On the tails of certain speculative prices

Author:
Thijs van der Valk

Supervisor:
Prof. Dr. Dick van Dijk

ERASMUS UNIVERSITEIT ROTTERDAM

Transtrend

February 9, 2011

Abstract

This thesis contains an empirical review (financial case study) of two high frequency stock return distributions and in particular an analysis of their tail-behavior under temporal aggregation. In line with a long econometrical tradition starting with [Mandelbrot, 1963], I study the unconditional return distribution of (stock) returns. I show that the (ultra) high frequency returns of Dutch Koninklijke KPN N.V. (ticker: KPN NA) and Hong Kong's Pccw Limited (ticker: 0008 HK) show no signs of stable distributions, which is in line with [Lau et al., 1990] but in contrast with [Pictet et al., 1996]. Using simulation, I point out that market microstructure characteristics, i.e. tick size and (il)liquidity, largely explain tail behavior of ultra high frequency under temporal aggregation and show that the tail estimators kurtosis and tail index may give contra-dictionary results. Handling ultra high frequency data, or data in general, has many pitfalls. In the thesis I discuss these pitfalls and describe methods to get around, such as tick data filtering. Overall, this thesis is an advocate for thorough data analysis and econometrical research with a clear focus on practicality.

Acknowledgement

I would like to thank anyone who contributes to making (my) life interesting and worthwhile. No, really, I appreciate the effort.

Contents

Chapter 1

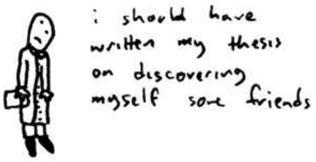
Introduction

1.1 Research Objective

The goal of this thesis is to provide a financial case study of two specific high frequency stock return distributions[1]. The stocks, KPN and Pccw, are comparable as they are both publicly held telecommunication companies, traded on developed market places, Euronext Amsterdam and Hong Kong Stock Exchange. However, both markets work differently which becomes apparent when looked at their return distributions. I thoroughly investigate all aspects of the high frequency return distributions, including the data quality and stylized facts.

More specifically, I will focus on their tail behavior under temporal aggregation. The importance of tail behavior, and other aspects of return distributions, under temporal aggregation lies in the heart of the hypothesis of scalable markets. With this hypothesis, one assumes that financial markets behave similarly on different time scales. Proper examination of tail behavior under temporal aggregation is needed to reject or accept this idea.

I will measure the tails according to standard kurtosis and the tail index using the Hill estimator and using a simple OLS regression method. The *quantitative* versus *qualitative* usability of the estimators will be investigated and I will try to make plausible that illiquidity and tick size largely explain the tails of the high frequency stock return distributions at the highest frequencies (< 1 hour).



[1] See [Eisenhardt, 1989] for a comprehensive paper on the process of inducting theory using case studies. Table 1 in this paper provides a simple framework for building sound case studies.

Relevance High frequency data is essential to financials firms in general and the Commodity Trading Advisor (CTA) Transtrend in particular. This CTA applies a systematic trading strategy that exploits medium term trends[2] in financial markets. Although Transtrend is not known as a high frequency trading specialist, high frequency data is used in many areas of its operation, including but not limited to research and development, daily pricing and algorithmic trading. Other financial firms, like banks and investment advisors, may use high frequency data in similar areas.

As with all data, working with high frequency data knows many challenges. First and foremost, the size of an average high frequency dataset is restrictive for complex modeling on a daily basis. Nevertheless, financial firms like Transtrend, that have a clear focus on innovation, always look for ways they can use new sources of information to improve their business. As such, understanding the full characteristic of high frequency data is of great interest to financial academics and practitioners.

Methodology I choose to make this research particularly narrow, because of both a practical as well as a theoretical reason. The main *practical* reason is that high frequency data is plentiful by nature and by focussing on two specific stocks, I confine the amount of computer power and time needed to handle the data. Naturally, the disadvantage of this narrowing-down is that the amount of *quantitative* statements that can be made about high frequency stock return distributions in general, is limited to these two names. However, the amount and quality of the *qualitative* statements increase with this focus on two names.

The second reason, which is more *theoretical* of nature, is because I agree with John Maynard Keynes[3], who regards economics as a moral science, and not a natural science. According to [Skidelsky, 2009], Keynes acknowledges that "there are areas statistical analysis [econometrics] may be a useful tool, but they are limited to simpler, less abstract, relations." Keynes: "The notion of testing the quantitative influence of factors suggested by a theory as being important is very useful and to the point. The question to be answered, however, is whether the complicated model ... does not result in a false precision beyond what the method ... can support."

My impression of econometrics, at least at the (under)graduate level - at which the foundation for additional study is laid anyway, is that there is little attention for *ideas* behind the models and *implications* of the models. I think economics is a hard subject to catch in numbers, especially because economics expressed as a financial time series is the outcome of the whimsicality of human nature. Citing [Skidelsky, 2009], "Keynes would have said that it is absurd to rely on risk models based on past data at the moment bankers were creating complex new products ever day." I agree. This may sound strange from a researcher working at a systematic, trend following CTA, whose business model it is to use past data to exploit trends in financial markets. However, the fact that an econometrician relies on markets depending on unreliable human beings, does not mean he cannot *try* to explain. He should keep it mathematically simple

[2] Medium term means several weeks.

[3] John Maynard Keynes (1883 - 1946), British economist and revolutionizer of modern economics.

though, and focus on know-why besides know-how. In the same vein, Keynes told Jan Tinbergen[4] in their dispute in writing, that one has to demonstrate first of all that the methods used are applicable, instead of just applying them. See e.g. [Garonne et al., 2004] and [Keuzenkamp, 1995].

In this line of thought, the main objective of this research will be to explain and not to predict.

1.2 This thesis

In order to interpret the high frequency data of the two stocks properly, I will discuss the following subjects throughout this thesis.

High frequency data In the age of computers, information is a blessing and a curse.

The availability of cheap storage and high speed fiber optics has made information plentiful. Google, high speed downloading and unlimited (free) storage of data has become an integral part of our lives.

In financial markets, this is not different. Millions of trades take place every day. Data providers like REUTERS send a constant stream of high frequency data all over the world. For many alternative investment managers, hedge funds or high frequency traders this stream is indispensable to generate their income.

One could argue that high frequency time series, i.e. prices that are recorded more often than daily, reflect a continuous arrival of news. These large datasets are for example, tick-by-tick prices and quotes of the S&P500 stock index or prices of the $/¥-foreign exchange rate sampled every five minutes. They contain however, a lot of erroneous data. This data needs to be filtered to be useable[Falkenberry, 2002].

The question with high frequency data arrises, what is valuable and what is not? What can one do with this tsunami of information? This integral part of good research is most often omitted. Some authors describe simple, ad-hoc filtering rules as a side note ([Muller et al., 1990], [Dacorogna et al., 1993], [Pictet et al., 1996]).

Because I think all research should start at the base, I will elaborately discuss high frequency data, stylized facts, filtering and value in chapter 2.

Fat tails at lower frequencies According to [Bartolomeo, 2007] there are three broad schools of thought in the discussion of fat tails in stock returns.

1. The returns have stable distributions and thus infinite variance.
2. The returns come from specific, identifiable, fat-tailed distributions (Gamma, Student-t, et cetera).

[4] Jan Tinbergen (1903 - 1994), Nobel Prize winner and respectfully regarded as founder of modern econometrics.

3. The returns are conditionally normal at all time, but show fat tailed behavior due to time varying variance and volatility clustering.

The first school of thought starts with [Mandelbrot, 1963] and [Fama, 1965], which will be discussed in chapter 3. The latter finds proof for Mandelbrot's stable Paretian hypothesis in daily US equity returns, but finds volatility clustering as well. [Muller et al., 1990] find empirical evidence of a price change scaling law in intraday foreign exchange rates. See [Rachev and Mittnik, 2000] for a elaborate review of stable distributions.

The second school of thought uses asset pricing and risk free probabilities to derive the theoretical distribution of stock returns. [Bartolomeo, 2007] refers to [Gulko, 1999] who find risk neutral probabilities that are equivalent to returns having a Gamma distribution. The literature on this idea is less comprehensive.

The third school of thought forms the foundation for the rich literature on (G)ARCH models[5]. These models assume volatility clustering which is predictable. See e.g. [Engle, 1982] and [Bollerslev, 1986]. (G)ARCH models are used on high frequency data to obtain volatility (and higher moments) estimates on a lower frequency, like daily. Market microstructure (i.e. bid-ask spreads and and tick sizes) however, heavily influence the usability to estimate intraday volatility and kurtosis.

Research on fat tails is often applied to low frequency data, including daily, weekly and monthly returns. High frequency data is mostly used to estimate volatility at lower frequencies, using measures as realized volatility. Some literature on lower frequency can be used to research higher frequency data, although one should recognize that the (trading) agents that act on one time scale are different in nature from those that act at another time scale. Whether this leads to the rejection of scalability of financial markets, is to be examined.

Stable distributions There are many fat-tailed distributions that have been fitted to financial time series, such as log-normal, Student-T, Cauchy and mixture distributions. These distributions often have the advantage that they are well understood and as there exist analytical expressions for their *pdf*s, they are easy to estimate.

A less widely used distribution is the stable distribution. Stable distributions have the attractive property that the sum of n copies of independent, identically distributed random variables is identically distributed as well. So, in a financial context this means that in considering asset returns, it does not depend which time scale one is looking at, i.e. hourly, daily or monthly returns will all have the same distribution.

In fact, stable distributions are the only distributions (with the normal distribution as a special case of the stable distribution) that have this property. This property forms the heart of a generalized Central Limit Theorem (CLT): the distribution of the sum a large number of independent identically distributed random variables belongs to a family of distributions known as stable distributions. This is the main reason that price returns should have scale-invariant properties in the first place.

[5] (G)ARCH, (Generalized) Autoregressive Conditional Heteroskedasticity

The relative unpopularity of stable distributions is due to two major drawbacks. Firstly, the variance and higher moments do no exist for stable distributions[6], and secondly, there is generally no analytical expression for the pdf available. See chapter 3 or [Rachev and Mittnik, 2000].

The reason I will look at the stable distribution is because it offers an interesting fat tailed alternative for the normal distribution. The stable distribution is a so called *power law* distribution. Because of the research objective of this thesis (which is to explain), the use of the stable distribution suffers no drawback from the lack of finite variance and lack of analytical pdf. Stable distributions provide a very general approach and give a lot of insight into the tails of distributions. The abundant availability of computational power has made the lack of analytical expression of the pdf less important. The goal is to test the tails and stability of high frequency returns. For this purpose, the stable distribution is very well suited.

I will discuss stable distributions in more detail in chapter 3.

Organization This thesis is structured as follows.

First I will look at high frequency data in chapter 2. I will discuss the stylized facts and show empirically what the influence of outliers and unfiltered data is. I will embed my empirical findings in the literature. As the data is the timber of the house I am building, this data analysis will be substantial.

In chapter 3, I will give an overview of stable distributions and tail indices, both from a theoretical a well as an empirical point of view. This part is needed to obtain a better understanding of stability and tail behavior. I will discuss the literature on (unconditional) stable return distributions of (stock) returns, starting with [Mandelbrot, 1963] and [Fama, 1965].

In chapter 4, I will report my findings on the tail behavior of the high frequency data. This part will contain an in depth analysis on a individual level for the two stocks KPN and Pccw. In chapter 5 I will describe a simulation of high frequency data that captures the findings of chapter 4 on tail behavior under temporal aggregation.

Finally, I will conclude in chapter 6.

[6] This does not count for the limiting case of the normal distribution

Chapter 2

High frequency data

Since the age of computers, larger and larger datasets have become available. In finance, this has lead to large datasets of prices recorded more often than daily. This what we call *high frequency data*.

High frequency data covers everything between prices recorded tick-by-tick and daily prices. Most often seen are tick-by-tick prices, prices recorded over five minutes intervals and hourly prices.

As mentioned in chapter 1 these datasets contain a lot of information, of value for both academics and practitioners. However, there are many challenges associated (in handling the information) as well.

The most important challenge is that high frequency datasets are not as freely available as daily datasets. The problem is *budget*. Well known and most studied datasets have been provided by Olsen & Associates (1995, foreign exchange data) and the Trade and Quotation database which consists of quotes and prices traded on NYSE, AMEX (now part of NYSE) and NASDAQ. The first dataset is freely accessible. Other datasets however are very costly and most often only collected by commercial investment firms, as opposed to universities and research groups. In this research I will use in house TRANSTREND data. It contains both spot and futures prices of almost all commodities, stocks, fixed income and foreign exchange markets from all over the world.

2.1 Value of high frequency data

In general, high frequency data is valuable at any stage of the investment process. Some examples of how high frequency data can be used, are given below.

1. High frequency data as primary source of price information. If you would follow the adage ‘the market is always and ultimately right’, collecting as much information about market prices as sensible, is of extreme importance. For high frequency traders, collecting as much prices as possible is viable, since the more they know, the quicker they can react, which ultimately will give them the edge over competition.

2. High frequency is not solely of interest for high frequency trading strategies. Even for lower frequency investments, looking at higher frequencies may give a better understanding (i.e. pricing) of the market at the desired frequency. A clear example of how higher frequencies may influence pricing at a lower frequency, is shown in figure 2.1. In this figure, we see that some outliers at the end of the day, provide a wrong daily price. Next to that, the volatility estimate of that day will be wrongly biased by those outliers.

3. High frequency trading as a source of secondary information. For instance, the data can be used for estimation of daily volatility and daily correlation. It can even be used for the estimation of trading costs, i.e. slippage, which depends on bid-ask spreads and intraday liquidity.

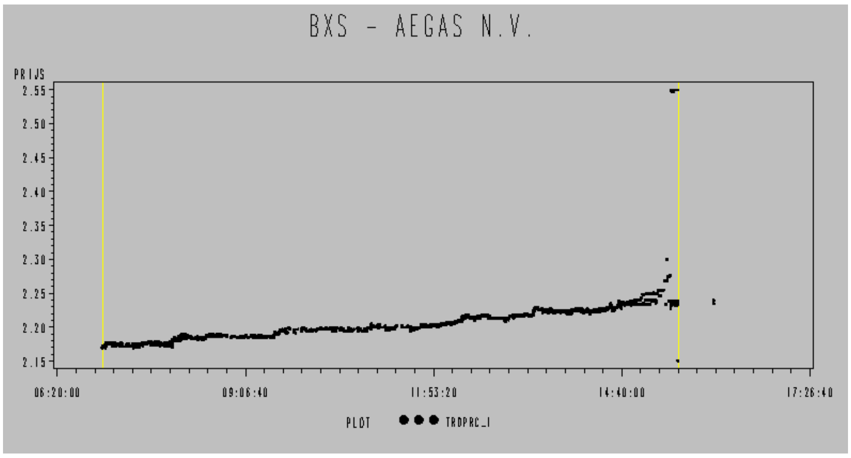


Figure 2.1: Traded prices of the common stock of Ageas (former Fortis), traded on Oct. 13 2010 on Euronext Bruxelles. Both high and low of the day are clearly influenced by some severe outliers at the end of the day. A look a this figure reveals a lot of valuable insight in the behavior of this stock on that day: the *real* volatility is a lot lower than the *observed* (uncorrected) volatility. Daily data provided by Reuters and Bloomberg may have been false in this particular case.

These examples are far from complete. In general, high frequency data is important for anything from data collection to making investment decisions, trading, portfolio construction, estimation of executing costs, et cetera.

2.2 Stylized facts

The stylized facts of high frequency returns are similar to, but distinct from daily returns, see [Taylor, 2005]. Intraday returns are defined equivalently to daily returns, i.e. the logarithm of the price change during some interval (or from trade to trade). The stylized facts of daily returns are as follows.

1. *The distribution of returns is not normal.* It is approximately symmetric; it has fat tails and is high peaked.
2. *There is almost no correlation between returns.* This is regardless of the lag between the returns.
3. *There is positive first order dependence between absolute returns.* This is valid for squared returns as well.

These stylized facts cover (low frequency) financial returns and are mostly undisputed. For high frequency data they seem to be valid as well.

Throughout this thesis I use high frequency data of two stocks: Dutch KONINKLIJKE KPN N.V and PCCW LIMITED from Hong Kong.

