreturn variation is attributed to jumps on average. This highlights the main importance and the impact of jumps to volatility.

Focusing on the estimates for skewness and kurtosis one has to reject the assumption of normally distributed jump sizes, as parametric literature often assumes. The Jarque-Bera test statistic (not shown in the table) rejects the normality of jump sizes for all six jump size series.

Finally, and as the main goal of this work, Figures 5.7 and 5.8 present a jump free volatility estimate, obtained using *ALGO1* and *ALGO2*.

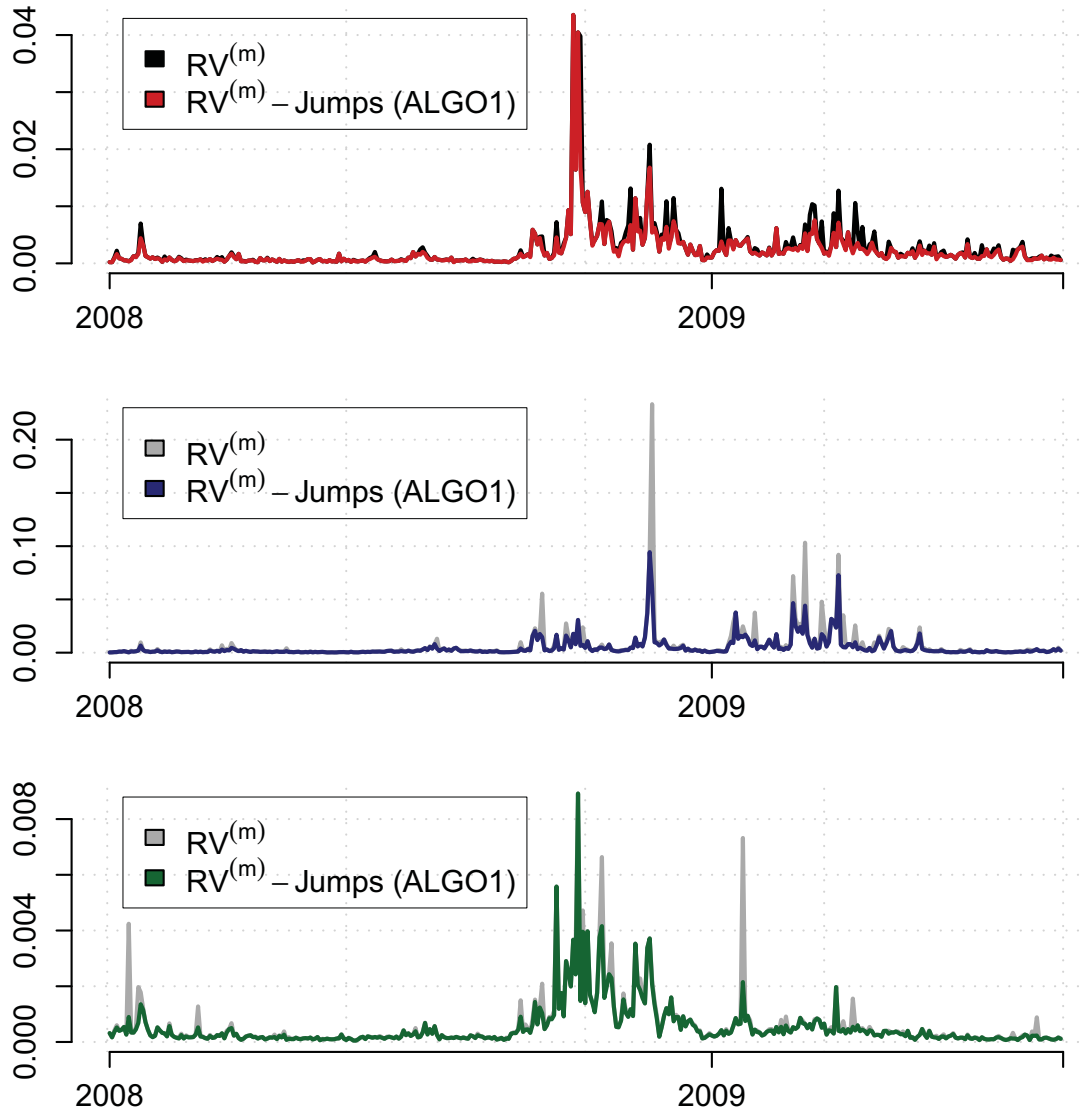


Figure 5.7: Volatility estimate minus estimated jump size using *ALGO1* for the stocks of Alcoa Inc. (red), Citigroup (blue) and IBM (green) from January 2008 – July 2009 based on 5 minute logreturns.

