ON THE MODELLING OF ULTRA HIGH FREQUENCY FINANCIAL DATA ON THE JOHANNESBURG STOCK EXCHANGE

by

HERCULAAS FREDERIK VAN ROOY

THESIS

submitted in fulfillment

of the requirements for the degree

PHILOSOPHIAE DOCTOR

in

MATHEMATICAL STATISTICS

in the

FACULTY OF SCIENCE

at the

UNIVERSITY JOHANNESBURG

PROMOTER: PROF. F. LOMBARD

MAY 2006

Acknowledgements

An endeavour of this nature could never have been completed without ample guidance, assistance and encouragement. In this regard, I would like to thank the following persons and institutions:

My promoter, Professor Freek Lombard, who knows (virtually) everything and finds joy in sharing. Thank you for sharing some of your knowledge and plenty of your time and wisdom. I am indeed fortunate to learn from a scholar of your calibre.

My sincere appreciation to my school teachers at Moorivier Primary and Potchefstroom Gimnasium and lecturers at the former Potchefstroom University for Christian Higher Education for your contributions in my growth and development.

My colleagues in the Department of Statistics at University of Johannesburg, especially Urbanus Human and Wickes Robbertse, who spent many hours listening to my rambling to help me clarify my algorithms and arguments.

Many friends for support and encouragement throughout.

My late grandfather, Bertus van Rooy, who showed a keen interest in my studies and shared my passion for mathematics. He would have enjoyed this. Equally, the encouragement and interest of my other grandparents remain a fond memory.

My parents-in-law, Fanie and Rina van Wyk, my brothers and sisters and their spouses, Bertus and Susan, Bennie and Nicola, Annerie and De Wet, and Sinette and Willem, as well as my brothers- and sister-in-law, Fanus, Madeleine and Maarten, for your encouragement and support.

A special word of thanks to my parents, Herrie and Jacoba van Rooy, for all the effort you put in to my education, both as a person and a student. Thank you for all the time you invested and the example you set.

My wife, Elize, who by now probably knows more about statistics than many of my students. I can hardly express my gratitude in one paragraph. Thank you for bearing with me through late nights and the many hours of being a computer widow. Thank you for your love, support and companionship.

God in heaven.

Herrie van Rooy

Johannesburg

August 2006

Table of contents

Abstract

This thesis considers the modelling of ultra high frequency (UHF) financial data from South African markets. The approach to be taken is that such irregularly spaced data can be viewed as a realization of a marked point process. We propose a statistical model that incorporates both the unequally spaced transaction times (the points) as well as the movements of the associated returns (the marks). In all data sets investigated, no change in the value of the mark accounts for more that half the observations. If "no change" is considered as the censoring of some underlying process, we can explicitly model both the censoring of marks and the underlying process by utilizing methods for Markov chains and missing values.

All models considered hitherto in the literature assume homogeneity of structure within a UHF data set. Data analyses indicate strongly that such an assumption is not justified. The proposed model aims to exploit this observation. The diurnal (time of day) effect is a form of non-stationarity commonly found in UHF data sets. We show that the method currently considered standard practice is inadequate and we will propose modifications of it. Consideration is given to the classification of heterogeneous subsets that arises naturally in UHF data, for instance daily subsets of a UHF data set. We find evidence in support of some market microstructure theories, but no theory is supported by all data sets considered.

We pay attention to technical issues surrounding the application of certain tests to large samples. As large samples are common in UHF data sets methods that are sensitive to large sample size, for example the Ljung-Box test, are not suitable.

Chapter 1

Review of relevant literature

Our aim is to find a model that suitably describes the financial point process first given the name ultra high frequency data (UHF data) by Engle (2000). Essentially this refers to the limiting case in econometrics where all transactions have been recorded. The analysis of UHF data is a relatively recent development as only now are we able to process the vasts amounts of data involved. Processing data sets in the order of tens or hundreds of thousands of transactions was once unthinkable, but in modelling the assets for which UHF data is recorded this is exactly the aim. While such a scenario could be considered asymptotic utopia it bears complications of its own.

To see why it is necessary to have models able to describe assets prices at a transaction level it is informative to consider the evolution of models for asset prices. Earlier studies in the pricing of derivatives, like the seminal paper by Black and Scholes (1973), assume that asset prices follow geometric Brownian motion. This is a very attractive assumption for various reasons, foremost its convenient mathematical properties. The downside of the use of geometric Brownian motion, and other diffusion models, has been widely studied, see for example Cont and Tankov (2004, Chapter 1) or Rydberg (2000). To mention but a few drawbacks, the assumptions of complete markets, continuity and stationary independent increments in the price process are contradicted by empirical facts.

This led to a body of research on the use of Lévy processes as models for assets; for an overview, see Cont and Tankov (2004) or Schoutens (2003). These models relax some of the assumptions that make the use of diffusion models so restrictive. Amongst others, Lévy models now provide for incomplete markets and jumps in the asset price. Heavy tails in the distribution of increments are incorporated generically in Lévy models, rather than by the inclusion of nonlinear structures for volatility in diffusions. Hence, we obtain a less restrictive framework while still retaining some of the analytical tractability that makes diffusion models so attractive. However, the assumptions of stationary independent increments in the price process is still retained. Another serious drawback of Lévy models from a mathematical

finance perspective is that prices implied for derivatives are not unique, rather they depend on the choice of martingale measure.

To relax the assumption of stationary independent increments a more dynamic parameterization of asset prices is needed. Duan (1995) introduced the idea of using time series models for the pricing of derivatives by using a $GARCH$ process for asset returns. While this model sacrifices analytical tractability it is able to explain some of the systematic biases found in the Black-Scholes and other models. Duan, however, still retains the idea of sampling asset prices at fixed intervals.

The sampling frequency is what distinguishes the analysis of UHF data from other forms of econometric data. By including all transactions UHF data is inherently irregularly spaced. In econometric applications considered prior to the advent of UHF models (Engle and Russell, 1998) the sampling frequency was chosen to be a fixed interval, e.g. the daily closing price of a share, and data points could merely be indexed by integers. In contrast, the arrival times of transactions in an irregularly spaced setting needs to be modelled explicitly before, or in conjunction with, other attributes of interest. It is here where UHF models can contribute to mathematical finance by providing a realistic model for asset price processes. Another potential benefit from studies on UHF models relates to their application to market microstructure models, which we will consider later in this chapter.

For a more thorough understanding of the problem at hand it is necessary to review some of the research relevant to our study. It is not our aim to provide an exhaustive discussion, since extensive reviews are already available. We merely wish to point out a few relevant studies and provide the reader with references for more general discussions.

This chapter is divided into three sections. The first section considers duration modelling, the second includes prices and the last looks at studies related to market microstructure effects.

1.1 Duration modelling

In treating the intraday transactions authors initially ignored the irregular spacing by indexing transactions according to the order in which they arrived without explicitly accounting for calender time, see for example Hasbrouck (1991). This method of indexing is commonly referred to as transaction time. Transaction time, however, ignores possible information contained in intertrade durations and the implications that duration has for both parameters and models. For example, Easley and O'Hara (1992) found that time is endogenous to the price process. Hence, models for durations in calendar time were needed. We mention two frameworks for the modelling of intertrade durations in calender time, namely the autoregressive duration approach and the conditional intensity approach. Both these frameworks model stationary counting processes. However, authors are in agreement that the arrival times of transactions are non-stationary, see for example Hautsch (2004), Gourieroux and Jasiak (2001, Chapter 14), Andersen and Bollerslev (1998) and Engle and Russel (1998) amongst others. This includes well known causes of seasonality or, more precisely, diurnality like the "lunchtime effect" which states that trade durations tend to increase daily around lunchtime only to decrease again when traders "return from lunch". It is therefore necessary to detrend the data in some or other fashion. Engle and Russell (1998) proposed the use of a deterministic multiplicative diurnal effect as a function of the time of day. To preserve their notation, if t_i denotes the arrival time of the i-th trade, then $x_i = t_i - t_{i-1}$ is the associated intertrade duration. Under the multiplicative effect we have

$$x_i = \breve{x}_i s\left(t_{i-1}\right)$$

where $s\left(t_i\right)$ denotes the diurnal component and $\breve{x}_i$ denotes the diurnally adjusted duration. This basic approach is followed by Hautsch (2004), Veredas, et al. (2002), Gramming and Maurer (2000), Engle (2000), Andersen and Bollerslev (1998) and Engle and Russel (1998) amongst others. There are at least three methods proposed for the estimation of $s\left(t\right)$. Engle and Russell (1998) proposed that durations for individual days be superposed and $s\left(t\right)$

estimated by linear splines placing one node for each hour, on the hour, with the exception of the last hour which gets an additional node on the half hour. Andersen and Bollerslev (1998) proposed a Fourier series approximation, while Veredas, et al. (2002) estimate a semi-parametric form for $s(t)$ which is estimated jointly with other model parameters. Whichever method authors prefer, the important assumption inherent in all these methods is that the diurnal effect, $s(t)$, is homogeneous across trading days. This assumption will be subjected to scrutiny in proposing our approach to diurnal effect modelling (Chapter 2).