KPN is a Dutch telecommunication company. Its main listing is on EURONEXT AMSTERDAM (Reuters Identification Code (RIC): KPN.AS) and has a market capitalization of 2.5b USD. Due to European regulation (MIFID), KPN is traded on many other exchanges, multilateral trading facilities (MTF) and other (dark) trading pools. To grab the full market, I use a composite data stream to capture all trades on the primary exchange and all alternative venues. The RIC for this stream is KPNEUR.xbo.

PCCW is a telecommunication company as well. Its main listing (and only listing besides a US ADR) is on the HONG KONG STOCK EXCHANGE (RIC: 0008.HK). PCCW has a market capitalization of 2.5b USD.

The data for botch stocks runs from January 1, 2010 through June 30, 2010. I will use prices on business days between the official opening hour and closing hour, and where applicable, I respect trading breaks. Overnight returns are omitted. I do not use after hours data.

Fat tails Intraday returns are leptokurtic. This means they have fat tails and high peaks. This is shown for KPN in figure 2.2. In this figure we see four histograms for four different measurement intervals. With increasing interval (decreasing frequency) we see a return distribution that gets less discrete. This is because tick size gets less important with decreasing frequency.

Figure 2.3 shows the intraday returns for PCCW. The discrete steps in all histograms arise as a result of discrete price changes due to fixed tick sizes. The high peak at 0 for the highest frequencies is mainly due to the fact that the prices are more often recorded than there are trades in the particular intervals. This is a form of illiquidity.

Figure 2.2 and 2.3 nicely show that two apparently comparable telecom stocks, show completely different return distributions. This is a first sign, that we have to be careful with general, quantitative statements about the return distributions. Or in other words, we should know what we are doing.

Kurtosis Although kurtosis may not be the right measure to estimate the tails (I am getting a little ahead of myself), it can be insightful to see how the tails of the distribution (in the sense of kurtosis) behave under aggregation.

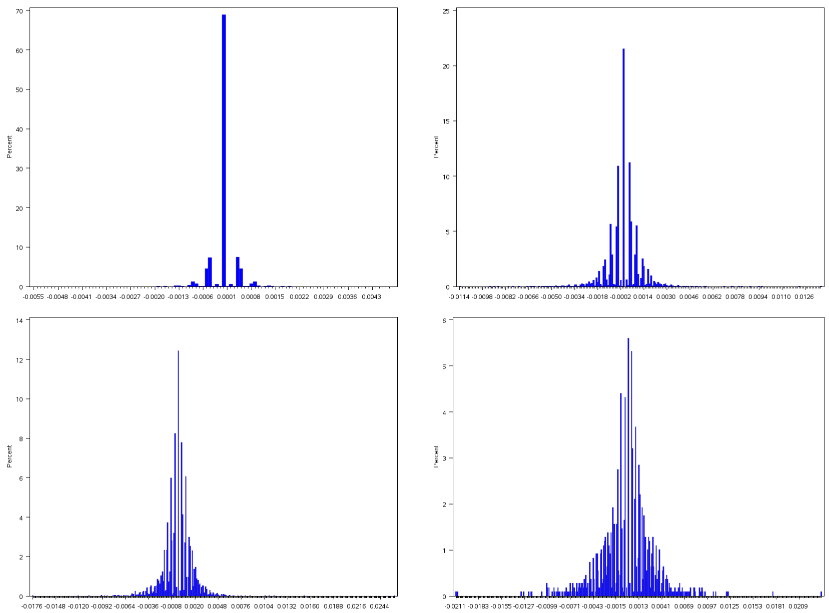


Figure 2.2: Histogram of log-returns of KPN for different frequencies. The data runs from January 1, 2010 through June 30, 2010. From top left clockwise: 15 second returns, 5 minute returns, 15 minute returns and 1 hour returns.

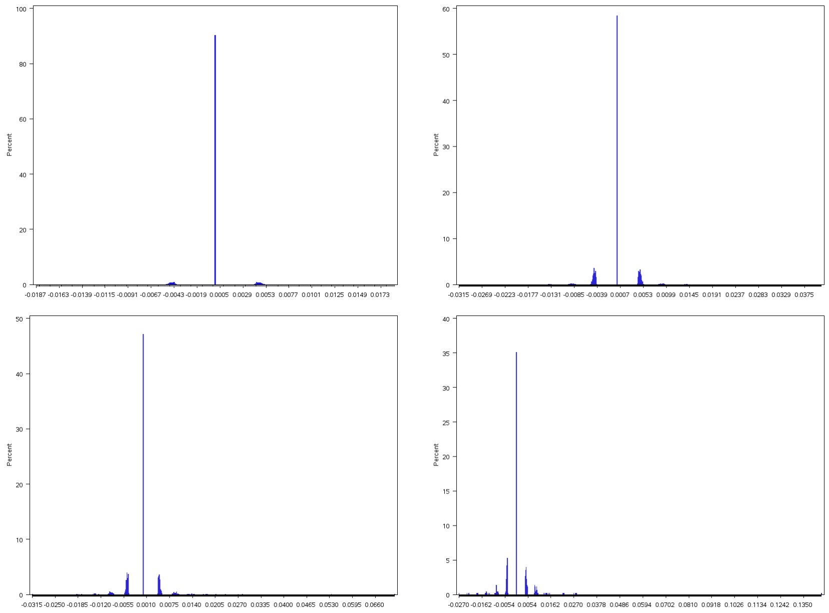


Figure 2.3: Histogram of log-returns of Pccw for different frequencies. The data runs from January 1, 2010 through June 30, 2010. From top left clockwise: 15 second returns, 5 minute returns, 15 minute returns and 1 hour returns.

The kurtosis increases with the frequency of the returns, which is to be expected if finite fourth moments and i.d.d. returns are assumed. In figures 2.4 and 2.5 kurtosis is plotted vs frequency for KPN and Pccw. It shows three lines.

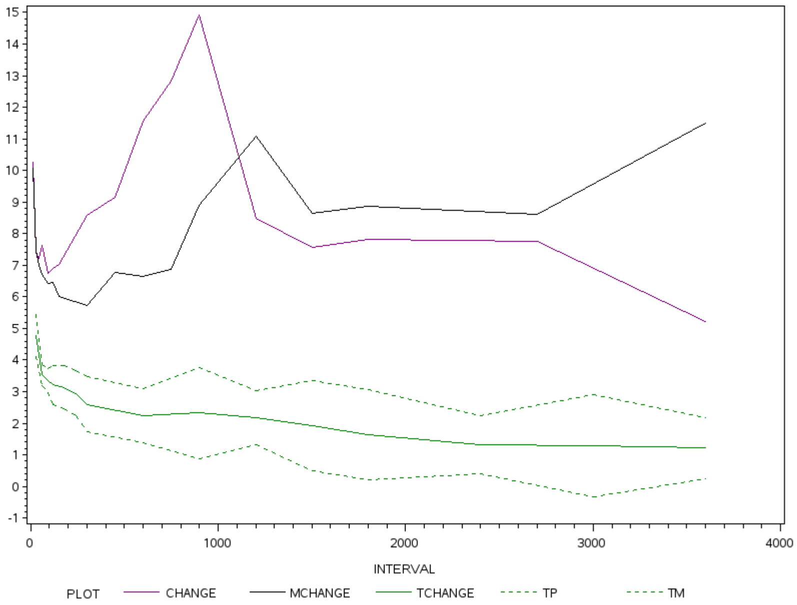


Figure 2.4: Kurtosis vs frequency (expressed in interval seconds) for KPN. The data runs from January 1, 2010 through June 30, 2010. The line labeled CHANGE shows kurtosis calculated using non-overlapping time intervals. The line labeled MCHANGE shows kurtosis calculated using overlapping time intervals. The line labeled TCHANGE is the theoretical kurtosis based on MCHANGE, calculated using a simple bootstrap. The dashed lines show the 95% confidence interval.

First line, labeled CHANGE, shows kurtosis that is calculated on normal log-returns over non-overlapping time intervals.

The second line, labeled MCHANGE, shows kurtosis calculated on log-returns over overlapping intervals. This means the following. Assume we are interested in 15 second returns and kurtosis. In that case we calculate the return over the first 15 seconds of trading (starting the first second of trading), followed by the returns over the 15 seconds starting from second 2 of trading, followed by the returns over 15 seconds starting in second 3, and so forth. I divided the interval in 15 parts, so in total you get 15 times more returns than in the normal case. For example, for the 1-hour interval returns, you get returns starting at 0 min, 4 min, 8 min, through to 56 min. The kurtosis is estimated by calculating the kurtosis over the 5 parts individually and then averaging over these 5 kurtoses. The advantages of this method are that the start of the sample choice does not influence the estimation and that we have more robust estimator. The line MCHANGE runs smoother and clearly shows an increase of the kurtosis with increasing frequency.

The third line is the theoretical kurtosis, calculated using a simple bootstrap

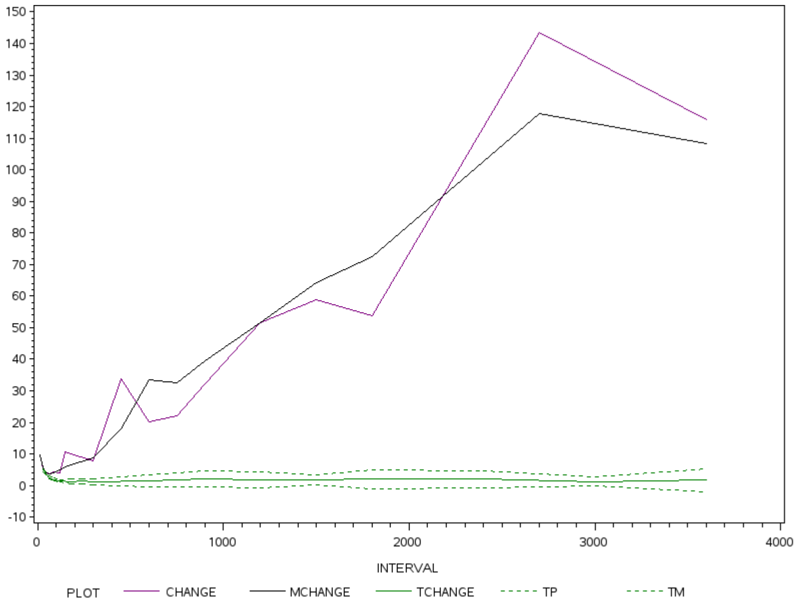


Figure 2.5: Kurtosis vs frequency for Pccw. The data runs from January 1, 2010 through June 30, 2010. The line labeled change shows kurtosis calculated using non-overlapping time intervals. The line labeled mchange shows kurtosis calculated using overlapping time intervals. The line labeled tchange is the theoretical kurtosis based on mchange, calculated using a simple bootstrap. The dashed lines show the 95% confidence interval.

and assuming i.d.d. returns. So for example, the theoretical 10 minute kurtosis $\kappa^t_{10\text{m}}$ is given by $3 + (\overline{\kappa}_{5\text{m}} - 3)/2$ in which $\overline{\kappa}_{5\text{m}}$ is the estimate of the 5 minute kurtosis (and in this case based on overlapping returns).[1]

Clearly, the estimated kurtosis is a lot bigger than can be expected if we assume i.d.d. returns. The daily (sample) kurtosis of KPN between January 1, 2010 through June 30, 2010 is estimated on 4.15. The daily (sample) kurtosis of PCCW for the same date interval is 15.09.

These findings are in line with prior research. For example [Pictet et al., 1996] report kurtoses of foreign exchange rate returns of some major currencies against the USD. For 30 minutes intervals the kurtosis of the USD/CHF is 66.76 and for 6 hours and 24 hours intervals the kurtosis is 10.09 and 5.68. The used sample runs from January 1, 1987 to June 30, 1996. This is a lot bigger than supposed normality would suggest.

The kurtosis of PCCW plotted in figure 2.5 seems to explode for lower frequencies up to one hour. Close look at the histogram for hourly returns shows extreme movement on January 20, 2010, and in particular between 07:00 GMT and 08:00 GMT. Figure 2.6 shows this movement of 17.5 % in little less than one hour and the sharp correction just after the start of trading the next day. Even on a daily scale this outbreak of volatility is extreme.

If we omit all data on January 20 and January 21, we get a clearer picture of the relation between frequency and kurtosis. Figure 2.7 shows this adjusted sample. It exhibits the same pattern as can be seen for KPN in figure 2.4: for

[1] Let X_i be stochastic variables with zero mean. Let S_N be a sum of N i.d.d. copies of X_i, so $S_N = \sum_{i=1}^N X_i$. Let σ^2_X be the unconditional variance and κ_X the unconditional kurtosis of X_i.

The variance of S_N is given by

$$\mathrm{E}(S_N^2) = \mathrm{E}((\sum_{i=1}^N X_i)^2) = \mathrm{E}(\sum_{i=1}^N X_i^2) + \mathrm{E}(\sum_{i=1}^N \sum_{j=1,i\neq j}^N X_i X_j) = N\mathrm{E}(X_i^2) = N\sigma_X^2 \tag{2.1}$$

in which the last part follows because of the independency of the copies. The kurtosis of S_N is given by $\mathrm{E}(S_N^4)/\mathrm{E}(S_N^2)^2$. The first part is given by

$$\begin{aligned}\mathrm{E}(S_N^4) =& \mathrm{E}((\sum_{i=1}^N X_i)^4) = \mathrm{E}(\sum_{i=1}^N X_i^4) + \mathrm{E}(3\sum_{i=1}^N \sum_{j=1,i\neq j}^N X_i^2 X_j^2) + \\ & \mathrm{E}(4\sum_{i=1}^N \sum_{j=1,i\neq j}^N X_i^3 X_j) + \mathrm{E}(6\sum_{i=1}^N \sum_{j=1,i\neq j}^N \sum_{k=1,k\neq i,k\neq j}^N X_i^2 X_j X_k) + \\ & \mathrm{E}(\sum_{i=1}^N \sum_{j=1,i\neq j}^N \sum_{k=1,k\neq i,k\neq j}^N \sum_{l=1,l\neq i,l\neq j,l\neq k}^N X_i X_j X_k X_l) \\ =& \mathrm{E}(\sum_{i=1}^N X_i^4) + \mathrm{E}(3\sum_{i=1}^N \sum_{j=1,i\neq j}^N X_i^2 X_j^2) = N\mathrm{E}(X_i^4) + 3N(N-1)\sigma_X^4\end{aligned} \tag{2.2}$$

. such that the kurtosis of S_N, κ_S is given by

$$\kappa_S = \frac{\mathrm{E}(S_N^4)}{\mathrm{E}(S_N^2)^2} = \frac{N\mathrm{E}(X_i^4) + 3N(N-1)\sigma_X^4}{N^2\sigma_X^4} = \frac{1}{N}(\kappa_X - 3) + 3. \tag{2.3}$$

See [Lau and Wingender, 1989] for a review of this so-called intervaling effect.

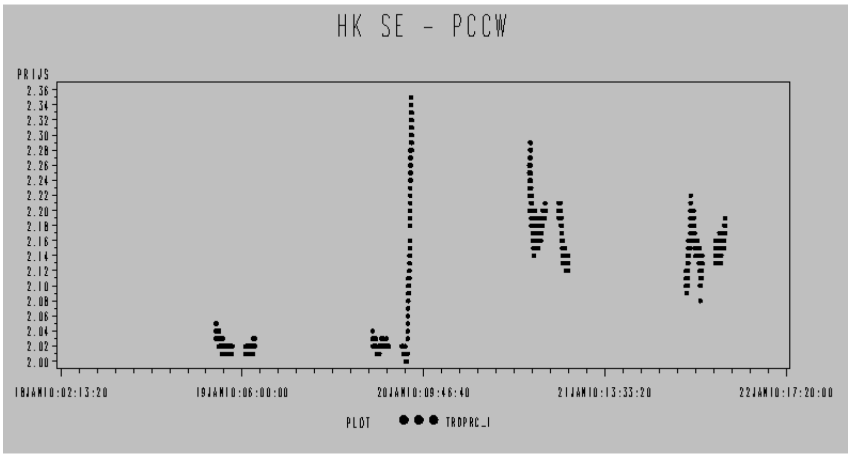


Figure 2.6: Four days of tick data for Pccw. The data runs from January 19, 2010 through January 22, 2010. The extreme hourly movement (extreme autocorrelation) in the last hour of trading on January 20 heavily influences the kurtosis for frequencies up to one hour.

the highest frequencies the kurtosis is very high, for lower frequencies it first decreases and then increases slightly and finally decreases slowly for the lowest frequencies. In chapter 4 and 5 this behavior will be examined more closely.