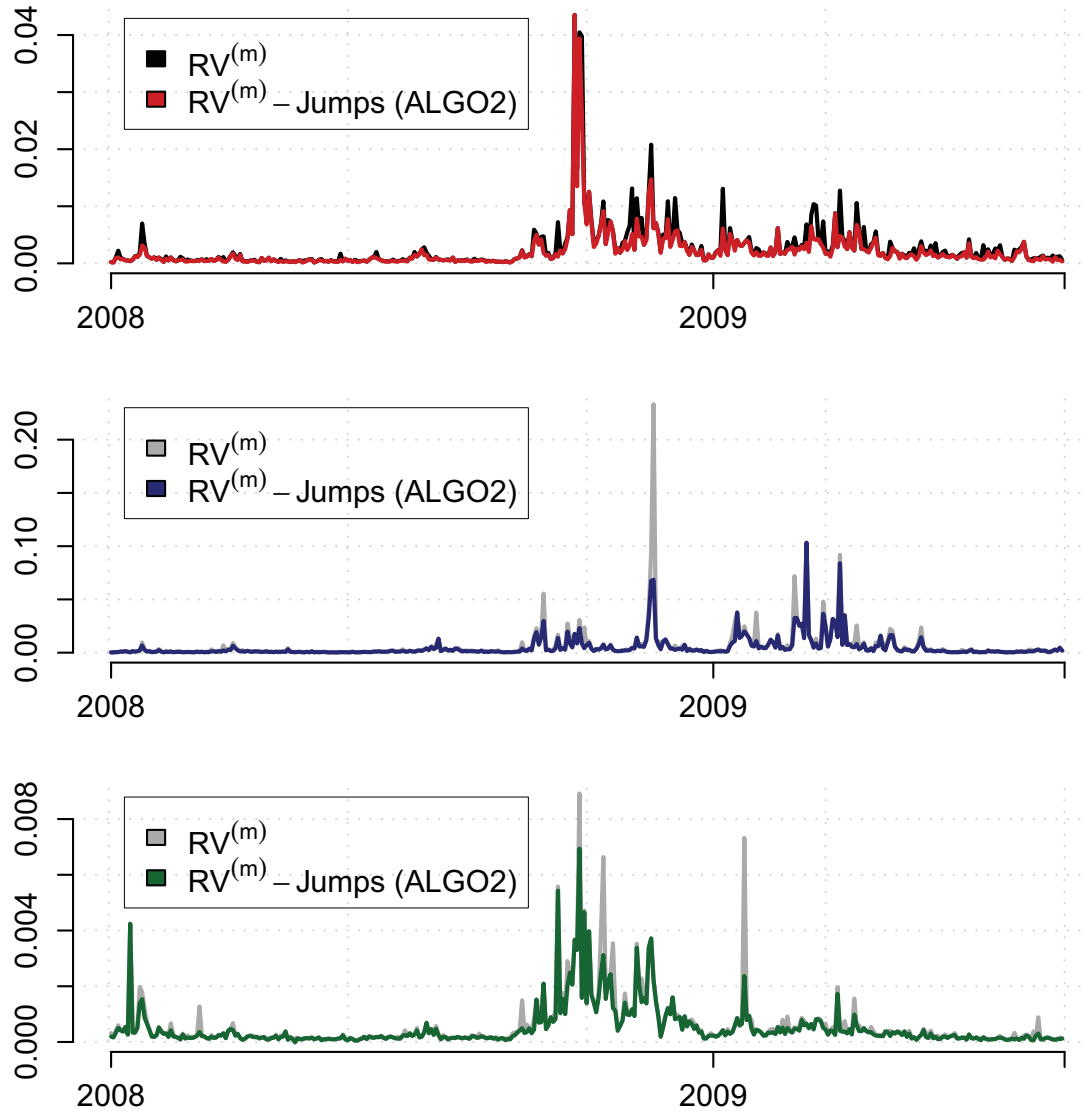


Figure 5.8: Volatility estimate minus estimated jump size using *ALGO2* for the stocks of Alcoa Inc. (red), Citigroup (blue) and IBM (green) from January 2008 – July 2009 based on 5 minute logreturns.

6 Simulation study

In this chapter the experimental design of the simulation used to study the estimators discussed above is described. In the first Section the data-generation process is defined, whereas Section 6.2 contains more information about the implemented procedure. In Section 6.3 the results of the simulation study are discussed.

6.1 Data-generating price process with jumps

Empirical studies have shown that an assets logreturn distribution is non-Gaussian. In contradiction to Gaussian distributed asset logreturns, observed log returns are characterised by heavy tails and high peaks (see Andersen et al. (2002)). Referring to Bouchaud et al. (2001), there is also empirical evidence that suggests that future volatility and past returns are negatively correlated. This fact becomes known as the *leverage effect*. It is therefore appropriate to account for these effects in a simulation. In this simulation study the logprice process of an asset is modelled using the Heston stochastic volatility model. In this model, conceived by Heston (1993), it is assumed that the variance is a random process, which

1. exhibits a tendency to revert towards a long-term mean at a given rate,
2. possesses a volatility proportional to the square root of its level,
3. and whose source of randomness is correlated with the randomness of the underlying price process.

It is therefore possible to generate a price process which comes very close to observed price processes. The Heston stochastic volatility model is given by the following system of stochastic differential equations

$$dS_t = \mu S_t dt + \sqrt{V_t} dW_t^1 + k_t dq_t, \quad (6.1)$$

$$dV_t = \kappa(\theta - V_t)dt + \sigma\sqrt{V_t}dW_t^2, \quad (6.2)$$