The first framework we mention is the autoregressive duration approach introduced by Engle and Russell (1998) with their autoregressive conditional duration (ACD) model. The aim is to find a dynamic parameterization for the conditional mean duration. Let

$$\psi_i \equiv \psi_i\left(\breve{x}_i, \breve{x}_{i-1}, \ldots, \breve{x}_1; \theta\right) = E\left[\breve{x}_i | \breve{x}_{i-1}, \breve{x}_{i-2}, \ldots \breve{x}_1\right].$$

Then the class of ACD models are parameterizations of the form

$$\breve{x}_i = \psi_i \varepsilon_i$$

where the ε_i are independent and identically distributed (iid) innovations with density $p(\varepsilon; \phi)$ and ϕ and θ are variation free parameter vectors, i.e. the parameter space does not depend on the data. Engle and Russell initially proposed an $ARMA$ specification for ψ_i and the exponential distribution for the ε_i. Extensive research has been done on this class to test for the presence of ACD and improve fit. Proposed improvements centre around either the specification of the density or the parameterization of the ψ_i. For an extensive review of different types of ACD models see Hautsch (2004, Chapter 5). The most significant drawback of ACD models is that they do not readily extend to a multivariate context, see Hautsch (2004, Chapter 2).

The specification of durations (intervals between points) is one of three possible methods to fully describe a point process, see Cox and Isham (1980, p. 11). The second is related to the joint distribution of points in arbitrary sets, while the third is related to the intensity

function. An approach based on the latter more readily extends to a multivariate context, hence the advent of the conditional intensity approach. Loosely speaking, the intensity function of a point process is the instantaneous probability density of an arrival at a specific point in time (or space) given some history filtration. To formalize, the following definition is taken from Brémaud (1981, p. 27):

Definition 1 *Let N_t be a point process adapted to some history $\mathcal{F}_t$, and let λ_t be a nonnegative $\mathcal{F}_t$-progressive process such that for all $t \geq 0$*

$$\int_0^t \lambda_s ds < \infty$$

almost surely. If for all nonnegative $\mathcal{F}_t$-predictable processes C_t, the equality

$$E\left[\int_0^\infty C_s dN_s\right] = E\left[\int_0^\infty C_s \lambda_s ds\right]$$

is verified, then we say: N_t admits the $\mathcal{F}_t$-intensity λ_t.

Russell (1999) proposed that an $ARMA$ type specification be used for the log intensity, which he named the autoregressive conditional intensity (ACI) model. For illustration we provide the full specification of the bivariate $ACI(1,1)$ model in his notation. The conditional intensity function for process k is given by

$$\lambda^k\left(t|\mathcal{F}_t\right) = \omega^k \exp\left(\boldsymbol{\phi}^k_{N(t)}\right)$$

where $\omega^k > 0$ and $\boldsymbol{\phi}^k_{N(t)}$ is a measurable function of the bivariate filtration of all past arrivals, $\mathcal{F}_t$. In a bivariate setting the outcome of a pooled process can only be one of two types, say a and b. Let us define a variable y_i to be an indicator of the outcome type. An $ARMA(1,1)$ type specification is then used for $\boldsymbol{\phi}^k_{N(t)}$, i.e.

$$\boldsymbol{\phi}^k_{N(t)} = \begin{cases} a_a \varepsilon_{N(t)-1} + B\boldsymbol{\phi}_{N(t)-1} & \text{if } y_{N(t)-1} = 0 \\ a_b \varepsilon_{N(t)-1} + B\boldsymbol{\phi}_{N(t)-1} & \text{if } y_{N(t)-1} = 1 \end{cases}$$

where ω, a_a and a_b are 2×1 vectors, B is a 2×2 matrix and the ε_i are iid unit exponential random variables. As with the ACD model other parameterizations of the conditional intensity have been proposed, see Hautsch (2004).

1.2 Modelling of prices

When considering the evolution of the price process there are two competing frameworks. The first models price changes in a continuous state space, while the second takes them to be discrete.

The use of $GARCH$ models for UHF data was proposed by Ghysels and Jasiak (1997) in their $ACD-GARCH$ model. The idea is to use the ACD model for the duration, while using a $GARCH$ specification for returns. To compensate for the fact that returns are not evenly spaced, they use the temporal aggregation formula of Drost and Nijman (1993)[1]. To remove autocorrelation Ghysels and Jasiak (1997) utilize an $AR(3)$ filter. It should be mentioned that their aim was the measurement of volatility rather that the possibility of prediction and hence such a two step procedure for the modelling of returns would be acceptable, but it is not optimal. The use of the temporal aggregation formula in an irregularly spaced setting was expanded on by Gramming and Wellner (2002)

In the same line of thought Engle (2000) proposed the so called $UHF-GARCH$. Again the ACD model is used for duration modelling. To "standardize" returns, Engle (2000) conditions on the history and current duration and divides the returns by the root of the current duration to get a return per time unit. This can the be modelled by a $GARCH$ process. As with $ACD-GARCH$ the aim was volatility estimation.

The motivation for using a discrete time model for returns stems from the use of the tick system, whereby prices are quoted in fractions of a currency unit, typically 1/8 or 1/16, rather than in decimals (cents). Authors[2] argue that most price changes fall within two or three ticks on either side of the current price and hence can adequately be described

[1] Drost and Nijman derived a formula for relating $GARCH$ parameters when the sampling frequency is changed. The approach is adapted to cater for irregularly spaced observations.

[2] See Russell and Engle (2005, 2002), Tay et al (2004) and Rydberg and Shephard (2003) amongst others.

by a discrete state space model. We will mention three models of this type, namely the autoregressive conditional duration - autoregressive conditional multinomial ($ACD-ACM$) model of Russell and Engle (2002, 2005), the price decomposition model of Rydberg and Shephard (2003) and the autoregressive conditional marked duration ($ACMD$) model of Tay, et al. (2004). As it happens, the price decomposition model and the $ACMD$ were fitted to NYSE data originating prior to the changeover from a minimum tick size of \$1/8 to \$1/16 on June 24, 1997. The $ACD-ACM$ model was fitted to data using tick sizes of \$1/16. It is worth keeping in mind that by January 29, 2001 the NYSE had converted all securities to the decimal system. The decimal system could have detrimental implications for a discrete space model[3].

The $ACD-ACM$ uses a two step procedure where, in the first step, an ACD model is fitted to durations. The possible movements of the price process is taken to be a multinomial variable with states corresponding to tick movements in the price. Russell and Engle propose an $ARMA$ type specification for the inverse logistic function of the probability vector associated with the multinomial cell probabilities at a given point in time. Their formulation also makes provision for the inclusion of additional information such as quotes or the timing of trades both lagged and contemporaneous. In the application of the model Engle and Russell use five cells corresponding to movements of zero, one and more that one tick moves to either side. Hence, the coverage of price movements is equivalent to one tick in the \$1/8 tick system. The inclusion of contemporaneous duration in the price process leads to some complications because the implication is that "... from an economic perspective, the $ACD-ACM$ model has an investor observing the time of a trade without its corresponding tick movement" (Tay, et al., 2004).