After some ‘google-ing’, I found a post on a tech blog[2] on January 20, 2010 that reports of some rumors about the possible introduction of the Google Nexus S smart phone on the Hong Kong market by Pccw. On January 21, Pccw publishes a press release on its website[3] its introduction of the first Android powered smart phone in Hong Kong. The rumors might have triggered the extreme movement on January 20 (and the correction on January 21).

Autocorrelations Intraday returns are almost uncorrelated and if they are, it is mostly negative. There are two plausible explanations for the (small) negative dependence: the bid-ask spread and relatively high trading costs for high frequency trading. Largest negative dependence can be found for highest frequency (tick-by-tick).

As an example, tick data for one day for the two individual stocks is plotted in figure 2.8. The average tick size for these stocks over the period January through June was for KPN 4.42 bps and for Pccw 45.5 bps. The average bid-ask spread was for KPN 4.45 and for Pccw 45.5 bps.

Table 2.1 makes clear that tick size and bid-ask spread have a negative effect on the autocorrelation (first lag), For Pccw the tick-by-tick returns are highly negatively correlated. Others (e.g. [Lin et al., 1999]) find this relatively high negative correlation for single stocks as well. They find for IBM and Intel

[2] see http://cn.engadget.com/2010/01/20/nexue-one-pccw-launch-rumor/

[3] see http://www.pccw.com/

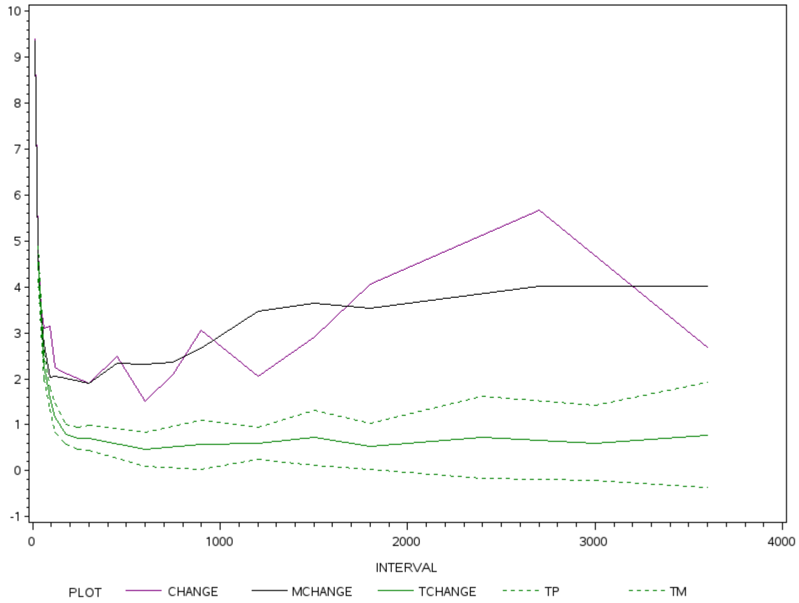


Figure 2.7: Kurtosis vs frequency for Pccw. The data runs from January 1, 2010 through June 30, 2010 and leaves out the data for January 20 and January 21. The line labeled change shows kurtosis calculated using non-overlapping time intervals. The line labeled mchange shows kurtosis calculated using overlapping time intervals. The line labeled tchange is the theoretical kurtosis based on mchange, calculated using a simple bootstrap. The dashed lines show the 95% confidence interval.

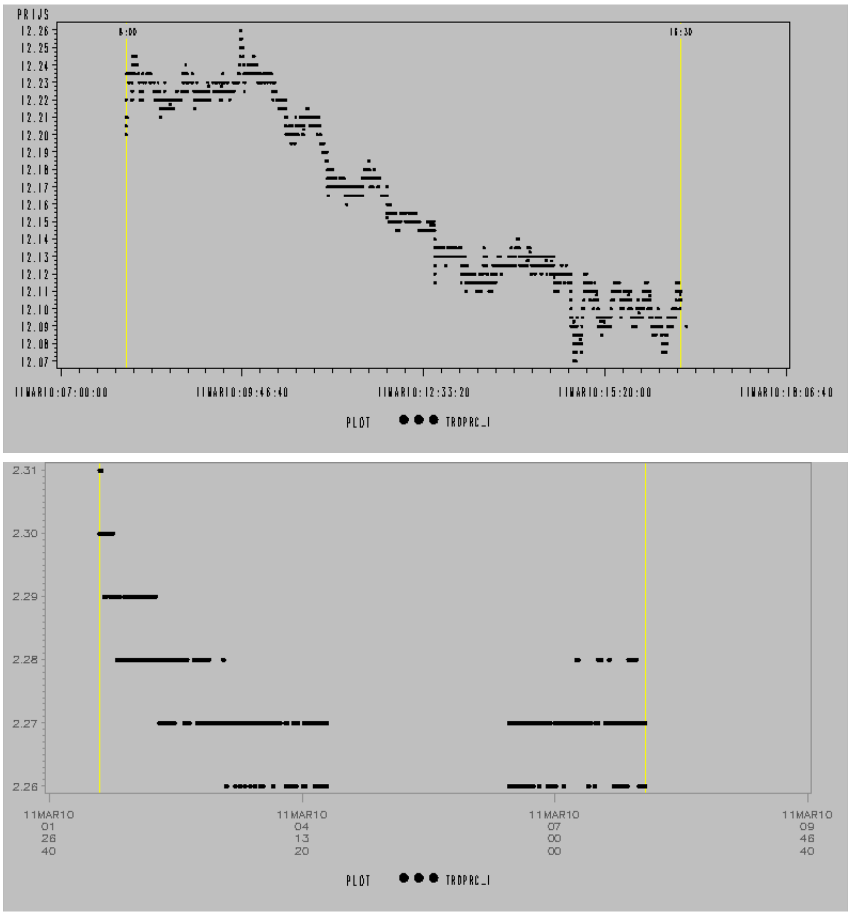


Figure 2.8: One day (March 11. 2010) of high frequency data for KPN (top) and Pccw (bottom).

autocorrelation of -0.27 and -0.48. For foreign exchange data and stock indices, the autocorrelation is a lot smaller, almost negligible.

Table 2.1: First order autocorrelation of high frequency returns

Frequency (in s)	KPN	Pccw
15	-0.13920	-0.17466
30	-0.15471	-0.23268
45	-0.14926	-0.26607
60 (1 minute)	-0.14227	-0.28761
90	-0.13088	-0.30851
120	-0.11637	-0.31895
150	-0.11528	-0.32669
300 (5 minutes)	-0.09218	-0.31968
450	-0.08006	-0.29424
600	-0.08004	-0.29068
750	-0.07172	-0.27241
900 (15 minutes)	-0.05544	-0.26538
1200	-0.04161	-0.23528
1500	-0.03278	-0.21580
1800	-0.03141	-0.19596
2700	-0.05640	-0.15744
3600 (1 hour)	-0.06199	-0.10390

Intraday volatility As for low frequency returns there is substantial positive dependence between absolute and squared returns for high frequencies. The dependence for high frequencies even seems to be more persistent and reaches out over multiple days. Table 2.2 shows this high dependency for KPN and Pccw. It is very intuitive to assume that high tick size and bid-ask spread have a positive effect on this dependancy.

The autocorrelation of absolute returns is not confined to first lags. Table 2.3 and 2.4 show the autocorrelation for multiple lags. Clearly, the dependency is significant for many lags, even for lags that capture a number of days. This stylized fact has been study by, amongst others, [Andersen and Bollerslev, 1997a], [Andersen and Bollerslev, 1997b] and [Dacorogna et al., 1993]. They find significant seasonal patterns for DM/$ absolute returns over thirty-minute and twenty minute respectively. See figure 2.9 from [Chang and Taylor, 2003].

2.3 Intraday seasonality

Seasonality is found for intraday volatility as well. [Taylor, 2005] express these finding in a fourth and fifth stylized fact.

4. *The volatility depends on the time of day.* This variation is significant.
5. *There may be shot bursts of high volatility in intraday prices.* These bursts are (mainly) caused by major macroeconomic announcements.

Table 2.2: First order autocorrelation of high frequency absolute returns

Frequency (in s)	KPN	PCCW
15	0.13677	0.11169
30	0.15385	0.13310
45	0.15970	0.14460
60 (1 minute)	0.16337	0.14951
90	0.16741	0.14894
120	0.17520	0.14854
150	0.18196	0.15390
300 (5 minutes)	0.18076	0.15092
450	0.17369	0.14798
600	0.17327	0.15657
750	0.17788	0.16552
900 (15 minutes)	0.17140	0.15966
1200	0.16426	0.16319
1500	0.15208	0.15693
1800	0.14964	0.16297
2700	0.13468	0.09954
3600 (1 hour)	0.12064	0.05775

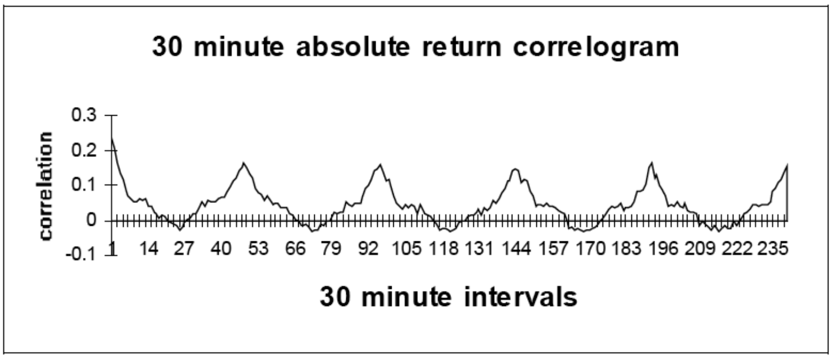


Figure 2.9: Autocorrelation for DM/$ thirty-minute absolute returns, see [Chang and Taylor, 2003].

Table 2.3: Autocorrelation of KPN absolute returns for multiple lags.

Frequency (in s)	lag 1	lag 2	lag 3	lag 4	lag 5	lag 10	lag 15	lag 20	lag 25	lag 30
15	0.13677	0.10678	0.09423	0.09416	0.08788	0.08426	0.08543	0.08813	0.08258	0.08387
30	0.15385	0.11769	0.10841	0.10844	0.10439	0.10417	0.09934	0.09646	0.09614	0.09612
45	0.15970	0.12543	0.12052	0.11692	0.10864	0.11010	0.10736	0.10565	0.10218	0.10425
60 (1 minute)	0.16337	0.13135	0.12622	0.11887	0.11920	0.11566	0.11140	0.11061	0.11313	0.11172
90	0.16741	0.14498	0.13596	0.13130	0.12519	0.12650	0.12704	0.12092	0.11770	0.11809
120	0.17520	0.15100	0.14407	0.13326	0.13412	0.13072	0.12443	0.12535	0.11677	0.11926
150	0.18196	0.15507	0.14458	0.14254	0.13734	0.13256	0.12958	0.12665	0.12151	0.12315
300 (5 minutes)	0.18076	0.15646	0.15019	0.14598	0.13537	0.13331	0.12702	0.13072	0.12272	0.11379
450	0.17369	0.15926	0.14304	0.13554	0.13239	0.12504	0.11484	0.11888	0.11477	0.10508
600	0.17327	0.15676	0.14083	0.13306	0.12213	0.12139	0.11004	0.10913	0.10368	0.10106
750	0.17788	0.15178	0.13801	0.13289	0.11840	0.10998	0.11001	0.10451	0.09491	0.09302
900 (15 minutes)	0.17140	0.14593	0.14077	0.12016	0.10698	0.11026	0.09360	0.09653	0.08021	0.07570
1200	0.16426	0.12878	0.11781	0.10169	0.10778	0.07973	0.07838	0.06277	0.05345	0.04716
1500	0.15208	0.12331	0.10305	0.10294	0.08133	0.06133	0.05282	0.05438	0.03011	0.02591
1800	0.14964	0.13198	0.09896	0.07436	0.06108	0.05327	0.04343	0.02607	0.01879	0.04137
2700	0.13468	0.09969	0.07098	0.07301	0.03672	0.03390	0.02756	0.06845	0.07074	0.07836
3600 (1 hour)	0.12064	0.05729	0.06239	0.04363	0.03910	0.06482	0.06339	0.08996	0.09355	0.05619

Table 2.4: Autocorrelation of Pccw absolute returns for multiple lags.

Frequency (in s)	lag 1	lag 2	lag 3	lag 4	lag 5	lag 10	lag 15	lag 20	lag 25	lag 30
15	0.11169	0.07255	0.06013	0.05705	0.05456	0.04449	0.04281	0.04369	0.04371	0.04444
30	0.13310	0.07594	0.06115	0.05481	0.05552	0.05152	0.05080	0.05487	0.05106	0.05030
45	0.14460	0.07382	0.06745	0.06079	0.05719	0.05665	0.05006	0.05187	0.04831	0.04132
60 (1 minute)	0.14951	0.07363	0.06362	0.05923	0.05436	0.05386	0.04719	0.04728	0.04879	0.04354
90	0.14894	0.07135	0.05465	0.06046	0.05471	0.04958	0.04971	0.04836	0.04316	0.04084
120	0.14854	0.06623	0.06684	0.06091	0.05119	0.05141	0.04021	0.04383	0.04349	0.04740
150	0.15390	0.06991	0.06923	0.05317	0.05368	0.05163	0.04731	0.05000	0.04498	0.04109
300 (5 minutes)	0.15092	0.07449	0.07505	0.06630	0.06728	0.05314	0.04792	0.05718	0.05610	0.05054
450	0.14798	0.08830	0.07096	0.06823	0.07625	0.06361	0.06378	0.06928	0.07599	0.05638
600	0.15657	0.09091	0.09479	0.09133	0.07892	0.08751	0.08057	0.06060	0.06774	0.08441
750	0.16552	0.09713	0.09846	0.09067	0.09188	0.08613	0.06581	0.06645	0.08011	0.07587
900 (15 minutes)	0.15966	0.11129	0.09047	0.10546	0.09155	0.08275	0.08871	0.08856	0.06238	0.08309
1200	0.16319	0.11108	0.08797	0.07834	0.06898	0.05884	0.06840	0.07198	0.07181	0.08512
1500	0.15693	0.10053	0.09836	0.08182	0.08481	0.07780	0.08472	0.10676	0.10396	0.09884
1800	0.16297	0.10140	0.10905	0.09566	0.10446	0.09559	0.11399	0.13184	0.11195	0.08920
2700	0.09954	0.03747	0.02536	0.05749	0.07364	0.21690	0.05347	0.00278	0.00265	0.05240
3600 (1 hour)	0.05775	0.01468	0.06749	0.20933	0.01777	0.02155	0.11147	0.15451	0.01868	0.00490

Following [Taylor and Xu, 1997b] the intraday volatility patterns are estimated by assuming the periodic pattern repeats itself every day. Suppose the return for day t, denoted by r_t, is the sum of N intraday returns, $r_{t,j}$, $1 \leq j \leq N$. Let the return from market close on day $t-1$ to open on day t be given by $r_{t,1}$. The latent volatility for day t is σ_t, so

$$r_t = \sum_{j=1}^{N} r_{t,j} \quad \text{and} \quad \operatorname{var}(r_t|\sigma_t) = \sigma_t^2 \tag{2.4}$$

Simple estimates of the variance proportions, if zero mean and uncorrelated intraday returns are assumed, are given by

$$\hat{\lambda}_j = \frac{\sum_t r_{t,j}^2}{\sum_t \sum_{k=1}^{N} r_{t,k}^2} \quad \text{and} \quad \hat{\kappa}_j = \frac{\sum_t r_{t,j}^2}{\sum_t \sum_{k=2}^{N} r_{t,k}^2} \tag{2.5}$$

for all day and market open volatility respectively.