$$dW_t^1 dW_t^2 = \rho dt, \quad (6.3)$$

where $\{S_t\}_{t \geq 0}$ and $\{V_t\}_{t \geq 0}$ are the price and volatility processes, respectively. The pure jump process can be broken down into its counting process $\{k_t\}_{t \geq 0}$ and the jump size

process $\{q_t\}_{t \geq 0}$. μ denotes the rate of return of the asset. To take into account leverage effect, $\{W_t^1\}_{t \geq 0}$ and $\{W_t^2\}_{t \geq 0}$ are correlated Brownian motion processes with correlation parameter $\rho \neq 0$. $\{V_t\}_{t \geq 0}$ is known as a square root mean reverting process with long run average price variance θ and rate of reversion κ , where σ is referred to as the volatility of volatility, i.e. the variance of $\{V_t\}_{t \geq 0}$. According to Mikhailov and Noegel (2003) the variance of the square-root process in Equation (6.2) is always positive and if $2\kappa\theta > \sigma^2$ then it cannot reach zero. The authors also mention that the deterministic part of this process is asymptotically stable if $\kappa > 0$.

The model in Equation (6.1) was simulated with and without jumps in order to be able to compare the false detection rate of days with no jumps. The jumps were modelled in two ways. In a first run, mathematical jumps were implemented with a starting time, $s_t \sim U_{[300,23100]} \forall t \geq 0$, representing a time point between 9:35 and 15:55. The jump size corresponds to a 3 – 5% change of the assets value with a randomly determined sign. In the second run gradual jumps were implemented. The jump length was modelled by a shifted binomial distribution, according to the simulations of Kloessner (2010), taking values from one minute to 12 minutes with a starting time, $s_t \sim U_{[300,22680]}$, representing a time point between 9:35 and 15:48 to ensure the perceptibility of the full gradual jump. The jump size is comparable to the one used for the mathematical jumps and the sign of the jump was also determined randomly.