The price decomposition model of Rydberg and Shephard (2003) breaks down price movements into three multiplicative parts. The first part, A_t, indicates whether the price moves ($A_t = 1$) or not ($A_t = 0$). The other two parts are defined to take the value 0 if $A_t = 0$. If $A_t = 1$, the second part, D_t, indicates the direction of the price move, i.e. $D_t = 1$ if the price move is upwards and $D_t = -1$ if the price move is downwards. The third part,

[3]For a more detailed account on the evolution of the NYSE pricing system see Bacidore et al (2003).

S_t, indicates the size of the move measured in number of ticks, i.e. $S_t = 1, 2, \ldots$. The combined effect for a price move is then given by

$$A_t D_t S_t.$$

For A_t Rydberg and Shephard propose a generalized linear $ARMA$ model, for D_t an autologistic model and for S_t a negative binomial distribution. Rydberg and Shephard did not initially explicitly model duration. Rydberg and Shephard (2000) expands on the model by proposing the use of a Cox (doubly stochastic) process for the durations. An ACD model would be a specific case of a Cox process.

The $ACMD$ approach of Tay, et al. (2004) assumes that the movements in price are generated by three competing independent Poisson processes, one each for no move in the price and a one tick move in either direction. The given transaction is then recorded as an event of the type given by the Poisson process with the shortest duration. The three intensities are updated subsequent to every transaction. The implication for the durations process is an ACD type model with the constant intercept from the $ARMA$ specification for log mean duration (Russell and Engle) replaced by an intercept that varies according to previous tick movements. They also consider other renewal processes for the distributions of the three arrival processes. The price process is merely an accumulation of tick movements added to the known price at some earlier point in time.

To our knowledge no modifications to the discrete state price models have yet been proposed to account for the fact that the NYSE now also uses the decimal system.

1.3 Market microstructure theory

The theory of market microstructure is concerned with the frictions that cause the behaviour of asset prices to deviate from full-information (complete market) expectations. Microstructure literature provides theoretical arguments as to the possible nature of trading behaviour that could explain such deviations. An argument can be substantiated in one of two ways,

either by causal reasoning or by statistical testing. While causal reasoning might be ambiguous, hypothesis testing provides for more sound evidence. To be able to employ statistical testing, models have to be formulated in such a way that hypotheses surrounding theories can be tested in a formal manner. As some of these hypotheses could be tested with the help of a UHF model we consider theories for which our model could potentially provide scientific evidence either in support or to the contrary. The benefit of a UHF model lies therein that factors that are identifiable at a transaction level might be lost in aggregations when sampling is fixed at regular intervals, such as analyzing daily or weekly prices. It is not our intention to provide a comprehensive review of market microstructure, but rather to highlight theories of direct concern. For a comprehensive review see Madhavan (2000) or O'Hara (1995). Other studies involving the use of UHF data in testing market microstructure theories include Russel and Engle (2005, 2002), Tay, et al. (2004), Gramming and Wellner (2002), Engle and Dufour (2000), Engle (2000) and Engle and Russell (1998) amongst others.

Since the introduction of the asymmetric information model by Glosten and Milgrom (1985), models usually follow the assumption of two types of traders. The first type, called informed traders, have superior information unknown to the general public while the second group of traders, called liquidity traders, are motivated by non-information related concerns such as inventory control or portfolio rebalancing. The counterparty to a transaction is called a market maker. The market maker sets the price to compensate for the possibility that the trader has superior information. This compensation is incorporated in the bid-ask spread. The spread is thus, at least in part, a premium for the risk that the trader's information is superior to that of the market maker. When the market maker thinks it more likely that traders are of the informed kind, he increases the spread with the implication that volatility increases due to, amongst other things, bid-ask bounce[4] at a wider spread. Over time the market maker infers the private information of the informed traders from the order flow and sets the bid-ask spread to "centre" on the new true value. This provides a possible

[4]Bid-ask bounce refers to the a change in the price caused by a change in the party initiating the transaction, i.e. buyer initiated vs. seller initiated, rather than a change in the value of a company.

explanation for the presence of a bid-ask spread even in an efficient market for a risk neutral market maker with zero expected profit.

Diamond and Verrecchia (1987) argue that a proportion of informed traders might not already own the stock to which private information applies. If the news is good, this should have little effect, since one who expects the price to increase would purchase the stock. However, if the news is bad, and a trader doesn't own the stock, he would have to sell short. If a proportion of these traders are restricted from short selling, they will merely refrain from transacting in the share in question. Therefore, under the Diamond and Verrecchia (1987) model, lower trade frequency might be an indication of bad news, so that prices would tend to decrease in periods of relative quiet.

The Admati and Pfleiderer (1988) model distinguishes between two types of liquidity traders. "Discretionary" liquidity traders who have some control over the timing of their transactions within a certain time interval and "non-discretionary" liquidity traders who transact in a random fashion. They find that it is advantageous for "discretionary" traders to concentrate their trades together. As their model takes the arrival of private information to be exogenous and random, this implies that the proportion of informed traders should on average be higher when transactions occur less frequently, with the implication that higher average duration should imply higher volatility.

The Easley and O'Hara (1992) model also assumes the arrival of liquidity traders in a random fashion, but does not distinguish between different kinds of liquidity traders. They argue that informed traders will want to trade while their information has value. This will cause informed trades to be clustered together. The market maker knows this and will be mindful of the order flow to determine the likelihood of informed trading. They therefore suggest that periods of higher trade frequency would imply that informed trading, with the associated implications for spread and volatility, is then more likely.

Engle (2000) summarizes the implications of the Diamond and Verrecchia (1987) model as *no trade means bad news* and finds empirical evidence in support of this in the $UHF - GARCH$ context. This is also corroborated by Russell and Engle (2002, 2005) and Tay, et al. (2004) in the discrete price context.

Engle and Russell (1998) finds evidence consistent with the Easley and O'Hara (1992) model, i.e. high trade frequency implies informed trading. By implication this contradicts the Admati and Pfleiderer model which Engle (2000) summarizes as *slow trading means informed trading and high volatility.* These findings are also corroborated by Engle (2000), Russell and Engle (2002, 2005) and Tay, et al. (2004) in their respective contexts. In contrast, also in a $UHF - GARCH$ context, Gramming and Wellner (2002) reported: "It was found that lagged volatility has a significant negative impact on transaction intensity which is consistent with predictions from the Admati / Pfleiderer microstructure model."

1.4 Concluding remarks

In this chapter we highlighted some of the literature relevant to our study to obtain a better understanding of modelling in UHF data context. We mentioned a few intertrade duration models and possible applications thereof to the evolution of an asset price process at a transaction level. We argued that such a study could contribute to the literature on market microstructure by identifying dependencies on a single transaction level, which we will consider in Chapter 4. Another potential benefit from a UHF model is the ability to relax unrealistic assumptions of models used for asset prices in the pricing of derivatives. This remains an avenue for future investigation.

The remainder of the thesis is structured as follows: In Chapter 2 we scrutinize the deterministic multiplicative diurnal effect mentioned earlier in this chapter and formulate our own approach to the modelling of the diurnal effect. Chapter 3 outlines our approach to the modelling of durations and asset prices. In Chapter 4 we apply the formulated approaches to four shares on the Johannesburg Stock Exchange (JSE) and consider the implications for market microstructure hypotheses. Chapter 5 considers a few hypothesis tests utilized throughout the thesis. We either mention tests that are applied infrequently or raise specific issues with regards to know tests and mention or propose remedies. Amongst other things, we spend some time on goodness-of-fit testing for large samples. The reader may choose to peruse Chapter 5 before proceeding further or refer to it as the need arises.

We conclude with some final remarks in the Epilogue.

Chapter 2

On diurnal adjustment

This chapter considers diurnal adjustment of UHF data in some detail. It is known that the time of day plays a role in the frequency of trades on a stock exchange, see for example Gourieroux and Jasiak (2001, Chapter 14). The (possibly deterministic) effect that time of day has on trade frequency is referred to as the diurnal effect and it is this effect that we aim to adjust for. We will show that the method commonly used for diurnal adjustment (see Section 1.1) does not fully remove the time of day effect and can introduce bias. It will also be shown that the assumption of a homogeneous time of day effect is violated, thus affecting all models based on this premise.

We will argue that the diurnal effect can be removed by differencing and propose a method for modelling the diurnal effect to enable us to retain a handle on the effect while modelling the differenced series.