Figure 2.10 illustrates this intraday volatility pattern for KPN and Pccw (time zone is local time). It shows for both stocks a (clear) U-shaped pattern. For KPN the mean (unscaled) absolute return is at its highest at the opening and closing of the market and at its lowest around midday. In addition to that, there are some spikes around 14:30,15:30 and 16:00, which are all due to the opening of the US futures and stock markets and US news releases. For Pccw volatility levels at the opening and closing of both trading sessions is higher than at the middle of the trading sessions.

2.4 Filtering

All data contains aberrant outliers, some caused by human and some by technical errors (which are constructed by human hand). Most of these outliers have a clear effect on data quality, see for example figures 2.1 and 2.11.

These errors need to be filtered before the data can be used for analysis. In filtering, there exist two risks. First, the risk of underfiltering and not flagging incorrect trades and quotes. Second, the risk of overfiltering and flagging to many trades and quotes. Both risk may severely influence your analysis by overstating or understating data quality.

Some authors regard data filtering vital for there research. For example, [Muller et al., 1990] find in their analysis of high frequency foreign exchange data the following types of high frequency foreign exchange specific errors.

1. The prices of our data source are quoted prices and not actual trading prices.
2. The prices come from many contributors in an irregular sequence. The market makers tend to publish new prices in order to attract traders in the direction in which they want to trade.

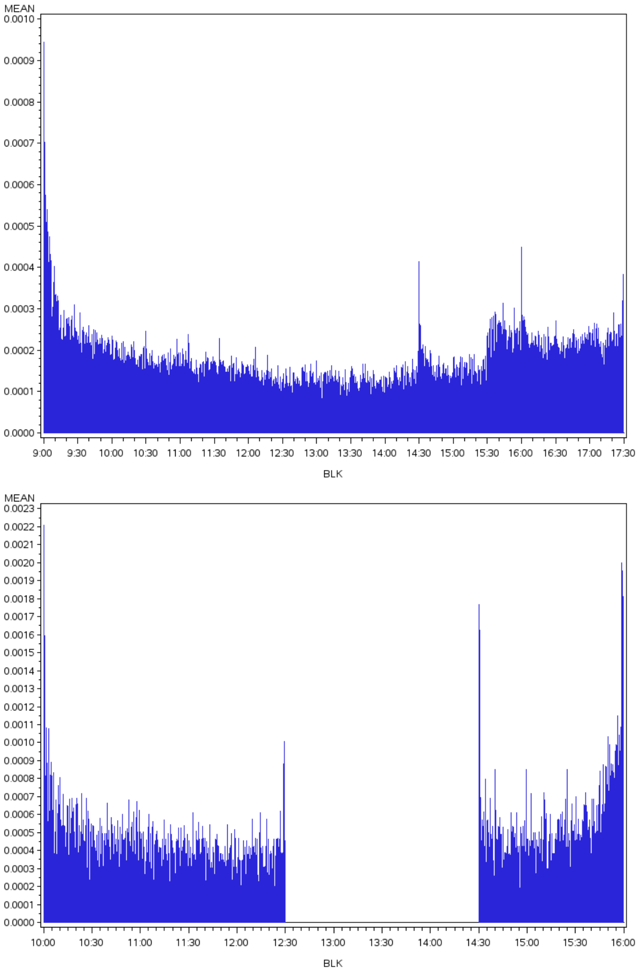


Figure 2.10: Intraday volatility for KPN (top) and Pccw (bottom). Time zone is local time. Volatility is estimated by 15 seconds absolute returns.

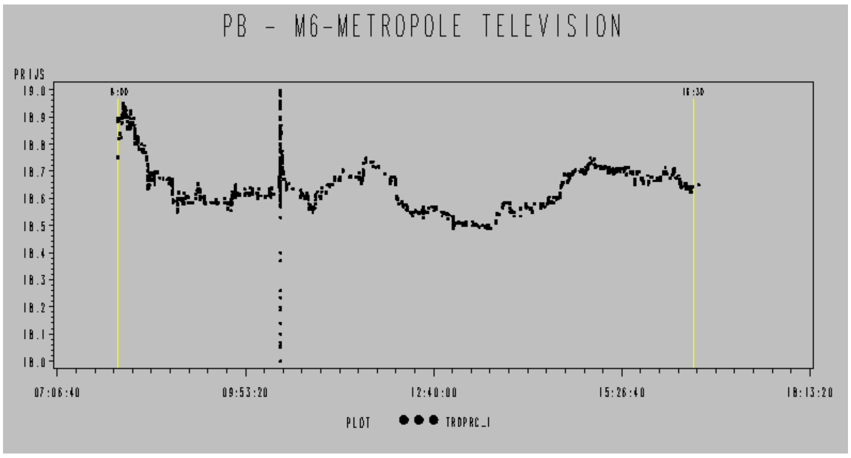


Figure 2.11: Traded prices of the common stock of M6-Metropole Television, traded on Nov. 5 2010 on Euronext Paris. A clear spike at 10:25 severely influences the profile of this stock on that day.

3. The main local markets can have different trading habits (e.g., different average volumes per transaction or bid-ask spreads) even if their active periods overlap.
4. There are transmission delays varying from few seconds up to few minutes.
5. There are transmission breakdowns or other failures that cause database holes.
6. Some very infrequent prices are completely aberrant (such as 100 times the normal price). These outliers are due to human and technical errors in the communication channels.

Many of these errors are very specific for the data and data sources that are used. Moreover, high frequency trading has much increased over the years, both in volume as in number of agents, i.e. trading specialists and trading strategies. This has led to new and different types of errors. As an example, consider the flash crash outliers of May 2010 due to human programming errors of ultrahigh frequency trading strategies. See for this example figure 2.12.

In the flash crash of May 6, 2010, at 2:42 pm, "the Dow Jones equity market began to fall rapidly, dropping more than 600 points in 5 minutes for an almost 1000 point loss on the day by 2:47 pm. Twenty minutes later, by 3:07 pm, the market had regained most of the 600 point drop. According to a report by the CFTC and SEC[4], the main cause of the crash was due to a failing computer algorithm: against a backdrop of unusually high volatility and thinning liquidity that day, a large fundamental trader (a mutual fund complex) initiated a sell

[4] *Findings regarding the market events of may 6, 2010*, report of the staffs of the CFTC and SEC to the joint advisory committee on emerging regulatory issues. See `http://www.sec.gov/news/studies/2010/marketevents-report.pdf`.

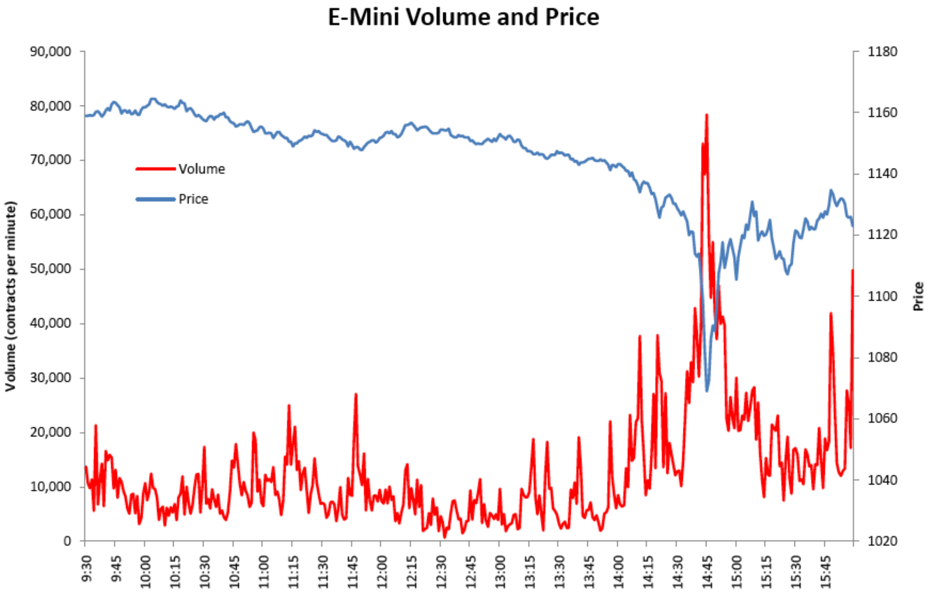


Figure 2.12: Flash crash of May 6, 2010. E-Mini volume and price.

program to sell a total of 75,000 E-Mini contracts (valued at approximately \$ 4.1 billion) as a hedge to an existing equity position. The computer algorithm was set to target an execution rate set to 9 % of the trading volume calculated over the previous minute, but without regard to price or time." [5]

[Falkenberry, 2002] gives an overview of high frequency equity data errors and filtering. In summary, bad high frequency data *emerges from the asynchronous and voluminous nature of financial data.* Errors occur due to human mistakes, both directly through trading as indirectly through designing and developing the technical infrastructure. A filter should do the following.

1. Create a time series for historical research that eliminates outliers in the trader's base unit of analysis without introducing concept and techniques that cannot be applied in realtime.
2. Not change the statistical properties of data relative to that which will be used in real time.
3. Not introduce excessive delay due to computation time or the need for excessive confirming data points, i.e. a suspected bad tick at time t being confirmed by future prices generated at time $t+1, t+2$, et cetera.
4. Be adaptive across securities with different tick frequency profiles
5. Be adaptive across securities with different price levels

[5] See `http://en.wikipedia.org/wiki/Flash_crash`.

He argues that

> The primary objective in developing a set of tick filters is to manage the overscrub/ underscrub tradeoff in such a fashion as to produce a time series that removes false outliers in the trader's base unit of analysis that can support historical backtesting without removing real-time properties of the data.

This kind of filter is more general and less dependent on the specific type of analysis. However, these requirements are not complete. For example, if you want to flag trades with a price of 0, which makes sense for normal securities or futures contracts, obviously you cannot use this filter for calendar spread futures contracts, such as wheat - wheat futures for example, tradable as calendar spreads on some exchanges. So being adaptive across securities is questionable.

In general, the ultimate goal must always be in mind when creating a correct filter. It is an illusion to think that a filter will only flag errors and leave the correct data totally intact. Consider this comparable with a physicist who changes a quantum system by observing it. This is fundamentally inevitable.

In this thesis, I will use in house TRANSTREND data. This data is filtered using a filter that is based on a two step approach.

1. Filter source specific, i.e. exchange and/or product/security specific, trades and quotes which are *labelled by the source* to be *fictitious*, *non representative* or just *wrong*. These include, amongst others, trades that are reported out-of-order or delayed, trades that are not single traded but part of a bunched trade or spread trade, quotes that are part of a opening or closing auction, et cetera. This includes cancelled trades as well. So, a traded that is cancelled by an exchange, will be filtered in this part.

2. Filter non-source specific trades and quotes that are outliers compared to trades and quotes in their direct proximity, in respect of time, level and volatility, using both information from the past and the (nearby) future. This means that in some cases, information that comes available after the fact (such as price levels after some news announcement) is used to filter. The boundaries implied by direct proximity will depend on profile, so 'normal' activity and volatility.

The first part is obvious and mostly easy. Trades and quotes can be flagged and deleted or adjusted. A concrete example of such a trade is a trade that is transmitted at 11:01 but has a time stamp of 10:59. This time stamp is created by the data vendor or even by the exchange. The trade is easily adjusted to 10:59, as long as there are no other trades or quotes after 10:59 and before 11:01.

Figures 2.13 and 2.14 show two examples of quotes that are indicative and not usable for statistical analysis.

The second part involves more subjective rule-making. The base assumption behind the rule-making is that any trade or quote could have been made by a rational trader. So, for example, an offer below the prevailing best bid price is

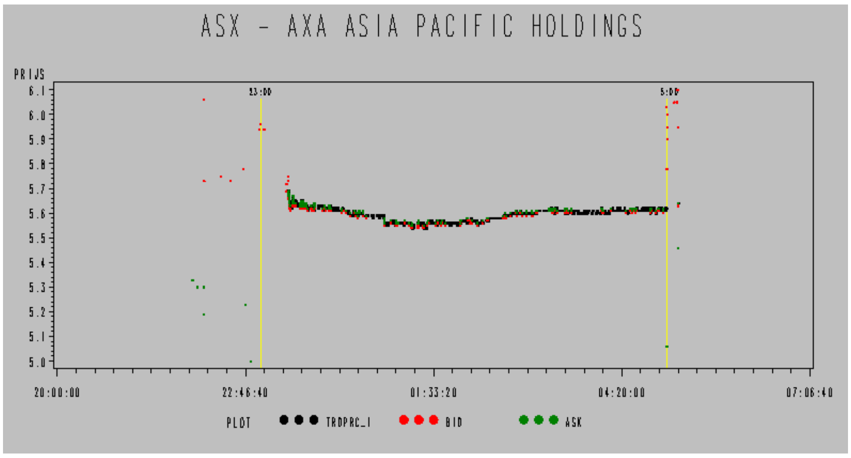


Figure 2.13: Traded prices and quotes of the common stock of AXA Asia Pacific Holdings, traded on Nov. 5 2010 on Australian Securities Exchange. Quotes pre-hours and after-hours are part of a opening and closing auction and are clearly wrong (bids are higher than offers). Even after the official opening some quotes are given that are wrong and could influence analysis.

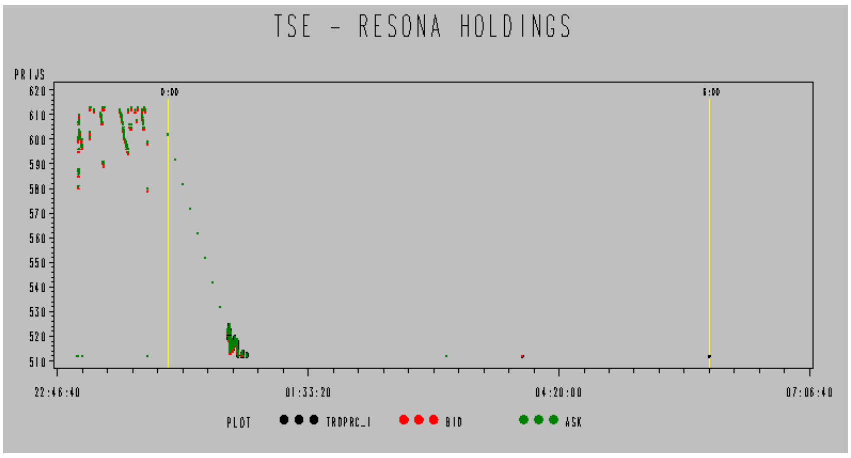


Figure 2.14: Traded prices and quotes of the common stock of Resona Holdings, traded on Nov. 5 2010 on Tokyo SE. Quotes pre-hours and after-hours are part of a opening and closing auction. The offers after the opening are wrong (you could buy at this prices). The stock went limit-down some minutes after the official opening.

not rational. However, a trade very much off the market is not irrational per se. It can be a trade by informed agents; this will only be clear after some other quotes at the same new level. If the market returns to the old level within certain time and range, the off-market trade will be deleted, as in retrospective it was not a rational trade (for one of the involved parties at least).

The acceptance levels are based on prior market activity and volatility.

Table 2.5 shows the percentage of filtered trades for some (groups of) stocks. In general between 0% (KRX Korea Exchange) and 20% (Tokyo SE) of quotes and trades are filtered, mostly depending on exchange and relevant trading platform (all electronic, composite, partly open out-cry, opening/closing auctions).

Table 2.5: Mean percentage of filtered trades and quotes per day between Oct 8, 2010 and Nov 5, 2010.

Stock	part 1 (%)	part 2 (%)
KPN	0.057	0.92
Pccw	5.9	0.13
Euronext AMS (37 stocks)	0.15	2.1
Hong Kong SE (222 stocks)	5.9	0.47

Chapter 3

Stable distributions

3.1 Mandelbrot and the stable hypothesis

In his seminal work on *The variation of certain speculative prices* [Mandelbrot, 1963] Mandelbrot critically reviews the work of Louis Bachelier on the independence and normality of price changes of the price of a stock, or of a unit of a commodity (see e.g. [Bachelier, 1900]). Bachelier argues that if the price changes from transaction to transaction are independent, identically distributed, random variables with finite variance, and if the transactions take place uniformly spaced through time, the central-limit theorem leads to price changes across days, weeks or months that are normally distributed (as they are sums of changes from transaction to transaction [Fama, 1963].