Additionally, days with more than one jump were simulated. In either case the detectability of each jump was ensured by prohibiting jumps covering the same time span.

6.2 Simulation details

Monte Carlo simulation was performed by discretising the stochastic process using the Euler method with an increment of one second per tick. The standard unit in this simulation is one trading day throughout. This resulted in

$$V_t = V_{t-1} + k \cdot (\theta - |V_{t-1}|)dt + \sigma \cdot \sqrt{V_{t-1}} \cdot \sqrt{dt} \cdot Z_t^1, \quad (6.4)$$

$$S_t = S_{t-1} - r \cdot S_{t-1}dt + \sqrt{|V_{t-1}|} \cdot S_{t-1} \sqrt{dt} \cdot Z_t^2 + S_{t-1} \cdot J_t, \quad (6.5)$$

where $|V_{t-1}|$ denotes the absolute value of V_{t-1} to avoid negative variances and J is a vector containing the jump sizes. This vector contains either zeros, representing no jump at that time, or a value corresponding to an increase (decrease) of 3 – 5% for mathematical jumps and 0.009 – 0.013% for gradual jumps. This is due to the fact that the gradual jump affects 60 – 720 subsequent values of S_t . Further, $\{Z_t^1\}_{t \geq 0}$ and $\{Z_t^2\}_{t \geq 0}$ are standard normal random variables with correlation factor ρ . To implement

Variable	Value	Variable	Value
S_0	100	V_0	.001
μ	0	σ	.001
r	0	θ	.001
κ	2	ρ	-0.620

Table 6.1: Experimental design for the simulation study.

the correlation factor ρ the Cholesky decomposition was used

$$Z_t^1 = \phi_t^1, \quad (6.6)$$

$$Z_t^2 = \rho\phi_t^1 + \sqrt{1 - \rho^2}\phi_t^2, \quad (6.7)$$

where $\{\phi_t^1\}_{t \geq 0}$ and $\{\phi_t^2\}_{t \geq 0}$ denoting independent standard random normal variables. Table 6.1 summarises the chosen values for Equations (6.1) and (6.2).

6.3 Results

To see the effect of mathematical and gradual jumps on the detection rate of the test statistics the confusion matrix is presented. The matrix consists of four cells: the upper left cell is the proportion of the test statistic smaller than the 95% standard normal critical value among days without jumps, the upper right cell is the proportion greater than the 95% critical value among the days with no jumps, the lower left cell is the proportion smaller than the 95% critical value among the days that jumps occur on, and the lower right cell is the proportion greater than the 95% critical value among the jump days (see Huang and Tauchen (2005)).

Therefore the off-diagonal elements represent the proportion of days when the statistic signals a wrong answer, whereas the diagonal elements represent the ability of the test statistics to identify correctly whether or not there is a jump on a given day. Hence, the row sums of the matrix are equal to one. According to the asymptotic results presented in Section 3.2, one can expect that the value in the upper left cell is close to 0.95 for sufficient values of m , i.e. for $m \rightarrow \infty$.

Table 6.2 shows the results based on five minutes logreturns of 1,000 simulated days. It can be seen that the test statistic used in *ALGO1* holds the suspected confidence level of 95% in contradiction to the test statistic implemented in *ALGO2*, which has an incredibly huge false detection rate. This result holds for mathematical as well as for gradual jumps. Secondly, days with a mathematical jump were quite well detected by both algorithms. Only 18.5% of the days were not detected by *ALGO1* and only

		mathem. jump		gradual jump	
		(NJ)	(J)	(NJ)	(J)
ALGO1	(NJ)	.965	.035	.947	.053
	(J)	.185	.815	.747	.253
ALGO2	(NJ)	.671	.329	.579	.421
	(J)	.149	.851	.093	.907

Table 6.2: Confusion matrix based on 5 minutes logreturns of 1,000 simulated days.