2.1 Problems with current practice

Engle and Russell (1998) proposed that diurnal adjustment be accomplished by fitting a multiplicative time of day effect to the intertrade durations of all trading days and then removing it from the data by division. The effect is implicitly assumed to be homogeneous over trading days. This can be seen in multiple ways. Firstly, in fitting the diurnal effect the durations of multiple trading days are superposed and used for the estimation of a single diurnal curve. Secondly, only the time of day plays a role in deciding which diurnal "factor" to remove from the observed duration. For convenience we preserve Engle and Russell's notation in the ensuing discussion . Formally, let t_i denote the time at which the i-th transaction took place, and let x_i denote the intertrade duration associated with t_i, i.e.

$$x_i = t_i - t_{i-1}.$$

Further, let $\phi(t_{i-1};\theta)$ denote the time of day effect associated with t_{i-1} containing parameters θ. Then the i-th diurnally adjusted intertrade duration $\tilde{x}_i$ is given by

$$\tilde{x}_i = x_i/\phi(t_{i-1};\theta).$$

In contrast to the assumption of homogeneity, it is well known that the number of transactions per day can vary substantially between trading days. On a day with higher transaction volume[5], the expected duration would typically be shorter than on a day with lower volume. We could then find a diurnal curve (effect) for each trading day which is the normal expected duration for a given time of that specific trading day. By implication the expected duration would have to differ for days with different transaction frequencies. For illustration, let the diurnal effect for the k-th day be given by $\phi_k(t_{i-1};\theta_k)$, where k denotes the day on which the particular transaction took place and θ_k the associated parameter vector. ϕ_k gives a true reflection of the expected duration for a specific day, as it takes into account that trading days have a stochastic number of trades. The function ϕ could be interpreted as a central location measure for the ϕ_k's. The following figure gives an indication of the relation of ϕ (the curve named $E\&R$ in Figure 2.1) with respect to two possible realizations for ϕ_k. We consider a day with high transaction volume, say curve $high$, and a day with low transaction volume, say curve low:

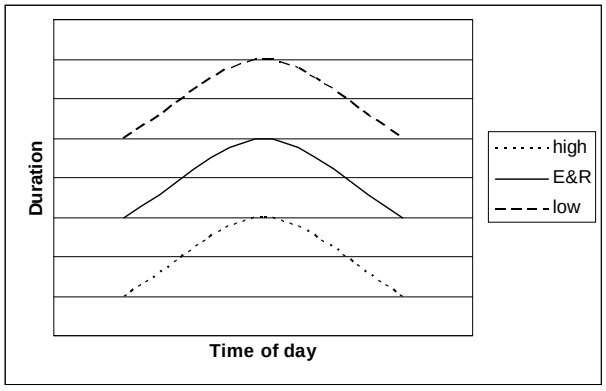


Figure 2.1. Illustration of the relative position of the diurnal curve for trading days with different levels of expected duration.

[5] With the term *higher transaction volume* we imply a higher trade frequency. This should not be confused with the number of shares changing hands in a specific transaction.

With these daily functions we can now once again adjust the durations for the diurnal effect, say

$$\begin{aligned}\dot{x}_i &= x_i/\phi_k(t_{i-1};\theta_k) \\ &= \tilde{x}_i\phi(t_{i-1};\theta)/\phi_k(t_{i-1};\theta_k) \\ &= \tilde{x}_i f(t_{i-1};\theta,\theta_k).\end{aligned}$$

In general, the function f would not be unity with the implication that $\dot{x}_i \neq \tilde{x}$. If the assumption of a multiplicative diurnal effect is correct, then

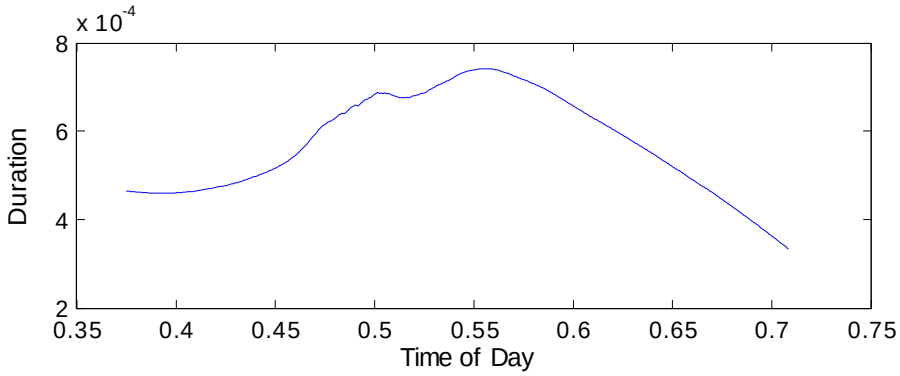
$$\phi_k(t_{i-1};\theta_k) = (1+\alpha_k)\phi(t_{i-1};\theta)$$

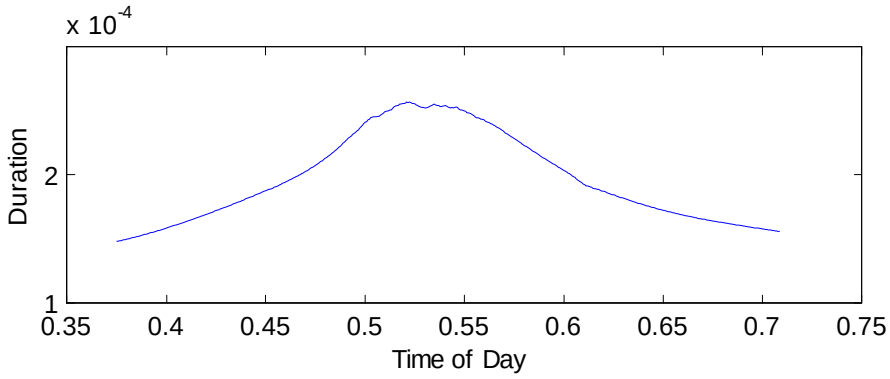
and

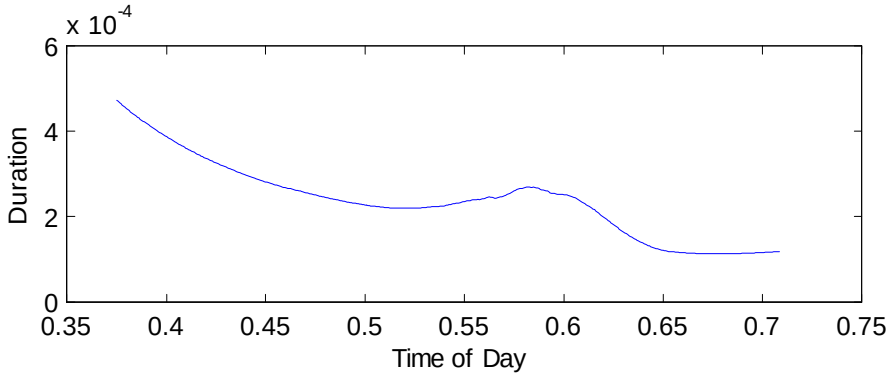
$$\tilde{x}_i = (1+\alpha_k)\dot{x}_i$$

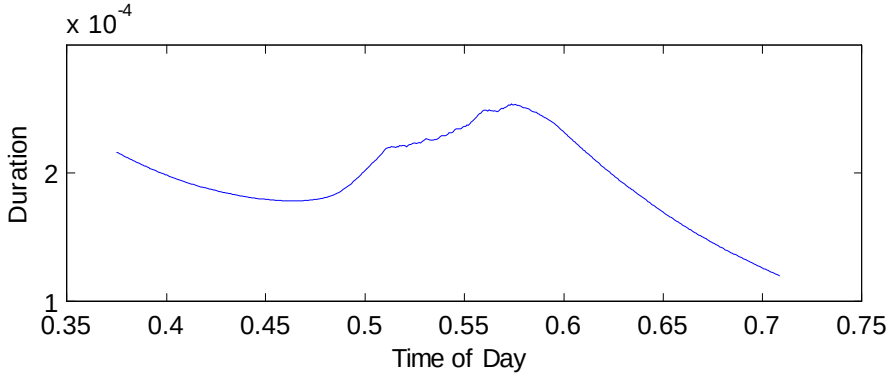
where α_k is the (stochastic) proportion by which ϕ_k differs from ϕ. Thus $\dot{x}_i$ is not the diurnally adjusted duration but the product thereof with a stochastic level component. The diurnal component has been reduced from a function to a stochastic level, but has not been removed altogether. Thus data transformed through the multiplicative procedure utilizing ϕ removes the diurnal component, but introduces (at least) a stochastic level component.