In more formal terms, Bachelier assumes that, if we let $Z(t)$ be the price at the end of time t, then the successive differences $Z(t + \Delta t) - Z(t)$ are independent, normally-distributed random variables wit zero mean and variance proportional to the differencing interval Δt. This process has become commonly known as *Brownian motion.* According to Mandelbrot is has been known to empirical economists since 1900 that price changes of most financial time series were too peaked to be samples from Gaussian distributions. This is equivalent to stating that histograms of price changes usually contain so many outliers that a normal distribution fitted to the histogram of data price change series is much lower and flatter than the distribution of the price changes themselves. Figure 3.1 is reprinted from Mandelbrot's work and shows a Bell curve fitted to histograms of fifth and tenth price changes of wool between 1800-1937.

Mandelbrot finds that the tails of distributions of price changes are so extraordinarily long that the second moment (variance) does not tend to any limit. These observations lead to an approach which he warrants to be radically new. He makes the following two assumptions: 1) the variances of empirical distributions of price changes are infinite, and 2) the empirical distributions are best described by a non-normal family of probability distributions first discovered by Paul Lévy [Lévy, 1925], which he calls *stable Paretian* (I will call them *stable* distributions).

The implications of these assumptions reach far. If the variance of price changes

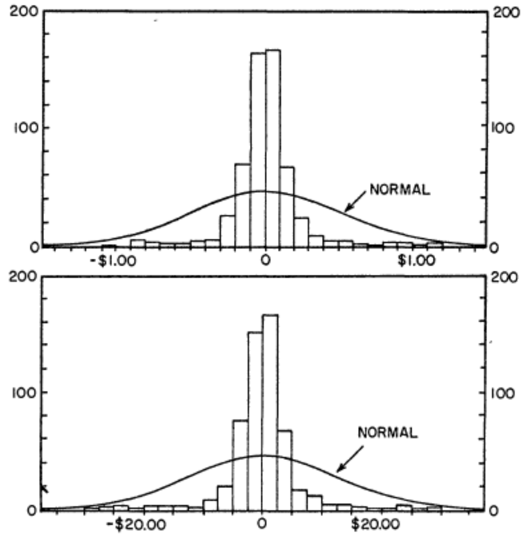


Figure 3.1: Two histograms illustrating departure from normality of the fifth and tenth difference of monthly wool prices, 1890-1937. In each case, the continuous bell-shaped curve represents the Gaussian "interpolate" based upon the sample variance. See [Mandelbrot, 1963].

is infinite, sample variance is a meaningless measure of dispersion and with this a meaningless measure of risk. Many modern statical tools in finance are based or dependent on the assumption of finite variance, which is misleading if the stable Paretian hypothesis holds.

See e.g. [Fama, 1963] for a comprehensive discussion of the theoretical and empirical implications Mandelbrot's findings. See [Fama, 1965] for Mandelbrot's ideas applied to US stock returns. This research finds fat tails and volatility clustering.

3.2 Stable paretian distributions

The logarithm of the characteristic function of *stable* distributions[1] is given by

$$\log f(t) = \begin{cases} i\delta t - \gamma^{\alpha}|t|^{\alpha}[1 - i\beta\frac{t}{|t|}\tan(\alpha\frac{\pi}{2})] & \text{if } \alpha \neq 1 \\ i\delta t - \gamma|t|[1 + i\beta\frac{t}{|t|}\frac{2}{\pi}\log(\gamma|u|))] & \text{if } \alpha = 1. \end{cases} \tag{3.1}$$

(The case if $\alpha = 1$ will be sometimes left out for convenience). The four parameters α, β, γ and δ are respectively called the *index of peakedness* (or *stability index* or *tail index*), *index of skewness, scale parameter* and the *location parameter.*

[1] There are actually many parametrizations, see e.g. [Nolan, 2010].

The skewness index β determines the symmetry of the distribution. Its values can be $-1 \leq \beta \leq 1$. If $\beta = 0$ the distribution is symmetric.

In review of extreme risks, so the tails of a distribution, the stability index α is crucial. It determines the total probability contained in the tails.

The values of α are in the interval between 0 and 2. If we let $\alpha = 2$, we get

$$\log f(t) = i\delta t - \gamma t^2. \tag{3.2}$$

which is the logarithm of the characteristic function of the Gaussian distribution. When $0 < \alpha < 2$, the extreme tails are higher than for the normal distribution. The variance only exists (i.e. is finite) is $\alpha = 2$. The mean of stable distributions exists as long as $\alpha > 1$. Mandelbrot assumes that $1 < \alpha < 2$, so that the mean of the distribution of price changes exists but that the variance is infinite. This is in contrast to the Gaussian hypothesis which asserts $\alpha = 2$, for which the variance does exist.

It should be emphasized that although the population variance does not exist, for finite samples the sample variance (and higher moments) can always be calculated. The information contained in this is however very limited since the sample moment does not converge to the population moment.

There are three important properties of stable distributions. First of all, stable distributions are fat-tailed, or *leptokurtic*, so more probability mass is located in the tails of the distribution than for the normal distribution. Secondly, stable distributions are invariant under addition (this is actually called stability), this means that the linear combination of two independent identically distributed random variables has the same distributions as the two variable up to a location and scale parameter. Thirdly, a generalized version of the Central Limit Theorem states that sum of a large sum of independent, identically distributed random variables has a stable distribution[Gnedenko and Kolmogorov, 1954]. So for distributions with finite variance this yields the Central Limit Theorem with a limiting normal distribution for sums, and for distribution with infinite variance this yields a limiting stable distribution for sums.

In terms of characteristic functions, a more precise definition of invariance under addition is that the sum of i.i.d. stable variables is given by the logarithm of the characteristic function

$$\begin{aligned}\log \prod_{n=1}^{N} f(t) &= \sum_{n=1}^{N} \log f(t) \\ &= i(N\delta)t - (N\gamma^{\alpha})|t|^{\alpha}[1 - i\beta \frac{t}{|t|} \tan(\alpha \frac{\pi}{2})]\end{aligned} \tag{3.3}$$

So the sum has the same distribution up to the scale, which is multiplied by $N^{\frac{1}{\alpha}}$, and the location parameter, which is multiplied by N. The skewness index and the stability index are constant under addition.

In practice, this would mean that if daily price changes follow a stable distribution, the sum of daily price changes, i.e. weekly, monthly or yearly price changes, would follow the same distribution with location and scale parameters multiplied by 5, 21 or 250 (i.e. the number of weekdays in a week, month or year).

It can be easily shown that the stability still holds when the N individual members n of the sum have different location δ_n and scale γ_n parameters. The logarithm of the characteristic function of the sum is then given by

$$\begin{aligned}\log \prod_{n=1}^{N} f_n(t) &= \sum_{n=1}^{N} \log f_n(t) \\ &= i(\sum_{n=1}^{N} \delta_n)t - (\sum_{n=1}^{N} \gamma_n^\alpha)|t|^\alpha [1 - i\beta \frac{t}{|t|} \tan(\alpha \frac{\pi}{2})]\end{aligned} \tag{3.4}$$

The parametrization of the distributions given above actually follows from the definition of stability (see e.g. [Nolan, 2010]). A distribution X is called *stable* if the following holds. Let $X, X_1, X_2, \ldots, X_N$ be independent, identically distributed stable random variables, then

$$X_1 + X_2 + \ldots + X_N \overset{d}{=} c_N X + d_N \tag{3.5}$$

in which $c_N > 0$ and d_N are some constants. A distribution is called *strictly stable* if $d_N = 0$. Obviously the normal distribution is stable, as are the Cauchy and Lévy distributions. The class of all stable laws that satisfy this property is given by (3.1).

3.3 Tail approximation

Except for $\alpha = 2$, stable distributions have a asymptotic power-law (also called *Pareto* or *Zipf* law) tails. For the right tail, $x \to \infty$

$$P(X > x) \sim \gamma^\alpha c_\alpha (1 + \beta) x^{-\alpha} \tag{3.6a}$$

$$f(x) \sim \alpha \gamma^\alpha c_\alpha (1 + \beta) x^{-(\alpha+1)} \tag{3.6b}$$

in which $c_\alpha = \sin(\frac{\pi\alpha}{2})\Gamma(\alpha)/\pi$. Equivalently, for the left tail, $x \to -\infty$

$$P(X < -x) \sim \gamma^\alpha c_\alpha (1 - \beta) x^{-\alpha} \tag{3.7a}$$

$$f(-x) \sim \alpha \gamma^\alpha c_\alpha (1 - \beta) x^{-(\alpha+1)}. \tag{3.7b}$$

If $\beta = 1$ (-1) the left (right) tail decays faster than any power.

See figure 3.2 below for some probability density plots. These plots have been made using the algorithm described in [Chambers et al., 1976] and [Weron, 1996]. This algorithm works as follows:

1. Generate a random variable U uniformly distributed on $(-\pi/2, \pi/2)$, so $U \sim U(-\pi/2, \pi/2)$.
2. Generate an independent exponential random variable E with mean 1, so $E \sim \exp(1)$.
3. For $\alpha \neq 1$, the random variable $X \sim S(\alpha, \beta, 1, 0)$ is given by

$$X = S_{\alpha,\beta} \frac{\sin(\alpha(U + B_{\alpha,\beta}))}{(\cos(E))^{1/\alpha}} \left(\frac{cos(U - \alpha(U + B_{\alpha,\beta}))}{E} \right)^{(1-\alpha)/\alpha} \tag{3.8}$$

in which

$$S_{\alpha,\beta} = \left(1 + \beta^2 \tan^2 \frac{\pi\alpha}{2}\right)^{1/(2\alpha)} \tag{3.9a}$$

$$B_{\alpha,\beta} = \frac{\arctan(\beta \tan \frac{\pi\alpha}{2})}{\alpha} \tag{3.9b}$$

For $\alpha = 1$, X is given by

$$X = \frac{2}{\pi} \left[\left(\frac{\pi}{2} + \beta U \right) \tan U - \beta \log \left(\frac{E \cos U}{\frac{\pi}{2} + \beta U} \right) \right] \tag{3.10}$$

4. Let $Y \sim S(\alpha, \beta, \sigma, \mu)$ given by

$$Y = \begin{cases} \sigma X + \mu & \text{if } \alpha \neq 1 \\ \sigma X + \frac{2}{\pi}\beta\sigma \log \sigma + \mu & \text{if } \alpha = 1 \end{cases} \tag{3.11}$$

3.4 The stability index

The stability index α thus essentially determines whether the stable or Gaussian hypothesis holds. So testing these hypotheses and measuring the true value of α is very important. However, the absence of analytic expressions for stable distributions except the Gaussian ($\alpha = 2$), the Cauchy ($\alpha = 1$ and $\beta = 0$) and the Lévy ($\alpha = 1/2$ and $\beta = 1$) distributions, makes this very hard to do.

OLS If we take the logarithm of $P(X > x)$ than $\log P(X > x) \sim -\alpha \log(x)$ for $x \to \infty$. A simple standard linear regression can be performed between (the logarithm of) the data and the corresponding probabilities to obtain the OLS estimator of α.

The advantage of such regression is that all aspects of least squares estimation are very well known. This includes bias, consistency and asymptotic normality. An additional advantage is that least squares regression is very easy to implement.

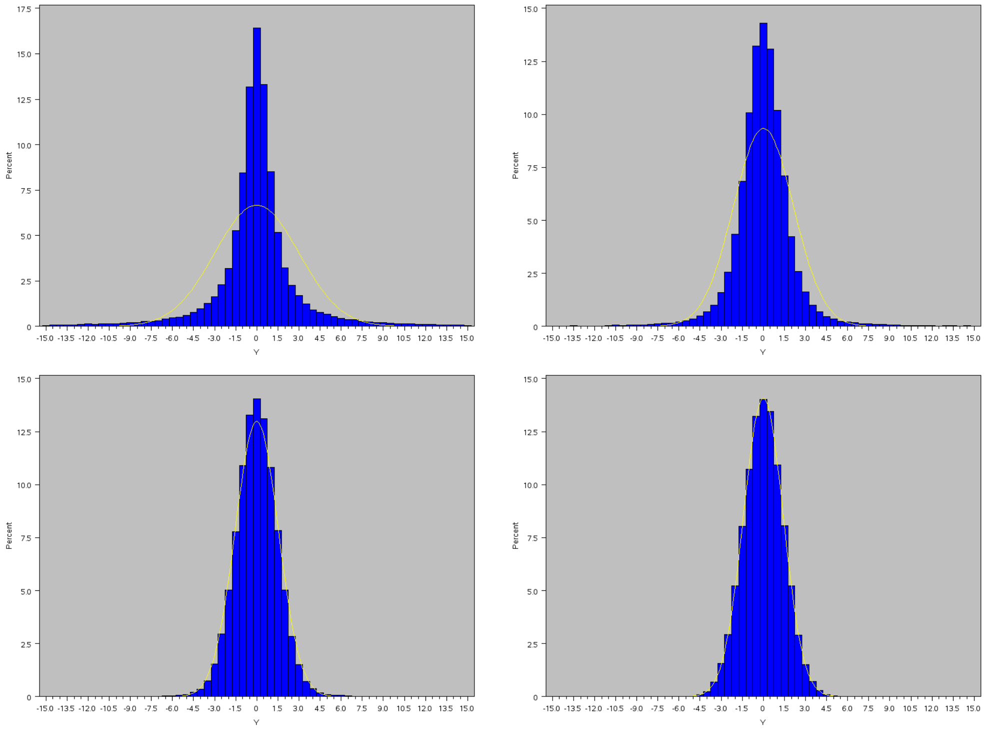


Figure 3.2: Random samples from stable distributions for $\alpha = 1, 1.5, 1.9$ and 2, $\beta = 0$, $\gamma = 1$ and $\delta = 0$. For the $\alpha < 2$ plots the tails have been truncated at $|X| < 15$. As a reference a fitted normal distribution is plotted as well.

The main disadvantage of estimating α of this regression is that it is (asymptotically) valid for the far tails $x \to \infty, x \to -\infty$. In practice there will be less observations in the tails, so the outcome will very much rely on the starting value of x. For high values ($\alpha > 1.5$) the estimator will overestimate the true value of α.

Hill Estimation A much more reliable and often used estimator is the *Hill* estimator, proposed by [Hill, 1975]. It is based on the evaluation of the conditional likelihood for the parameters describing the tail behavior (i.e. α) given the extreme order statistics.

Let $X^{(j)}$ be the j-th order statistic of X, so let $X^{(1)} \leq X^{(2)} \leq \cdots \leq X^{(N-j)} \leq \cdots \leq X^{(N)}$ be the order statistics. Assuming i.d.d. data, the Hill estimator (for the right tail) is given by

$$\hat{\alpha} = [\frac{1}{k}\sum_{i=0}^{k-1} \log X^{(N-i)} - \log X^{(N-k)}]^{-1} \tag{3.12}$$

conditional on $X^{(N-k)} \geq d$ in which d is some large enough threshold value. The mean and variance of $\hat{\alpha}$ are given by

$$E(\hat{\alpha}|X^{(N-k)} \geq d) = \frac{k\alpha}{k-1} \qquad \text{for } k > 1 \tag{3.13a}$$

$$\text{Var}(\hat{\alpha}|X^{(N-k)} \geq d) = \frac{(k\alpha)^2}{(k-1)^2(k-2)} \qquad \text{for } k > 2 \tag{3.13b}$$

The Hill estimator is proved to be asymptotically normally distributed, i.e. $k^{1/2}(\hat{\alpha}^{-1} - \alpha^{-1}) \sim N(0, \alpha^{-2})$ for large values of N and $k = k(N)$, see [Goldie and Smith, 1987]. The left tail estimator is obtained by multiplying the observations by -1 and rearranging the data in descending order.

As with linear regression the main drawback is that the estimation is very dependent on the choice of k. The order k has to be small enough to capture the tail of the distribution, and large enough to generate an appropriate sample size. There are several, sophisticated models for the choice of k, see e.g. [Mittnik and Paolella, 1999].