14.9% by *ALGO2*, respectively. On the other hand, only 25.3% of days with a gradual jump were detected by *ALGO1*. In contradiction to this result stands the detection rate of *ALGO2*, which performs much better by detecting 90.7% of the days containing a gradual jump. This supports the suspicion that the test statistic z_{QPLM} is unable to detect gradual jumps whereas test statistics based on intradaily highs and lows are able to find these type of jumps.

Another question is if both algorithms are able to identify the exact number of jumps per day. Therefore, each 1,000 days with one mathematical and one gradual jump (MG), two mathematical jumps (MM) and two gradual jumps (GG) were simulated. Table 6.3 reports the amount of days with no jumps (0), one jump (1J), two jumps (2J), three jumps (3J) and more than three jumps ($> 3J$) identified by each algorithm.

As already suspected, *ALGO1* is unable to detect both jumps in the data under the (MG) set up. Nevertheless, one jump is identified with high precision. This shows that the test statistic performs well in detecting mathematical jumps even if gradual jumps occur on the same day. *ALGO2*, on the other side, detects two jumps on about 47% of the days but also over estimates the number of jumps per day significantly. Interestingly, qualitatively the same also holds true for days with two gradual jumps. *ALGO1* also performs poorly in this setup. only 40.1% of the days with two gradual jumps were identified as days with at least one jump and only 1.3% of the days correctly. On the other hand, days with two mathematical jumps were reliable identified by *ALGO1*. Each day in this simulation run was identified as a day with at least one jump and only 22.3% of the days were misidentified in the sense that only one jump was found. Again, *ALGO2* performs worse as only 55.9% of the days were identified correctly and about 34% of the days were marked as days with more than two jumps.

In a nutshell, the simulation has shown the following important results:

1. *ALGO2* is unable to hold the suspected confidence level,
2. *ALGO2* is able to detect gradual jumps, but overestimates their number on a given day,

Algorithm	Number of detected Jumps	Simulation setup		
		MG	MM	GG
ALGO1	(0)	.018	.000	.599
	(1J)	.914	.223	.388
	(2J)	.068	.776	.013
	(3J)	.000	.001	.000
	(> 3J)	.000	.000	.000
ALGO2	(0)	.010	.005	.022
	(1J)	.125	.098	.188
	(2J)	.474	.559	.426
	(3J)	.208	.177	.185
	(> 3J)	.183	.161	.179

Table 6.3: Detection rate, based on 5 minute logreturns, for each 1,000 simulated days with one mathematical and one gradual jump (MG), two mathematical jumps (MM) or two gradual jumps (GG).

3. *ALGO1* is unable to identify days with at least one gradual jump correct, due to its inability to detect gradual jumps,
4. *ALGO1* detects the right amount of mathematical jumps on a given day with a high guarantee.

The results presented above were tested for their stability by using different values for $\sigma = \theta = \{0.001, 0.005, 0.0001\}$ and $\alpha = \{0.95, 0.99\}$ as well as for sampling frequencies ranging from 5 – 30 minutes. In all cases the findings are qualitatively the same and the main conclusions remain unchanged.

7 Conclusion

Based on the theory developed by Kloessner (2009), which builds on the idea of using intradaily highs and lows, this work presents a sequential algorithm to determine the number of jumps per day, the sign of the jump, the time at which the jump occurs and the corresponding jump size. The main idea of the algorithm was first presented by Andersen et al. (2007b) and later on in a slightly different version published by Ane and Metais (2010). In their version, the algorithm, which is referred to as *ALGO1* in this work, is based on the classic approach where the key element is the difference between two estimators for the volatility of the price process: the realised volatility, $RV^{(m)}$, first introduced in Andersen et al. (2001), and the bipower variation, $BPV^{(m)}$, developed by Barndorff-Nielsen and Shephard (2004). As $RV^{(m)}$ approximates the quadratic variation of the price series for higher sampling frequencies whereas $BPV^{(m)}$ only approximates the continuous-time component, it is possible to identify jumps by a significant difference between these two estimators.