However, if the diurnal effect is not purely multiplicative, the stochastic function f will take on a more complicated form than merely a stochastic level, removing the adjusted value $\tilde{x}_i$ even further in complexity from $\dot{x}_i$. The following figures (Figure 2.2) were created by smoothing the durations of the first six trading days in November 2004 for Anglo American PLC using a LOESS (Cleveland, 1979) with the data span selected to include 60% of the data. Both duration and time of day are given as a proportion of a full day.


x 10^-4
8
6
4
2
Duration
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
Time of Day


x 10^-4
2
1
Duration
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
Time of Day


x 10^-4
6
4
2
0
Duration
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
Time of Day


x 10^-4
2
1
Duration
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
Time of Day

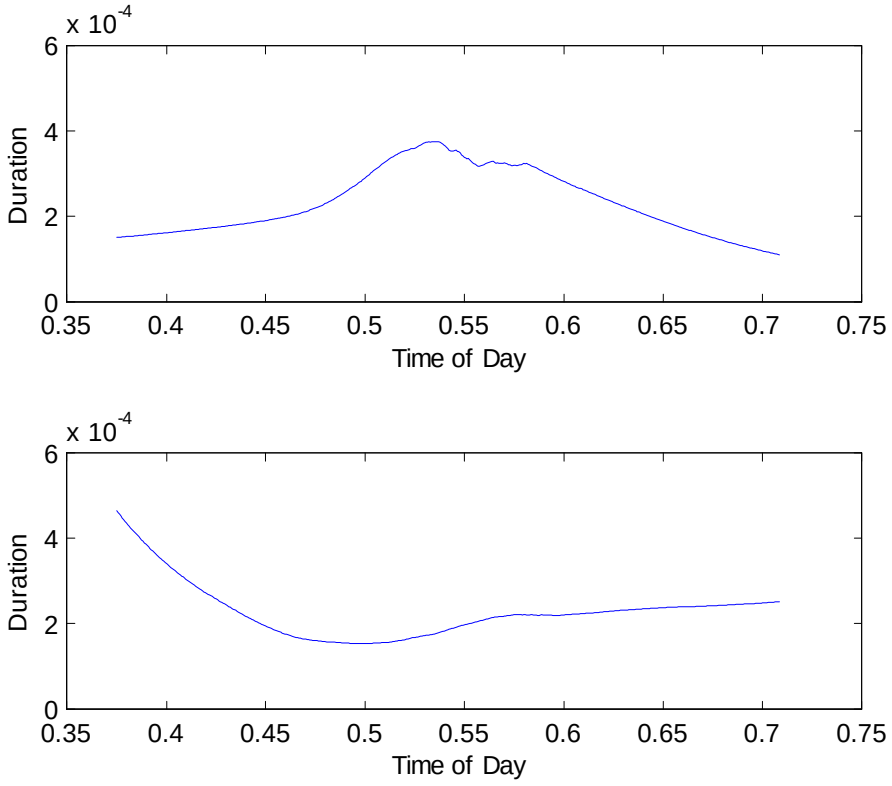


Figure 2.2.1. Six panels with smooths of the durations of the first six trading days in November 2004 for Anglo American PLC in ascending order. These clearly show that all diurnal curves do not have the same basic form.

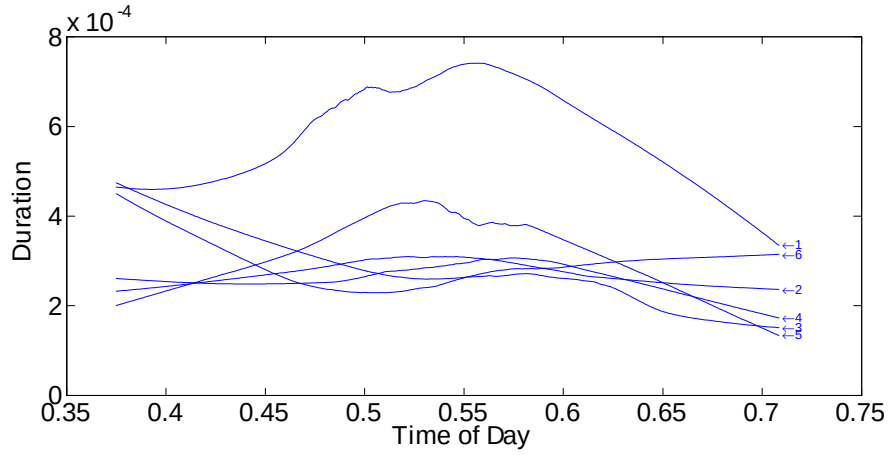


Figure 2.2.2. Superposed smooths of the durations of the first six trading days in November 2004 for Anglo American PLC.

While it could be argued that days 1, 2 and 4 have more or less the same form, notice the difference of scales on the duration axis illustrated more clearly by Figure 2.2.2. For these days possibly only a stochastic level component is required in the specification of the

stochastic function f. From the remaining days it is clear that the structure of the function of f is more complex than merely a stochastic level component. Notice, for example, that the curvature of day 6 is counter to expectation, i.e. in contrast to the "lunchtime" effect mentioned in Section 1.1 the trade frequency actually peaked around lunch. These smooths provide further evidence contradicting the assumption of a homogeneous multiplicative diurnal effect. Later in this chapter we will propose a method for the modelling of a heterogeneous effect, i.e. the ϕ_k.

2.2 Diurnal adjustment by differencing

Engle and Russell (1998) modelled the multiplicative diurnal effect, ϕ (see Section 2.1), with nodes fixed on the hour, except for the last trading hour, which had an additional node on the half hour. A value of ϕ for arbitrary time of day was then estimated by linear splines / interpolation. Subsequent to the discussion in the previous section we concluded that this method does not effectively remove the diurnal effect, irrespective of the method of estimation of ϕ (for examples see Section 1.1), since the homogeneous diurnal assumption is flawed. However, under the assumption that the trend can indeed be approximated locally by a straight line, removal can also be accomplished by first order differencing. In contrast, differencing is not sensitive to the heterogeneity of the trading days. Its validity can be expected to prevail as long as the assumption of a locally linear trend remains reasonable. We can assess the success of differencing as a method for diurnal adjustment in two ways. Firstly, test whether our new process is stationary in the mean, which would show that the diurnal component has been removed and, secondly, testing for over-differencing using an MA unit root test. For a discussion of the latter, refer to Section 5.3. To illustrate the success of differencing, the following two figures respectively show plots of the first 5000 durations for the share Anglo American PLC (ticker AGL, Figure 2.3.1) and the first 4000 durations for the share MTN Group Ltd. (ticker MTN, Figure 2.3.2) in November 2004, roughly corresponding to the first two trading weeks of that month.

Visually, at least, the proposition of constant mean seems realistic with the implication

that the diurnal effect is approximately removed. If our models (to follow in Section 4.2) adequately describes the differenced duration, this will provide more rigorous proof of the success of differencing in removing the diurnal effect. It is noteworthy that the differenced series shows volatility clustering commonly associated with $GARCH$ type processes. We will explore the phenomenon further in the next chapter.