The Hill estimator is easy to implement and easy to interpret (see [Hill, 1975]). [2]

In figure 3.3 the behavior of the Hill estimator is illustrated in a Hill-plot. A random sample of a symmetric stable distribution is simulated with $N = 100000$ and $\alpha = 1.5$. Obviously, the Hill estimator is very misleading for the true value of α. Only for $k/N < 0.1$ the true value of α lies within the 95%-confidence interval. A similar behavior is described in [Mittnik and Paolella, 1998].

[2] The Hill estimator is only valid for so-called Fréchet extreme value distributions. The Fisher-Tippet(-Gnedenko) Theorem states that the standardized maximum (or minimum) of

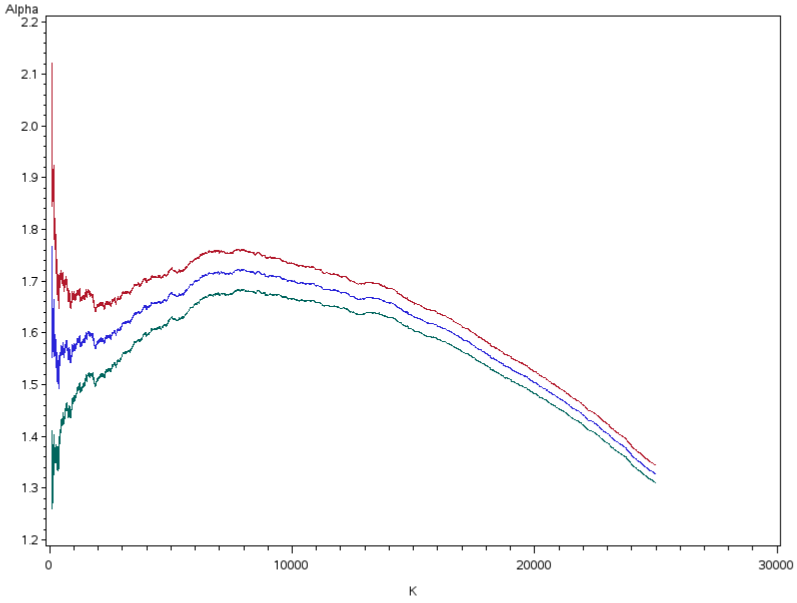


Figure 3.3: Hill-plot for $N = 100000$ and $\alpha = 1.5$ including an upper and lower band ($\pm 2 \cdot \text{MSE}$)

Other methods Other methods to estimate the tail index are based on quantile estimation, transformation of the characteristic function and maximum likelihood estimation.

a sample converges in distribution to one of three types of distributions: the *Gumbel* type, the *Fréchet* type or the *Weibull* type. These types are all cases of the generalized extreme value distribution and are defined as

$$F(x;\mu,\sigma) = e^{-e^{-(x-\mu)/\sigma}} \qquad \text{Gumbel} \qquad (3.14a)$$

$$F(x;\mu,\sigma,\alpha) = \begin{cases} 0 & \text{if } x \le \mu \\ e^{-((x-\mu)/\sigma)^{-\alpha}} & \text{if } x > \mu \end{cases} \qquad \text{Fréchet} \qquad (3.14b)$$

$$F(x;\mu,\sigma,\alpha) = \begin{cases} e^{-(-(x-\mu)/\sigma)^{\alpha}} & \text{if } x < \mu \\ 1 & \text{if } x \ge \mu \end{cases} \qquad \text{Weibull} \qquad (3.14c)$$

in which $x \in \mathbb{R}$. The Gumbel type domain of attraction contains thin tailed distributions, such as the normal, log-normal, exponential and gamma distributions. The Fréchet type contains fat tailed distributions like the Pareto, Cauchy and stable distributions. The Weibull type contains distributions with bounded support, such as the uniform and beta distributions ([Zivot and Wang, 2006]).

Chapter 4

The tails of KPN and Pccw

In section 2.2 the tails of KPN and Pccw are discussed in terms of kurtosis. The kurtosis of both stocks seems larger than normality would suggest and so kurtosis does not seem to be an appropriate measure. In order to get a better picture of the tails, I will focus on the asymptotic behavior of the stocks.

Moreover, I will assume that the appropriate distribution is stable and use the theory as described in sections 3.3 and 3.4.

4.1 The tails of KPN

For both KPN and Pccw I used all prices recorded between opening hours for all business dates excluding holidays from January 1, 2010 through June 30, 2010. I applied the filter as described in section 2.4 in order to clean the data. In the below analysis I used normalized log-returns. Moreover, I used overlapping returns (see figure 2.4). This means for example for the 15 second intervals, that I calculated 15 returns, starting from second 1 through second 15 of trading, after which I do 15 times the needed analysis and then take the median over the estimates.

First, I looked at the left tail of KPN. I applied two methods to estimate the tail index: OLS and the Hill Estimator. For both methods I used the 10% lowest returns.

Figures 4.1 and 4.2 show the tail index as a function of the frequency. Both figures show an increasing tail index with increasing frequency and seem to be mutually consistent with a small blip in highest frequency region. As the frequency goes down (so the interval increases) the distribution changes to a less fat-tailed distribution.

The stability of the estimators is shown in figures 4.3 and 4.4. Clearly, the window size (so which percentage of lowest returns) is very important. In both figures one can see that the mean estimate rises with frequency. The Hill estimator seems to be very unstable. Since both estimator show the same tendency of rising tail index with frequency and the OLS estimator seems to be more

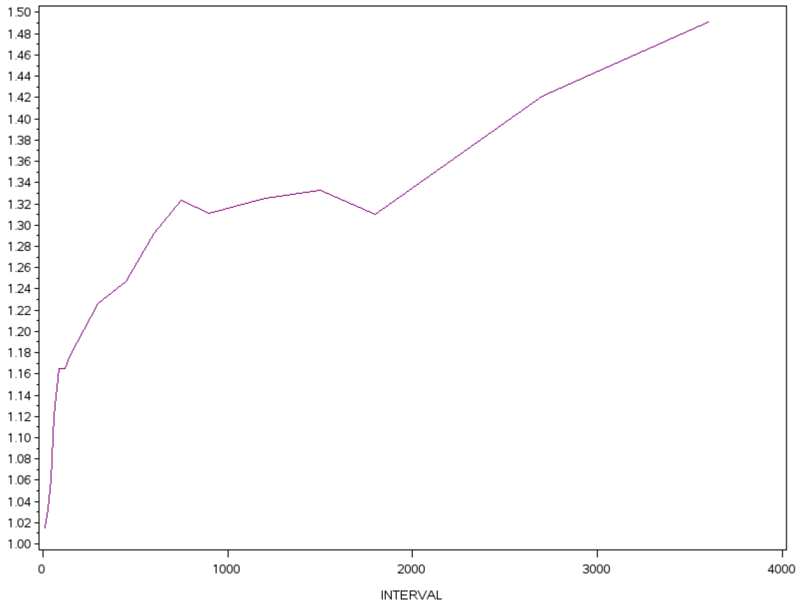


Figure 4.1: Tails index vs frequency for KPN estimated using the OLS method for a window size of 10%.

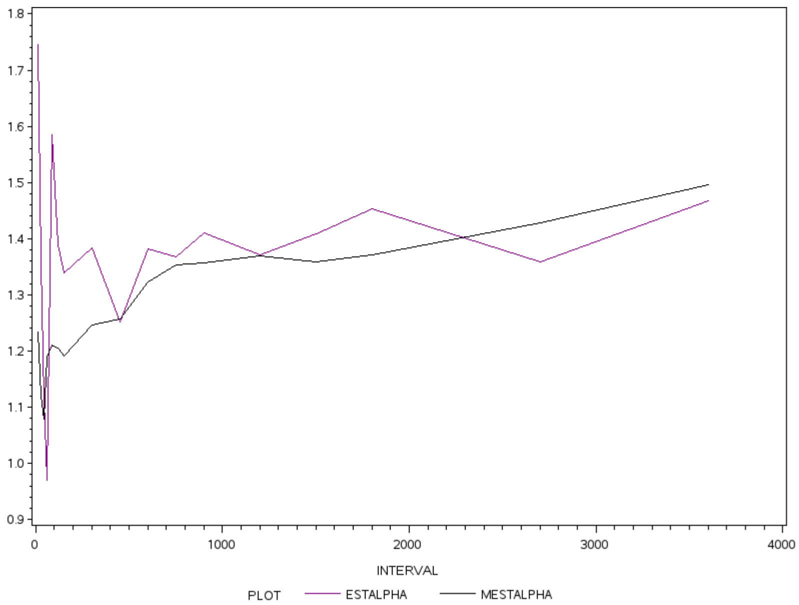


Figure 4.2: Tails index vs frequency for KPN estimated using the Hill method for a window size of 10% and a mean over all window sizes between 1% and 10%.

stable, I prefer the OLS estimator, although this is quite subjective and mainly motived by a preference for simplicity.

The choppy lines in figure 4.4 are due to the tick size, as the Hill estimator is very sensitive to the window size and the order statistic $X^{(N-k)}$ in equation 3.12. To a lesser degree, the choppy lines are visible in figure 4.3 as well. The OLS estimator is however less sensitive to tick size. Illiquidity (i.e. no price change over an interval) may have a more significant influence, as this could explain that the choppiness becomes less relevant as the frequency decreases.

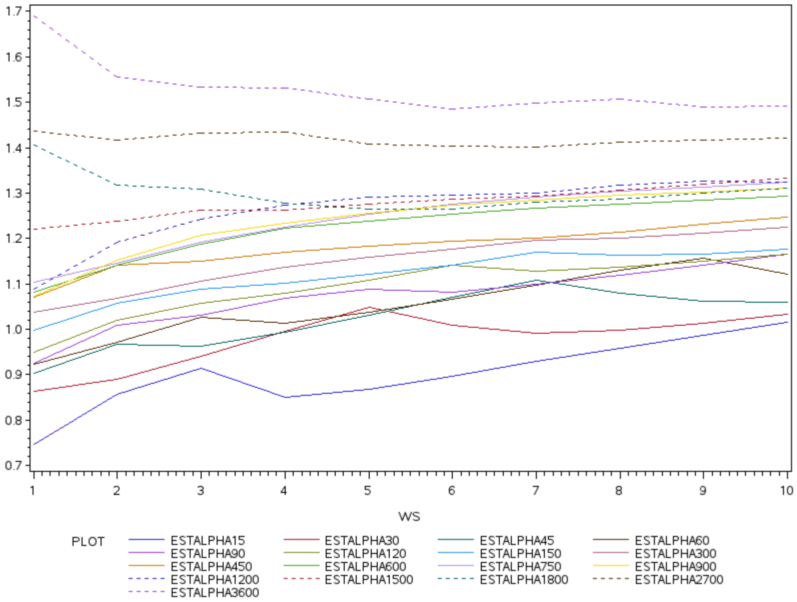


Figure 4.3: Tails index vs window size (%) for KPN estimated using the OLS method for different frequencies.

4.2 The tails of Pccw

The tails of Pccw are a little more extreme than for KPN. Figures 4.5 (OLS) and 4.6 (Hill) show the tail index as a function of the frequency. I did not leave out the data on January 20 and 21, 2010, since the OLS estimator and Hill estimator are not sensitive to the high autocorrelation peak seen in figure 2.6. This is due to the fact that the OLS and Hill estimator look at left tails only and are not symmetrical as is the kurtosis.

The stability of both estimators are shown in figures 4.7 and 4.8.

The estimate of a tail index above 2 is not to be expected as it would imply a thin(ner) tailed distribution. From both figures I conclude heuristically that for the highest frequency the tails are very fat (or kurtosis very high) and becomes less fat very fast with decreasing frequency towards normal tails.

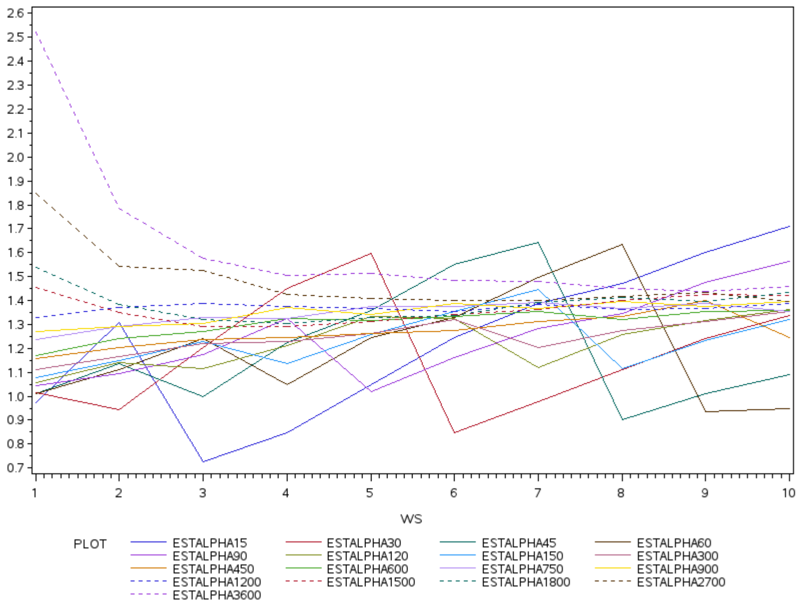


Figure 4.4: Tails index vs window size (%) for KPN estimated using the Hill method for different frequencies.

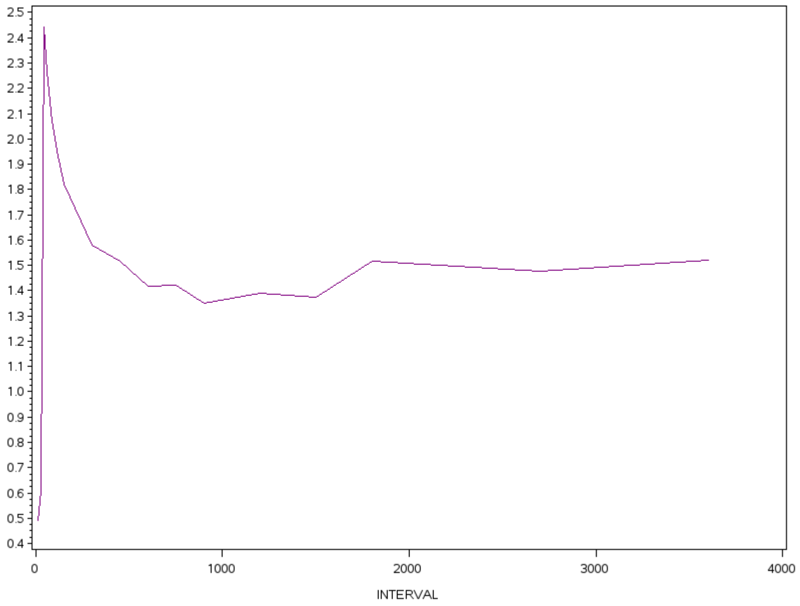


Figure 4.5: Tails index vs frequency for Pccw estimated using the OLS method for a window size of 10%.

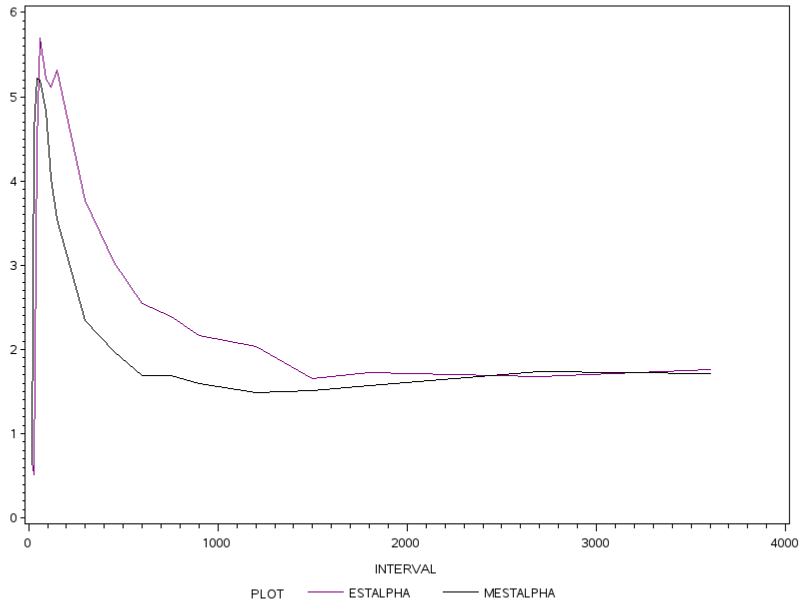


Figure 4.6: Tails index vs frequency for Pccw estimated using the Hill method for a window size of 10% and a mean over all window sizes between 1% and 10%.