In order to detect whether or not a jump occurred on a given day, the test statistic z_{QPLM} , first presented by Barndorff-Nielsen and Shephard (2006), was used as the simulations of Huang and Tauchen (2005) showed the good performance of this estimator. Nevertheless, as discussed in Kloessner (2010), this test statistic is unable to detect gradual jumps, a kind of jump which was somehow ignored up to now. In contrast, the test statistics presented by Kloessner (2010) are able to detect these kind of jumps and, furthermore, allow for positive and negative jumps to be separately tested.

Therefore, *ALGO1* was applied only with minor changes, i.e. with an additional procedure to detect split gradual jumps, and in order to measure the impact of gradual jumps, the same algorithm was adapted to the theory of Kloessner (2009), later on called *ALGO2*.

Both algorithms were applied to a high frequency dataset containing TAQ for stocks of Alcoa Inc., Citigroup and IBM from January 2008 until July 2009, covering the time span of the financial crisis in October 2008. All three stocks show the impact of the financial crisis, resulting in incredible high estimates of realised volatility for specific days.

The findings for both algorithms correspond with recent results from Ane and Metais (2010) and Andersen et al. (2007b) by identifying days with more than one jump. For all three stocks *ALGO1* mainly reports days with one or two jumps, whereas *ALGO2*

finds significantly more jumps (frequently four to six) per day. This difference indicates the influence of gradual jumps for days with a significant test statistic and is therefore an indicator of the importance of this kind of jump.

Consequently, *ALGO2* detects a higher jump intensity of 0.99, 0.83 and 0.53 for the stocks of Alcoa Inc, Citigroup and IBM, respectively. In comparison, *ALGO1* reports values of 0.44, 0.62 and 0.41. In both cases, positive and negative jumps are recorded nearly equally often.

Jumps occur mainly at the beginning of the trading day, uncovering the typical L-shape for US data. These findings are in line with the work of Ane and Metais (2010) and the main understanding of jumps as the result of news, which have to be incorporated into the current asset price.

Both algorithms detect that about 30 – 45% of the logreturn variation is attributed to jumps on average. The results of the empirical analysis highlight the importance and the impact of jumps on daily estimates of volatility.

As the two algorithms rely on the confidence level α and the chosen sampling frequency, the impact of different values for these quantities was checked in order to see how robust the results were. The results of *ALGO1* are not very robust to the chosen confidence level α , as the number of days with a jump as well as the number of jumps per day decreases drastically for higher values of α . The results of *ALGO2* seem to be stable in the sense of showing the same behaviour but with a much smaller decrease. As long as the sampling frequency is not chosen too high, i.e. is not exhibiting five minutes, the results of both algorithms are stable too.

In order to verify these results a simulation study was realised. Using the model presented by Heston (1993), 4,000 days were simulated with different amounts of mathematical and gradual jumps. A confusion matrix revealed that the test statistic of *ALGO2* is unable to hold the suspected confidence level in this simulation setup. It was also shown that, using this setup, z_{QPLM} is unable to detect gradual jumps whereas the test statistic used in *ALGO2* is truly able to detect this kind of jump.

Additionally, the simulation study revealed that *ALGO1* is also able to detect two mathematical jumps with a high guarantee, whereas *ALGO2* seems to overestimate the amount of jumps per day.

A good starting point for the latter would be the investigation of the resulting series of jumps as well as a multivariate analysis. An improvement of the asymptotic theory for the test statistic based on intradaily highs and lows would also be of value.