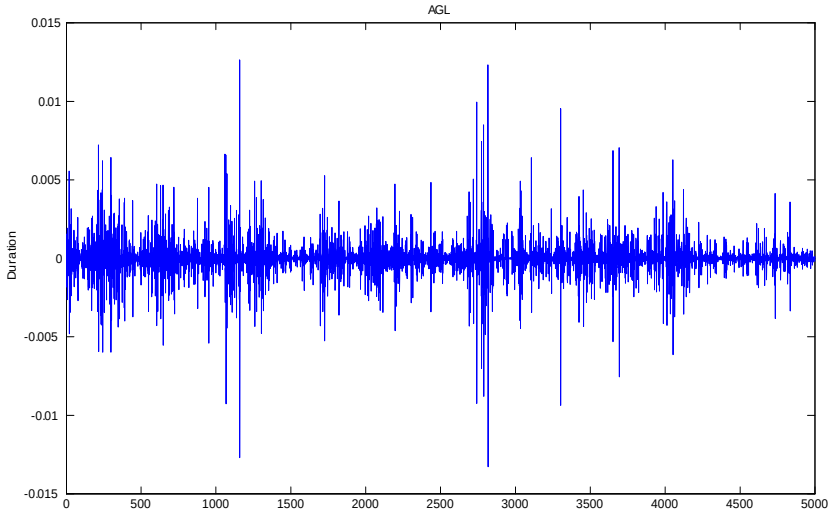


Figure 2.3.1. First differences of the first 5000 durations for Anglo American PLC in November 2004.

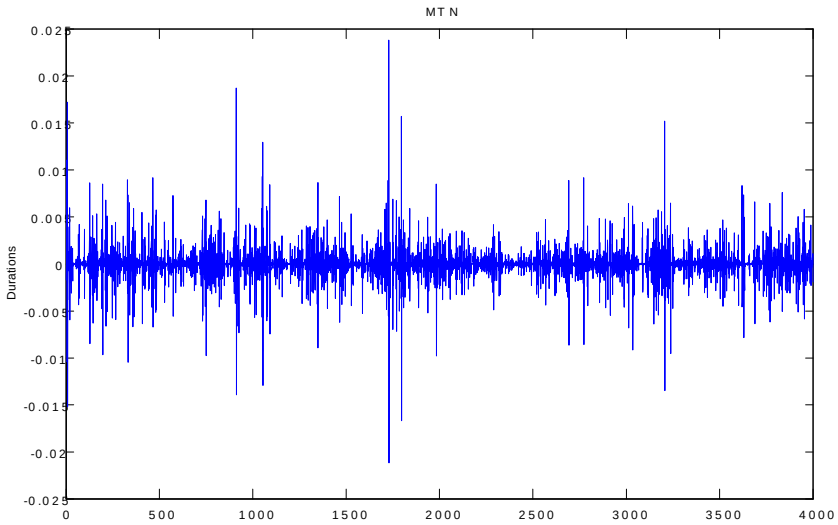


Figure 2.3.2. First differences of the first 4000 durations for the MTN Group in November 2004.

2.3 Modelling the diurnal effect

Our aim is to model the evolution of the diurnal effect over time, i.e. the evolution of the curve ϕ_k. The reason for doing so is that in differencing durations before modelling them, as we will do in Section 3.1, we sacrifice information about the diurnal effect. To retain a more accurate link with the real life situation knowledge of the diurnal effect is also required, hence the need for its modelling.

It is important to note the contrast between the superposition required for the methods of Section 1.1 and the individual diurnal curves we aim to model for each trading day, thereby relaxing the assumption of homogeneity. To this end two issues need to be addressed prior to modelling. Firstly, ϕ_k cannot be observed directly, thus we need to approximate it prior to any further analysis. Secondly, ϕ_k represents a distinct continuous curve for each trading day k, i.e. parametrically unspecified it represents a time series of infinite dimensional observations. Since we cannot model ϕ_k in an infinite dimensional space, we need to reduce the dimension of ϕ_k to make it tractable.

To approximate ϕ_k we will utilize a two step procedure. Firstly, we will run a robust smoother through the durations of each trading day. Secondly, we will fit a basis to the smooth. The reason for taking such a two step approach is to benefit from the advantages of both smoothers. In our first step we will apply the Robust Locally Weighted Regression algorithm proposed by Cleveland (1979). A local polynomial regression smoother does not reduce the dimension of our curve sufficiently to use "as is" in further analysis. However, it is more robust against outliers than fitting a basis globally. A peculiarity of duration data is that "outliers" to the large side, i.e. large intertrade durations, are by construction sparse to (at least) one side with the implication that global bases tend to follow such points. Here we benefit greatly from a more robust method. Once we have approximated the diurnal curve with this smoother, we can fit a basis globally to the smooth. A basis fitted to the robust smooth does not have to deal with possible outliers and, hence, can adequately describe the main features of the diurnal curve in relatively low dimension.

Our method of selecting the "optimal" data span for the robust smoother is largely

dictated by sample size. To avoid repeated application of the smoother at different data spans we opt for a plug-in method adapted from the "rule of thumb" proposed by Fan and Gijbels (1996, p. 111). Let p be the order of the estimator used to approximate a function of order ν. Fan and Gijbels proposed that a polynomial, say $m(x)$, of order $p+3$ be fitted *globally* to the scatter plot. Then an estimate for the bandwidth, h, is given by

$$\check{h}_{ROT} = C_{\upsilon,p}(K)\left(\frac{\check{\sigma}^2 \int w_0(x)\,dx}{\sum_{i=1}^{N}\left\{m^{(p+1)}(X_i)\right\}^2 w_0(X_i)}\right),$$

where $C_{\upsilon,p}$ is a constant dependent on the kernel function K, $\check{\sigma}^2$ is the MSE of the residuals from the polynomial fit and w_0 is some weight function to compensate for edge effects. Since we are interested in preserving the nearest neighbour property we take $X_i = \frac{i}{N}$, for $i = 1, 2, \ldots N$. Then h can be interpreted as the proportion of data to be included on either side of a given point. Since span is two sided, we can approximate the optimal span, d, with $\hat{d} = 2\check{h}_{ROT}$. As a k-nearest neighbour smoother is less susceptible to edge effects we use equal weighting. We will use a tricube kernel, therefore $C_{0,1} = 2.0262$. To make our span selector more robust against outliers, we will replace the MSE with an interquartile range scaled so that it would be equal to the MSE were the residuals normally distributed. The span, $\hat{d}$, can then be used in smoothing the original scatter plot, i.e. smoothing intertrade duration with respect to t_i rather than X_i.

As mentioned earlier, in the second step we fit a basis to the smooth. The reason behind this is to reduce the dimension of the data. The associated parameters can be modelled as representatives of the original functional observation. As the diurnal curves do not show explicit signs of periodicity and in some instances are monotone, we will make use of a polynomial rather than a sinusoidal basis. Our basis of choice is the unshifted Chebyshev polynomials of the first kind and we therefore proceed with a discussion of this topic (see also Weisstein (1999)).

2.3.1 Unshifted Chebyshev polynomials of the first kind

Chebyshev polynomials of the first kind, T_n, are defined through the identity

$$T_n(\cos(\theta)) = \cos(n\theta).$$

where n indicates the order. Alternatively, T_n is defined by

$$T_n(x) = \cos(n \cos^{-1}(x))$$

with $|x| < 1$. This constraint leads to no loss in generality since a one-to-one map can be set up between the time span of a trading day and the interval $[-1, 1]$. We provide the first six Chebyshev polynomials as we will truncate the approximation of the diurnal curve there. They are

$$\begin{aligned}
T_0(x) &= 1 \\
T_1(x) &= x \\
T_2(x) &= 2x^2 - 1 \\
T_3(x) &= 4x^3 - 3x \\
T_4(x) &= 8x^4 - 8x^2 + 1 \\
T_5(x) &= 16x^5 - 20x^3 + 5x.
\end{aligned}$$

Chebyshev polynomials possess two properties which make them attractive for use in our application. They are orthogonal polynomials with respect to the weight function $(1 - x^2)^{-1/2}$, i.e.

$$\int_{-1}^{1} \frac{T_i(x)T_j(x)dx}{\sqrt{1-x^2}} = \begin{cases} \pi & \text{for } i = j = 0 \\ \frac{1}{2}\pi\delta_{ij} & \text{otherwise,} \end{cases}$$

with δ_{ij} the Kronecker delta function, and they satisfy the discrete identity

$$\sum_{k=1}^{m} T_i(x_k)\, T_j(x_k) = \begin{cases} m & \text{for } i = j = 0 \\ \frac{1}{2} m \delta_{ij} & \text{otherwise} \end{cases} \tag{1}$$

where $i, j \leq m$ and x_k, $k = 1, 2, \ldots, m$, are the m zeros of $T_m(x)$ given by

$$x_k = \cos\left[\frac{\pi(2k-1)}{2m}\right].$$

2.3.2 Approach to diurnal effect modelling

One should distinguish between modelling the diurnal effect and modelling the duration. The former we considered in detail in this chapter. The latter we will explored in Section 3.1 in conjunction with the differencing proposed in Section 2.2. The following steps will be applied to all shares analyzed to model the diurnal effect:

1) Durations follow a one-sided distribution by nature. A log transformation is therefore applied to the duration to gain a more symmetric distribution. The log durations for each trading day are then smoothed. As covariate we use the time of day mapped onto the interval $[-1, 1]$ using the map

$$t_i \mapsto \frac{t_i - \frac{1}{2}(t_{\min} + t_{\max})}{\frac{1}{2}(t_{\max} - t_{\min})}$$

where $t_{\min}$ and $t_{\max}$ respectively denote the times of the first and last trades on the day in question. We use the zeros of the Chebyshev polynomial of the highest order to be considered, say order m, as the grid for which we require the smoothed response. The order m, and by implication the number of points in the grid, is selected to be equal to the number of minutes in a trading day.