Both estimators show to same curve: they start at $\alpha \approx 0.5$ for the highest frequencies, then rapidly rise to $\alpha \gg 2$ and finally decrease to values of α between 1.5 and 2.

The stability of both estimators is shown in figures 4.7 and 4.8. Again, notice the the choppy behavior in figure 4.8, which is even more extreme than for KPN in figure 4.4. This is because the tick size of Pccw is larger.

Stable distributions These findings lead to the conclusion that the distribution of KPN and Pccw are certainly not stable. Especially for the higher frequencies, the return distribution show very fat tails. For the lower frequencies, the tail index seem to flatten against frequency. For the lowest frequencies the tail index seems to be quite stable. Tick size and illiquidity both heavily influence the stability and therefore reliability of the Hill and OLS estimators.

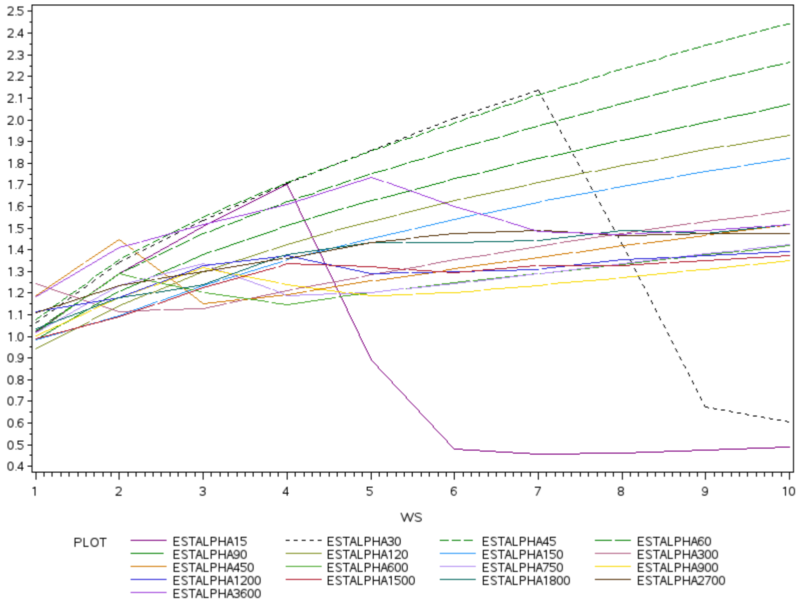


Figure 4.7: Tails index vs window size (%) for Pccw estimated using the OLS method for different frequencies.

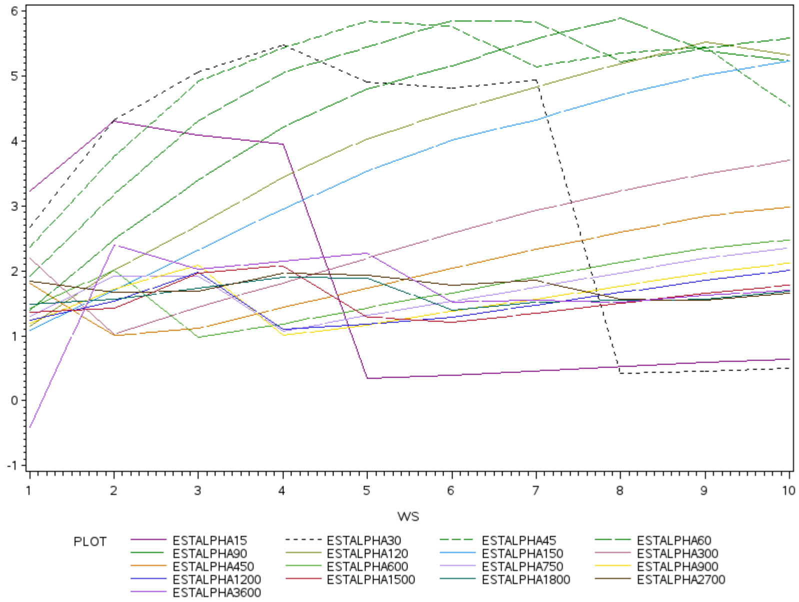


Figure 4.8: Tails index vs window size (%) for Pccw estimated using the Hill method for different frequencies.

Chapter 5

Simulation of the tails

The analysis of tail behavior of aggregated high frequency returns of both KPN and Pccw give us some interesting insights.

First of all, it is very hard to get a good estimate of the fatness of the tails of a distribution using a tail index. The estimates of α are quite unstable and not very reliable. This counts for the Hill estimator in particular. However, both the OLS estimator and Hill estimator are comparable if we look at the shape and level of the curve of tail index vs frequency.

Illiquidity The shape of the curve of α against the level of aggregation (the frequency) of KPN and Pccw show some commonalities. The highest frequencies show very high kurtosis and low tail index, both implying very fat tails. The reason for this is to be found not in the tails, but in the peaks of the distributions. Specifically, kurtosis as a estimator of tails, is very sensitive to a large number of 0 returns, and may not be appropriate to measure tails as such. This is what I refer to as illiquidity. This means that the estimation time scale is too small compared to the time scale of the pricing process.

The effect of illiquidity on kurtosis is the following.

1. Increasing or decreasing liquidity makes the distribution behave more respectively less like a Bernoulli distribution.

   The excess kurtosis of a Bernoulli distribution is given by

   $$\kappa_B = \frac{6p^2 - 6p + 1}{p(1-p)} \tag{5.1}$$

   which increases with very high and very low values of p and reaches its minimum at $p = 1/2$ for which $\kappa_B = -2$. This implies a *platykurtic* distribution. However, as p increases or decreases, so does the kurtosis.

   If the number of trades per second, minute, hour or day is increased, the chance increases that the price change over some interval smaller than the time between two prices, is zero (as there is no actual price update in that interval). This may be modeled by an increasing chance p.

Tick size The difference between the two stocks is that the curve of tail index vs frequency of Pccw shows a high peak just after the highest frequencies. The curve of KPN does not show such behavior. This is due to a different tick size (relative to the price level) for these two stocks. The effect of tick size on kurtosis is as follows.

2. Increasing or decreasing tick size makes the distribution behave more respectively less behave like a Bernoulli distribution as well. Higher tick size is the equivalent of increasing $p > 1/2$ in which p models the chance that price P_t rounds at price P_{t-1}, i.e. p models the chance that the return is zero.

In order to get a clear picture of the effect of tick size and illiquidity on the tails (and proof of the above hypotheses), I use simulation and create a dataset of high frequency returns.

5.1 Simulation in detail

The simulation is aimed to create a dataset that has approximately the same properties as the high frequency data of Pccw from January 1, 2010 through June 30, 2010. The algorithm is as follows.

1. Generate an equally spaced random time series S_t with t in seconds. S_t models the quasi-continuous price process. Let $S_0 = 1$ and S_t be a random walk given by

$$\log S_t = s_t = \log S_{t-1} + \epsilon_t \tag{5.2}$$

in which the shocks ϵ_t are normally distributed, i.e. $\epsilon_t \sim N(0, \sigma^2)$ with $\sigma^2 = 0.0001$. The variance σ^2 is chosen such that it approximately matches the unconditional variance of all log price-to-price changes of Pccw between January 1, 2010 through June 30, 2010, linearly adjusted to per-second price changes. The value of t runs from 0 through to 120 (days) x 4 (hours) x 60 (minutes) x 60 (seconds) which matches half a year of market hours of Pccw.

2. Let P_t model the price realization process (ignoring the tick size). P_t is updated with a snapshot (copy) of S_t, taken approximately every x seconds. The value of x is chosen to be $x = 45$ which matches the unconditional average time between two consecutive prices of Pccw between January 1. 2010 through June 30, 2010. This parameter models the liquidity of the stock. A trade takes place every second t that some random uniformly distributed variable $u \sim U(0, 1)$ is smaller than $1/x$. Consequently, the time series P_t is unevenly space through time just as real high frequency data.

3. Let $\tilde{P}_t$ be the price P_t rounded to the nearest tick below or above P_t. The tick size is set to 0.01 and matches the tick size of PCCW. The price is rounded to the nearest tick below if some random uniformly distributed variable $v \sim U(0,1)$ is smaller than a threshold $\tau = (P_u - P_t)/(P_u - P_l)$ in which P_u and P_l are the price P_t rounded to the nearest tick above respectively below P_t. This implies that the chance that P_t is rounded to a tick above, increases when P_t is closer to the price rounded to a tick above. This models the fact that buyers (sellers) are probably more (less) willing to pay a tick if the underlying (real) price is closer to that tick. The series $\tilde{P}_t$ models the actual trades.

4. Calculate m second interval log returns $r_{m,t}$ and $\tilde{r}_{m,t}$ from the price series P_t and $\tilde{P}_t$. If there is no price P_t (or $\tilde{P}_t$) at time $t = m$, use the latest available price. To make the analysis comparable with the analysis of the real tick data of KPN and PCCW, let m run from 15 through 3600 and start the intervals at $t = 0, (1/3)m, (2/3)m$.

5. Calculate the kurtosis and tail index (using the Hill estimator and the OLS estimator).

6. Repeat the previous steps $N = 25$ times and calculate the mean and standard deviation of the estimates. The number $N = 25$ is not that big, but given the size of one run of the simulation ($\approx 2M$ observations), larger would be impractical as regards to computer memory, disk space and computing time. It is large enough to give a good impression and make qualitative statements about the results.

5.2 Results

Figure 5.1 shows a realization of the price series P_t and $\tilde{P}_t$ for $0 \leq t \leq 4 \times 3600$ (4 hours or 1 business day). To illustrate that this is quite comparable with one business day of PCCW, see figure 2.8.

Kurtosis Figure 5.3 shows the kurtosis vs frequency of both the unrounded and rounded log returns. Both curves differ significantly in level but not in shape. Tick size does influence the kurtosis estimates as it lowers the level of the curve and makes the return distribution less fat tailed. This becomes more clear if we look at figure 5.2.

To answer the question why the bottom distribution is less fat tailed than the top distribution, we gave to consider kurtosis actually is. [Taylor and Xu, 1997a] tell us that kurtosis basically represents a movement of (probability) mass in the distribution that does not affect variance. If mass is moved from the shoulders of a distribution to the tails, the distribution will not have more positive kurtosis per se. This movement will simply increase the variance. The kurtosis will increase however, as mass is moved from the shoulders to the center as well.

If we start with the bottom distribution in figure 5.2 and move mass from the shoulders (which in this distribution are almost identical to the tails) to the tails

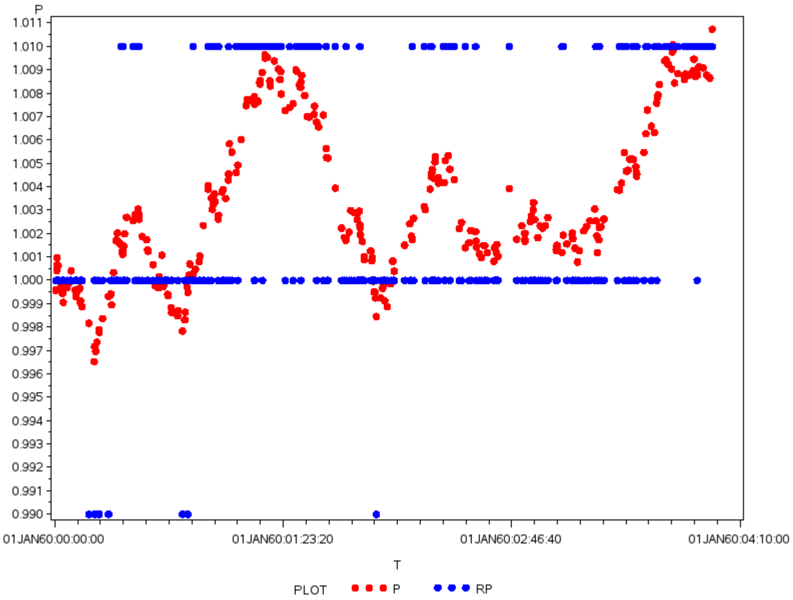


Figure 5.1: Four hours of simulated tick data. The series labelled p in red is the unrounded price realization process P_t and the series labelled rp in blue is the rounded price realization process $\tilde{P}_t$.

and the center, we will get the top distribution in that figure. This distribution will have a larger kurtosis.

The theoretical kurtosis, based on the assumption of i.d.d. returns, clearly explains the shape of both curves. This is expected as the price process S_t underlying P_t is a random walk and the log returns are i.i.d..

Tail index Figure 5.4 shows the tail index vs frequency of both the unrounded and rounded log returns, estimated using the OLS method. The curve of the unrounded prices series tells a similar story as the curve of the kurtosis in figure 5.3: for the highest frequency the tail index is at its lowest, so the tails are at its fattest, for lower frequencies the tail index increases and finally, for the lowest frequencies, the curve flattens at $\alpha \approx 2$.

The curve of the rounded price series tells a different story: for the highest frequency the tail index is at its lowest, but rapidly increases for the middle frequencies and for the lower frequencies, the tail index decreases and finally, for the lowest frequencies, the curve flattens below 2. This can be explained by noting that for the highest frequencies ($< 45s$) illiquidity is predominant, which means that there is a high peak at 0 return. As the frequency decreases, the tick size gets more dominant and makes the distribution look more like a Bernoulli distribution with a very low kurtosis. This peak at frequencies just after the highest frequencies is exactly what is seen for Pccw in figure 4.5 and 4.6.

Figure 5.5 shows the tail index vs frequency, estimated using the Hill estimator. The curves are very comparable to the curves estimated using OLS in figure 5.4,

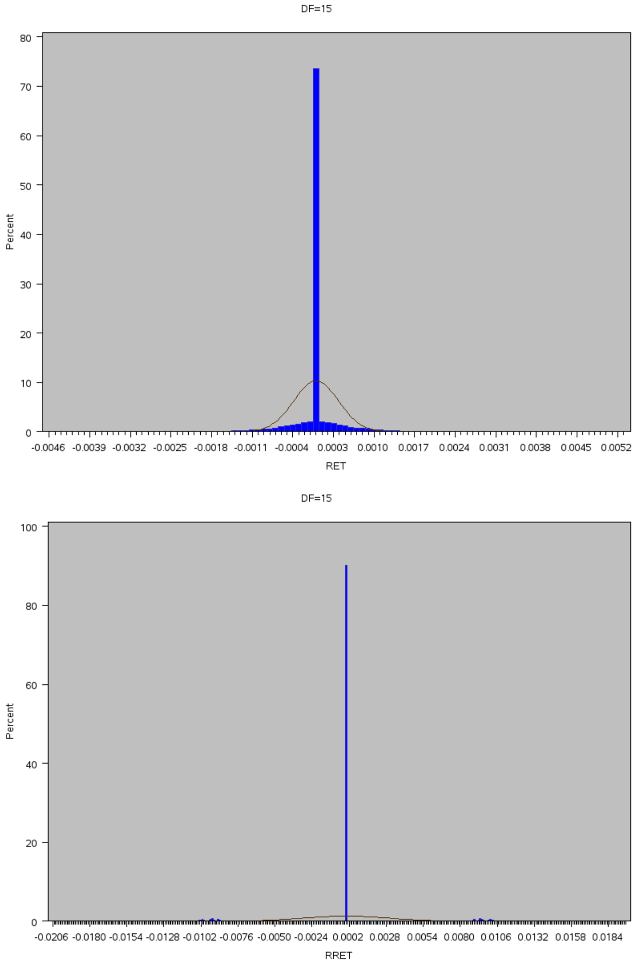


Figure 5.2: Simulated return distributions of a unrounded price series (top) and a rounded price series (bottom) for one 15 second intervals. The bottom distribution is less fat tailed than the top distribution.