Bibliography

- Ait-Sahalia, Y. and Jacod, J. (2009). Testing for jumps in a discretely observed process. *Annals of Statistics*, 37:184–222.
- Andersen, T. G., Benzoni, L., and Lund, J. (2002). An empirical investigation of continuous-time equity return models. *Journal of Finance*, 57:1239–1284.
- Andersen, T. G., Bollerslev, T., and Diebold, F. (2007a). Roughing it up: Including jump components in the measurement, modelling, and forecasting of return volatility. *The Review of Economics and Statistics*, 89:701–720.
- Andersen, T. G., Bollerslev, T., Diebold, F., and Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Econometrics*, 61:43–76.
- Andersen, T. G., Bollerslev, T., Diebold, F., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71:579–625.
- Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2000). Great realizations. *Risk*, 13:105–108.
- Andersen, T. G., Bollerslev, T., Frederiksen, P., and Nielsen, M. (2007b). Continuous-time models, realized volatilities, and testable distributional implications for daily stock returns. Technical report, Department of Economics, University of Aarhus.
- Andersen, T. G., Dobrev, D., and Schaumburg, E. (2010). Jump-robust volatility estimation using nearest neighbor truncation, working paper: Staff report no. 465.
- Ane, T. and Metais, C. (2010). Jump distribution characteristics: Evidence from european stock markets. *International Journal of Business and Economics*, Vol. 9, Nr. 1:1–22.
- Bandi, F. M. and Russel, J. R. (2006). Separating microstructure noise from volatility. *Journal of Financial Economics*, 79:655–695.
- Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2004). Regular and modified kernel-based estimators of integrated variance: The case with independent noise, discussion paper. Technical report.

Bibliography

- Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008a). Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and nonsynchronous trading. unpublished paper. Technical report, Oxford-Man Institute, University of Oxford.
- Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008b). Realized kernels in practice: Trades and quotes. *Econometrics Journal*, volume 4.
- Barndorff-Nielsen, O. E. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society*, Vol. 64, No. 2:253–280.
- Barndorff-Nielsen, O. E. and Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2:1–37.
- Barndorff-Nielsen, O. E. and Shephard, N. (2006). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics*, 4:1–30.
- Bouchaud, J.-P., Matacz, A., and Potters, M. (2001). Leverage effect in financial markets: The retarded volatility model. *Physical Review Letters*, Vol. 87:228–240.
- Brownless, C. T. and Gallo, G. M. (2006). Financial econometric analysis of ultra-high frequency: Data handling concerns. *Computational Statistics & Data Analysis*, 51:2232–2245.
- Hansen, P. and Lunde, A. (2006). Realized variance and market microstructure noise. *Journal of Business & Economic Statistics*, Vol. 24 (2):127–161.
- Hasbrouck, J. (2004). Empirical market microstructure: Economic and statistical perspectives on the dynamics of trade in securities markets. lecture notes. Technical report, Stern School of Business, New York University.
- Heston, S. (1993). A closed-form solutions for options with stochastic volatility. *Review of Financial Studies*, 6:327–343.
- Huang, X. and Tauchen, G. (2005). The relative contribution of jumps to total price variation. *The Journal of Financial Econometrics*, 3:456–499.
- Kloessner, S. (2009). Separating risk due to diffusion, positive jumps, and negative jumps. working paper. Technical report, Saarland University.
- Kloessner, S. (2010). Grasping economic jumps by sparse sampling using intradaily highs and lows. Technical report, Saarland University.

- Lee, T. and Ploberger, W. (2009). Rate-optimal tests for jumps in diffusion processes. working paper. Technical report, University of Rochester, University in St. Louis.
- Mikhailov, S. and Noegel, U. (2003). Hestons stochastic volatility: Model implementation, calibration and some extensions. *Wilmott*, July:74–79.
- Shiryaev, A. N. (1999). *Essentials of Stochastic Finance: Facts, Models, Theory*, volume Volume 3 of *Advanced Series on Statistical Science & Applied Probability*. World Scientific Publishing Company, 1st edition (april 15, 1999) edition.
- Zhang, L., Mykland, P. A., and Ait-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, Vol.100:1394–1411.

Declaration of Authorship

I hereby confirm that I have authored this masters thesis independently and without use of others than the indicated sources. Where I have consulted the published work of others, in any form (e.g. ideas, equations, figures, text, tables), this is always explicitly attributed.

Berlin, 10 June 2011

Martin Schelisch

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Berlin, den 10. Juni 2011

Martin Schelisch