2) Each smooth is regressed on the first six Chebyshev polynomials. We truncate the number of polynomials to capture the main features of the smooth, while retaining

tractability for modelling purposes. Applying the appropriate transformation to the design matrix, the discrete identity (1) reduces to an indicator function. The benefit from such a standardization lies in its numerical efficiency since the matrix to be inverted in the least squares estimation becomes an identity matrix. The complete (untruncated) design matrix is given by

$$X = \begin{bmatrix} \frac{1}{m} & \frac{T_1(x_1)}{\sqrt{m}} & \dots & \frac{T_m(x_1)}{\sqrt{m}} \\ \frac{1}{m} & \frac{T_1(x_2)}{\sqrt{m}} & \dots & \frac{Tm(x_2)}{\sqrt{m}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{m} & \frac{T_1(x_m)}{\sqrt{m}} & \dots & \frac{T_m(x_m)}{\sqrt{m}} \end{bmatrix}.$$

Hence, we obtain a multivariate time series of regression coefficients that describes the main features of the functional (infinite dimensional) diurnal curves for a sequence of trading days in a tractable dimension.

3) The time series obtained in step 2 is analyzed to determine its dependence structure. An appropriate, possibly vector, $ARMA$ process is selected and fitted to the series. To select the order of the process and the Chebyshev coefficients to retain we will iteratively first select the order of the model using the Schwartz's Bayesian criterion (SBC) and second omit the highest order Chebyshev coefficient that does not Granger cause any other Chebyshev coefficient in the series, i.e. the history of one coefficient does not contain information useful to predicting another (see Hamilton (1994, Chapter 11)).

This approach will be applied to four representative stocks listed on the JSE. Graphical illustrations of the first two steps of the approach applied to two trading days can be found in Section 4.1 where we analyze four data sets. We provide an introduction to the parts of the data required for diurnal modelling in the next section. The remainder of the data to be utilized will be discussed as it becomes relevant.

2.4 The data

Trade by trade data for all stocks listed on the JSE were purchased for the period of November 2004 to February 2005[6]. The period in question contains 82 trading days. We will utilize four shares in all subsequent modelling. We chose, according to the market capitalization on the ALSI40[7], numbers one, ten, twenty and thirty on the list. The reason for this was to assess the suitability of models for shares that have different trade frequencies. Especially, shares lower down the ranking on the ALSI40 would be considered low trade frequency shares on the NYSE. The four shares we will use are shown in Table 2.1 below.

Table 2.1. The four shares on the JSE that we will utilize in our analysis, with their corresponding ranks (by market capitalization) and tickers.

Rank	**Name**	**Ticker**
1	ANGLO AMERICAN PLC	AGL
10	MTN GROUP LIMITED	MTN
20	HARMONY GOLD MINING COMPANY LIMITED	HAR
30	TIGER BRANDS LIMITED	TBS

Trading hours on the JSE are from 9:00 until 17:00 from Monday to Friday. We will exclude the durations associated with all trades that occurred outside normal trading hours. Issues related to prices in the data set will be discussed in the next chapter where we will consider the modelling of log returns.

Results from the application of the approach outlined in the previous section to the four identified shares can be found in Tables A.1 – A.8 in the Appendix with the first four tables containing the span selected by the "plug-in" formula and the Chebyshev coefficients for the 82 trading days and the VAR models contained in the latter four. A discussion of the results will follow in Section 4.1.

[6] BFA MacGregor (2005)

[7] The ALSI40 is the South African all share index consisting of the forty shares on the JSE with the largest market captilisation.

Chapter 3

The Model

This chapter proposes a new statistical model for describing the irregularly spaced returns process for a financial stock represented by a UHF data set. The returns process may be considered a form of marked point process with the arrival times representing "points" and the stock returns representing "marks". We explicitly model both the arrival times of transactions and the returns, thus working in calender time. This is different from assuming the data to be evenly spaced, which is commonly referred to as transaction time. The proposed model is unique for a number of reasons. Firstly, the approach taken to remove non-stationarity in mean is not the generally utilized multiplicative model of Engle and Russell (1998) discussed in Section 2.1. Rather, we will detrend by differencing. Secondly, we model the stock returns directly, rather than relying on the midquote, a latent price process or a signal plus noise type model. Thirdly, we approach the fact that a substantial proportion of returns are zero, i.e. a large proportion of trades occur without an actual change in the stock price, from a censoring point of view and utilize methods for missing values in order to estimate our parameters. Finally, we do not attempt to force a single model on a heterogeneous data set. Rather, we attempt to exploit the heterogeneity in the hope of classifying the nature thereof. One could say that we allow a unique regime for each trading day.

3.1 Model for arrival times

In the previous chapter we showed that the multiplicative model for diurnal adjustment proposed by Engle and Russell (1998) does not attain the desired effect and proposed the use of differencing instead. This does not imply that the use of ACD models is inappropriate for intertrade durations. Rather, it means that more work needs to be done on methods that suitably adjust for the diurnal effect before progressing to an ACD specification. Of course, differencing would not be an appropriate method for diurnal adjustment if we intend using

ACD, because observations adjusted through differencing would not necessarily be strictly positive. Hence, we require an alternative specification which we provide in this section.

To formalize, let t_i denote the arrival time of the i-th transaction, then we propose the use of an $IMA - GARCH$ process to model the arrivals process, i.e.

$$\begin{aligned} \nabla^d t_i &= \boldsymbol{\theta}(L) e_i, \\ e_i &= \sqrt{\psi_i}\upsilon_i, \\ \psi_i &= a_0 + \sum_{j=1}^{n_a} a_j e_{i-j}^2 + \sum_{j=1}^{n_b} b_j \psi_{i-j} \\ \text{where } \nabla^d &= (1-L)^d, \\ \upsilon_i &\sim IID(0,1). \end{aligned} \tag{2}$$

Of course one may choose to include an autoregressive component in the model, however, in our analysis (see Section 4.2) we found this to be superfluous. In general we would expect to difference t_i at least twice, i.e. $d \geq 2$, as differencing once obtains the intertrade duration which would contain the diurnal effect. $GARCH$ effects in the innovations do not in general have to be present. In the presence of conditionally homoskedastic innovations an IMA model would suffice. $GARCH$ effects will be tested for with the Kuhn-Tucker multiplier test (see Section 5.3). In the event of the innovations being homoskedastic the specification reduces to the standard Box-Jenkins formulation, i.e. the e_i are iid.

Contrary to the definition of a duration, the above formulation does not explicitly exclude the possibility of negative values. In fitting a model of this form to market data, this is of little consequence since observed data will not contain negative values. However, in generating realizations from a model to, for example, bootstrap the variance of parameter estimates negative durations should be prohibited. We will briefly discuss three methods for addressing the possibility of negative durations.

The crudest method is to simply add a constant, sufficiently large to ensure positivity, to a realization of durations. This would have little or no adverse effect on parameter estimates as it would be removed by the differencing. It is, however, somewhat artificial

and the possible interpretation of such a constant is unclear. A more serious deficiency of this method is that it would influence the mean of the durations and, hence, the number of transactions in a specific trading day.