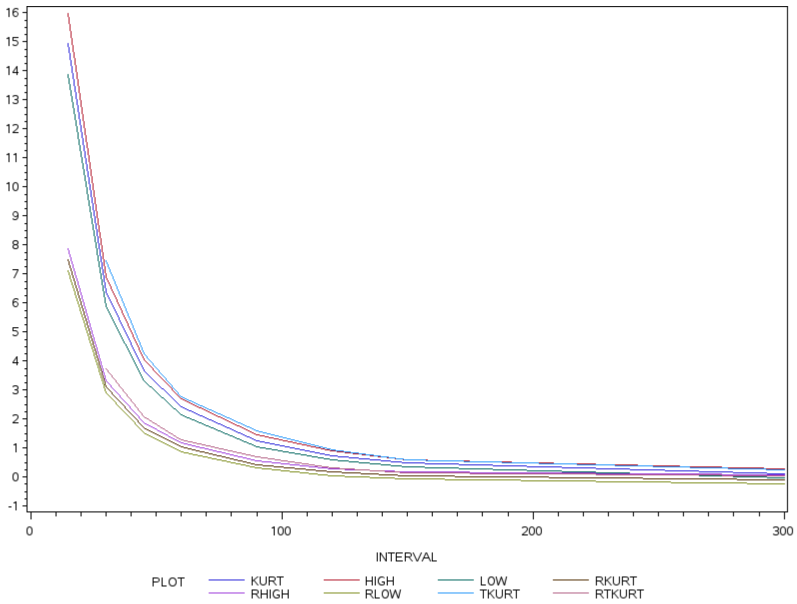


Figure 5.3: Kurtosis vs frequency of simulated tick data. The series labelled Kurt, High and Low show the kurtosis with upper and lower band of the unrounded log returns series r_t. The series labelled Rkurt, Rhigh and Rlow show the kurtosis with upper and lower band of the rounded log returns series $\tilde{r}_t$. The series labelled Tkurt and Rtkurt are the theoretical kurtoses based on the assumption of i.id. log returns.

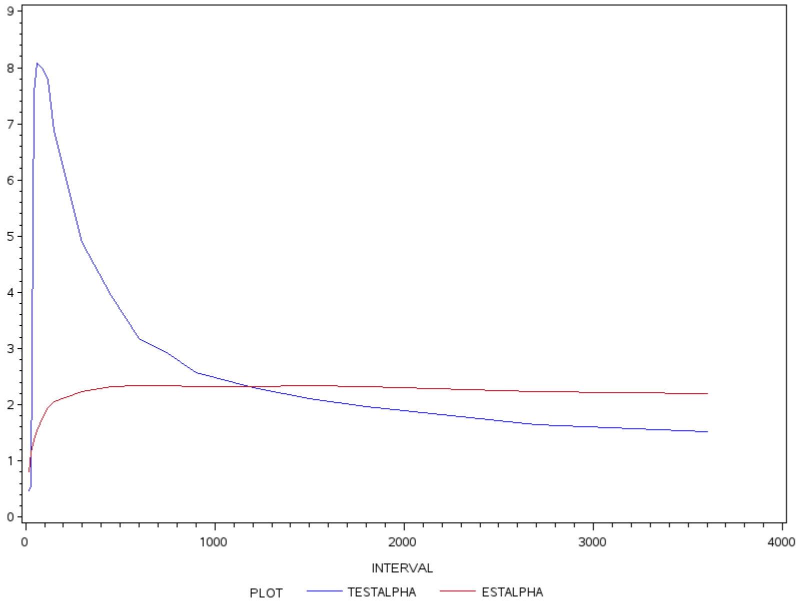


Figure 5.4: Tail index vs frequency of simulated tick data estimated using the OLS method for a window of 10%. The series labelled ESTALPHA shows the tail index of the unrounded log returns series r_t. The series labelled TESTALPHA shows the tail index of the rounded log returns series $\tilde{r}_t$.

although the peak in the tail index is a little higher with the Hill estimator.

Real high frequency data Simulation shows that tick size and illiquidity are partly able to explain tail behavior under temporal aggregation seen in real high frequency data. The fat tails (high kurtosis, low tail index) at the highest frequencies, the decreasing kurtosis (increasing tail index) at the lower frequencies and the high tail index peak at the frequencies just after the highest frequencies. Simulation does not explain the relatively high kurtosis at the lowest frequencies seen in figure 2.4 and 2.7 which is probably due to autocorrelation effects which are not considered in this simulation.

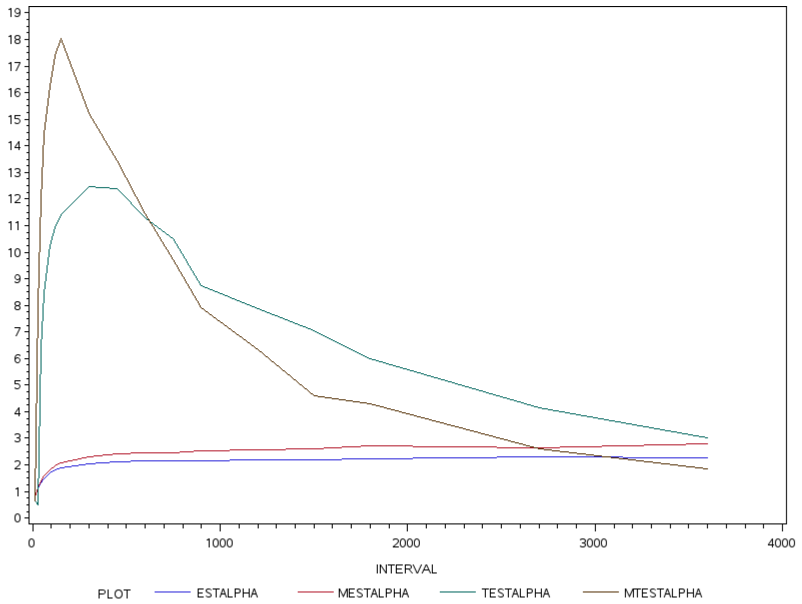


Figure 5.5: Tail index vs frequency of simulated tick data estimated using the Hill estimator for a window size of 10% and a mean over all window sizes between 1% and 10%.The series labelled Estalpha (window of 10%) and Mestalpha (mean over the windows) show the tail index of the unrounded log returns series r_t. The series labelled Testalpha and Mtestalpha show the tail index of the rounded log returns series $\tilde{r}_t$.

Chapter 6

Conclusion

In this thesis, the research objective was to write a financial case study of the high frequency return distributions of two comparable but different stocks, Dutch KPN and Hong Kong's Pccw, both telecommunication firms traded on developed market places, Euronext Amsterdam and Hong Kong Stock Exchange. First of all, I elaborately analyzed the high frequency data and discussed data quality and filtering. Secondly, I looked at the (left) tail behavior under temporal aggregation, using simple kurtosis and the tail index of stable distributions. Finally, I used simulation to show that the basic assumptions about market microstructure largely explain tail behavior under temporal aggregation for high frequency data.

High frequency data quality and filtering High frequency data has been hot in econometric research ever since the (real) rise of the computer. Many practitioners use high frequency data to speculate. Many researchers have used high frequency data as a source of information and for purpose of lower frequency applications.

I discussed the difficulties that arise with high frequency data in the first part of this thesis. I addressed data quality and filtering and stylized facts of two high frequency datasets. Rigorous data analysis is a prerequisite for proper research.

Tail behavior under temporal aggregation In the second part of this thesis, I examined the tail behavior of the high frequency return distribution of the two stocks. The importance of tail behavior, and other aspects of return distributions, under temporal aggregation lies in the heart of the hypothesis of scalable markets. As a 'measure' of tails, I used kurtosis and tail index which arises with stable distributions. In effect, I put the CLT vs the Generalized CLT.

Kurtosis and (fat tailed) stable distributions are mutually exclusive. The kurtosis of stable distributions is only defined (and zero) if the tail index α is 2. This is equivalent to a normal distribution. Another important difference is that kurtosis is a measure of both right and left tail, and the tail index is only for the right or the left tail. In this thesis I used both kurtosis and tail index to estimate the (left) tails. It was not a priori clear whether the kurtosis is defined and the return distribution is stable.

First, I measured the tail behavior under temporal aggregation using (sample) excess kurtosis as an estimator of the curve of tails vs measurement frequency. At the highest frequencies this results in a very high kurtosis, which decreases with decreasing frequency. If the tail index is used as an estimator, the results are similar for the highest frequencies.

However, the kurtosis for lower frequencies is relatively and absolutely higher than a normal distribution and the CLT suggest. The curves of kurtosis vs frequency for KPN and Pccw are shown in figures 2.4 and 2.7. The theoretical lines in these figures are only valid if finite variance is assumed. The fact that the theoretical curve does not explain the empirical curves, suggests that the variance is in fact not finite and that the high frequency return distributions are not in the domain of attraction of the normal distribution (which is what the CLT tells us). As an alternative one could assume a fat tailed distribution such as the stable distribution. Anyway, the tail index can be used as an measure of the tails.

The tail index is estimated using the Hill estimator and the OLS estimator. Both estimators show unstable results, especially for the lowest frequencies where the samples are too small to yield reliable results. Nevertheless, the curve that describes the relation between measurement frequency and the tail index show a corresponding shape: low tail index at the highest frequencies, followed by an increasing tail index with decreasing frequency. For Pccw we see an sharp peak in the tail index for the frequencies just below the highest frequencies (15-30 seconds).

I reject the stable hypothesis for both return distributions, because if the hypothesis would hold, it would imply a flat curve of tail index vs frequency. This is not observed. These findings are in contrast with prior research, such as by [Pictet et al., 1996]. They find stable tail indices ($>> 2$) for foreign exchange markets and interbank markets of cash interest rates for frequencies between 30 minutes and 6 hours. The markets in this thesis however showed even for the lower frequencies ($>$ 30 minutes) unstable tail indices. The main difference between that research and my research, is that they looked at 30 minute to 6 hour intervals, which is considerably longer that 15 seconds to 1 hour intervals. Rejection of the stable hypothesis implies that many risk models amongst others cannot be used to scale high frequency data to daily, monthly or yearly data.

For most frequencies, except for the the middle frequencies of Pccw, I find a tail index below 2. This indicates infinite variance. Amongst others, infinite variance implies that many popular regression models cannot be used as they require finite variance.

An alternative for the stable hypothesis, i.e. a flat tail index vs frequency curve, one could consider a parametric model for the tail index depending on frequency. You should always bear in mind however that tick size, illiquidity and other market microstructure characteristics heavily influence the analysis. This would require further research.

Illiquidity and tick size In the last part of this thesis, I used simulation to show that the market microstructure in terms of illiquidity (or inactivity) and tick size (or bid-ask spread) can largely explain the curve of tail index (and kurtosis) vs measurement frequency. I showed that the high tick size of PCCW explains the high peak in the tail index at the frequencies just below the highest frequencies. I made clear that *leptokurtosis* is as much about the peak as it is about the tails. Since kurtosis is symmetric in the tails and the tail index is not, kurtosis may give contra-dictionary results.

Further research Knowledge of and expertise in high frequency data may help any professional investor in understanding financial markets. Besides knowing what to do, is knowing what *not* to do as much important. Scaling markets is one such thing. TRANSTREND may use this research a starting point of more advanced data filtering and trading algorithms.

Bibliography

T. Andersen and T. Bollerslev. Heterogeneous information arrivals and return volatility dynamics: Uncovering the long-run in high frequency returns. *Journal of Finance*, 52(3):975–1005, 1997a.

T. Andersen and T. Bollerslev. Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance*, 4(2-3):115–158, 1997b.

L. Bachelier. *Théory de la speculation*. Gauthier-Villars, Paris, 1900.

D. Bartolomeo. *Fat Tails, Tall Tales, Puppy Dog Tails*, page 29. Newport RI, 2007.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

J. Chambers, C. Mallows, and B. Stuck. A method for simulating stable random variables. *Journal of the American Statistical Association*, 71:340–344, 1976.

Y. Chang and S. Taylor. Information arrivals and intraday exchange rate volatility. *Journal of International Financial Markets, Institutions and Money*, 13: 85–112, 2003.

M. Dacorogna, U. Muller, R. Nagler, R. Olsen, and O. Pictet. A geographical model for the daily and weekly seasonal volatility in the foreign exchange market. *Journal of International Money and Finance*, 12(4):413–438, 1993.

K. Eisenhardt. Building theories from case study research. *The Academy of Management Review*, 14(4):532–550, 1989.

R. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

T. Falkenberry. High frequency data filtering. *Technical report, tickdata.com*, 2002.

E. Fama. Mandelbrot and the stable paretian hypothesis. *Journal of business*, 36:420–429, 1963.

E. Fama. The behavior of stock market prices. *Journal of business*, 38:34–105, 1965.

G. Garonne, R. Marchionatti, and R. Bellofiore. Keynes on econometric method. a reassessment of his debate with tinbergen and econometricians, 1938-1943. CESMEP Working Papers 200401, University of Turin, 2004. URL `http://ideas.repec.org/p/uto/cesmep/200401.html`.

B. Gnedenko and A. Kolmogorov. *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley, Reading, Massachusetts, 1954.

C. Goldie and R. Smith. Slow variation with remainder: Theory and applications. *Quarterly Journal of Mathematics*, 38:45–71, 1987.

L. Gulko. The entropy theory of stock option pricing. *International Journal of Theoretical and Applied Finance*, 2(3):331–356, 1999.

B. Hill. A simple general approach to inference about the tail of the distribution. *The Annals of Statistics*, 3:1163–1174, 1975.

H.A. Keuzenkamp. Keynes and the logic of econometric method. Discussion Paper 1995-113, Tilburg University, Center for Economic Research, 1995. URL `http://ideas.repec.org/p/dgr/kubcen/1995113.html`.

A. Lau, H.S. Lau, and J. Wingender. The distribution of stock returns: New evidence against the stable model. *Journal of Business and Economic Statistics*, 8(2):217–224, 1990.

H.S. Lau and J. Wingender. The analytics of the intervaling effect on skewness and kurtosis of stock returns. *Financial Review*, 24(2):215–234, 1989.

P. Lévy. *Calcul des Probabilité.* Gauthier-Villars, Paris, 1925.

S.-J. Lin, J. Knight, and S. Satchell. *Financial markets tick by tick*, chapter Modelling Intra-day Equity Prices and Volatility Using Information Arrivals - A Comparative Study of Different Choices of Informational Proxies. John Wiley & Sons, Chistester, 1999.

B. Mandelbrot. The variation of certain speculative prices. *Journal of business*, 36:394–419, 1963.

S. Mittnik and M. Paolella. A tail estimator for the index of the stable paretian distribution. *Communications in Statistics - Theory and Methods*, 27:1239–1262, 1998.

S. Mittnik and M. Paolella. A simple estimator for the characteristic exponent of the stable paretian distribution. *Mathematical and Computer Modeling*, 29:161–176, 1999.

U. Muller, M. Dacorogna, R. Olsen, O. Pictet, M. Schwarz, and C. Morgenegg. Statistical study of foreign exchange rates, empirical evidence of a price change scaling law, and intraday analysis. *Journal of Banking and Finance*, 14:1189–1208, 1990.

J. Nolan. *Stable Distributions - Models for Heavy Tailed Data.* Birkhäuser, Boston, 2010. In progress, Chapter 1 online at academic2.american.edu/∼jpnolan.

O. Pictet, M. Dacorogna, and U. Muller. Heavy tails in high-frequency financial data. *Working paper, Olsen and Associates*, 1996.

S. Rachev and S. Mittnik. *Stable Paretian Models in Finance.* John Wiley & Sons, Chichester, 2000.

R. Skidelsky. *The return of the master.* Public Affairs, New York, 2009.

S. Taylor. *Asset Price Dynamics, Volatility, and Prediction*, chapter 12: High-Frequency Data and Models. Princeton University Press, Princeton, 2005.

S. Taylor and X. Xu. On the meaning and use of kurtosis. *Psychological Methods*, 2(3):292–307, 1997a.

S. Taylor and X. Xu. The incremental volatility information in one million foreign exchange quotations. *Journal of Empirical Finance*, 4(4):317–340, 1997b.

R. Weron. On the Chambers-Mallows-Stuck method for simulating skewed stable random variables. *Statistics & Probability Letters*, 28:165–171, 1996.

E. Zivot and J. Wang. *Modeling financial time series with S-PLUS, Volume 13*. Springer Science+Business, New York, 2006.