Another possibility is not to explicitly prohibit negative durations, but rather retain the information contained in the diurnal effect. By also generating a realization of the diurnal effect and adding that to a realization (possibly containing negative partial sum values) the occurrence of negative durations are unlikely. One such method was proposed in Section 2.3. This is a more satisfactory way of decreasing the possibility of negative durations. It would, however, not wholly exclude the event.

The final, and intuitively most appealing, method is by applying a log transformation. The transformation would be applied to the x_i's rather than the t_i's as it is positive durations, i.e. monotone increasing arrival times, we wish to ensure. This implies taking first differences of the t_i's, applying the log transformation to the first differences and then differencing once more to remove the diurnal effect. The implication of differencing log durations is that we assume the diurnal effect to be locally proportional rather than locally linear (assumption implied by differencing raw durations). We will compare the latter two methods in our subsequent analysis (see Section 4.2).

3.2 Model for the log returns

Let S_{t_i} denote the price at time t_i at which the transaction was concluded. Then the log return is given by

$$r_{t_i} = \log\left(\frac{S_{t_i}}{S_{t_{i-1}}}\right).$$

We will utilize a continuous state space model for the modeling of the log returns process, contrary to the discrete state space models proposed by Rydberg and Shephard (2002), Engle and Russell (2002) and Tay, et al. (2004). The reason for our choice is that the JSE, from which we derive our data, quotes prices in cents, rather that ticks. This is equivalent to a tick size of $\left(\frac{1}{100}\right)$ which is substantially smaller than the tick size of $\left(\frac{1}{16}\right)$ drawn from the

NYSE prior to 2001 and considered in Engle and Russel (2002). As mentioned in Chapter 1, since January 2001 the NYSE also employs a decimal system, influencing the relevance of a model considering a tick size of $\left(\frac{1}{16}\right)$. The coverage previously achieved by considering tick movements of $\left(\frac{1}{16}\right)$ for, say two ticks in each direction as in Engle and Russell (2002), would now require roughly 12 ticks of size $\left(\frac{1}{100}\right)$ in each direction. Considering 25 states becomes both tedious and numerically cumbersome. Tay, et al. (2004) only considered discrete price movements of one tick in either direction.

A substantial proportion of log returns in the four shares we will consider are zero, i.e. more that 50% of transactions occur without a price change from the previous level. This is consistent with the findings of Rydberg and Shephard (2002) and Engle and Russell (2002). This implies an atom in the density of the continuous state space and, as it contributes more than half the observed returns, ignoring it would constitute a miss-specification. In the $UHF-GARCH$ context (see Section 1.2) zero returns are lost by prefiltering the data, which casts additional doubt on the use of such a procedure. We will capture the atom by modelling the process "change" / "no change", i.e. zero return / non-zero return, with a Markov process, similar to Rydberg and Shephard's component that describes action. Our model for returns is given by

$$r_{t_i} = m_{t_i}\tilde{r}_{t_i}, \tag{3}$$

where m_{t_i} denotes the aforementioned Markov process with

$$m_{t_i} = \begin{cases} 0 \text{ if } S_{t_i} = S_{t_{i-1}} \\ 1 \text{ otherwise.} \end{cases}$$

We will call $\tilde{r}_{t_i}$ the dynamic returns process. Specific parameterizations of $\tilde{r}_{t_i}$ will be considered in due course. We can observe both r_{t_i} and m_{t_i} directly in the market. However, this is not the case for $\tilde{r}_{t_i}$. Rydberg and Shephard (2002) decompose the process $\tilde{r}_{t_i}$ into a directional and (discrete) size component. Both components are assumed to be zero if there is no change in the price level. We will make use of an alternative approach, namely that the process $\tilde{r}_{t_i}$ can only be *observed* when $m_{t_i} = 1$, thus avoiding any assumptions on the value

of $\tilde{r}_{t_i}$. This could be considered a form of censoring. We could therefore apply methods addressing missing values to obtain parameter estimates in parametric models considered for $\tilde{r}_{t_i}$. To this end we will make use of the EM algorithm of Dempster, Laird and Rubin (1977), which we will discuss later in this chapter.

We will model $\tilde{r}_{t_i}$ as an $ARMA$ type process. In addition, innovations may be made to conform to a $GARCH$ specification if required. The model for $\tilde{r}_{t_i}$ is given by

$$\begin{aligned} \boldsymbol{\phi}(B)\tilde{r}_{t_i} &= \boldsymbol{\theta}(L)\varepsilon_{t_i}, \\ \varepsilon_{t_i} &= \sqrt{h_{t_i}}\eta_{t_i}, \\ h_{t_i} &= \alpha_0 + \sum_{j=1}^{n_\alpha} \alpha_j \varepsilon_{t_{i-j}}^2 + \sum_{j=1}^{n_\beta} \beta_j h_{t_{i-j}} \\ \text{where } \eta_{t_i} &\sim IID(0,1). \end{aligned} \tag{4}$$

The above formulations of (3) and (4) comes down to a mixture between a discrete and a continuous state model. The implication is that moves in a stock price are modeled as inherently continuous while recognizing that a substantial proportion of transactions do not imply a change in the price.

We could generalize (4) to incorporate the arrival times process by, for example, letting the standard deviation of the innovations take on some functional form of the arrival times. As it stands, (4) only makes use of duration information to indicate the start and end of a trading day and the number and position (arrival time) of transactions in that day. We will consider one extension of (4) that incorporates duration more explicitly into the price process. The idea of Engle (2000) to model returns as a rate per unit time rather than model the observed returns directly can also be applied here to the dynamic returns process. By modelling the returns conditional on the duration we can standardize the return as follows:

$$\acute{r}_{t_i} = \frac{r_{t_i}}{\sqrt{x_i}}. \tag{5}$$

The standardization is based on the assumption that the variance of the ε_{t_i} is proportional to the duration, which is consistent with diffusion models. This does not affect zero returns

so that we can apply both (3) and (4) without additional modification.

We do not explicitly consider diurnal adjustment of the price process in our analysis, consistent with Rydberg and Shephard (2002). However, by introducing durations into the price model, information related to the diurnal effect is implicitly incorporated.

Diurnal adjustment by the multiplicative model would induce some interdependence between trading days. This could be interpreted incorrectly as conditional heteroskedasticity, while in actual fact it is just the stochastic diurnal curve considered in Chapter 2. As (2) does not require multiple days for diurnal adjustment it allows additional flexibility in modelling individual days with a parameterization best suited to its characteristics. We are, for example, able to model innovations with a $GARCH$ parameterization for one day, but exclude it for a day with homoskedastic innovations As with the arrival times model (2), in the event of homoskedastic innovations the ε's reduce to the standard iid innovations Box-Jenkins formulation. In addition to possible interdependence induced by the incorrect diurnal adjustment, the same could be argued for miss-specification of the dynamic returns process by ignoring irregular spacing. We will have the opportunity to study this phenomenon by fitting (4) both with and without employing the standardization (5).

One would be tempted to compare our model for log returns to the models mentioned in Section 1.1. However, the applicability of such a comparison is questionable for the following reasons: Firstly, it is not clear how to compare the models in a simulation study. For instance, would one simulate data from a continuous, a discrete or a mixture models? By implication this choice would favour models on the corresponding state space. Secondly, a comparison on observed market data would require a criteria to benchmark models against, but no clear criteria exists to compare models of such diverse natures. In addition, the discrete price models were formulated to explain a data generation system no longer in use and would therefore be applied out of context.

3.3 Model fitting with Quasi Maximum Likelihood

In our model specification (4) we assumed that the innovations η_{t_i} in the presence of conditional heteroskedasticity have mean zero and variance one. However, we did not assume normality. We will make use of a quasi-maximum likelihood estimator (QMLE) for our parameters. For the general $ARMA - GARCH$ model formulated for the log returns, the quasi-log likelihood function is given by

$$l(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{t}) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\left(\log(h_{t_i}) + \frac{\varepsilon_{t_i}^2}{h_{t_i}}\right).$$

Strong consistency and asymptotic normality of the QMLE were proven by Francq and Zakoian (2004) in their Theorem 3.1 and the asymptotic variance of the estimators can be found in their Theorem 3.2. As we will assume a censoring regime and make use of the EM Algorithm in fitting the $ARMA - GARCH$ models to the price process, we will rather estimate the variance of our parameter estimates by applying the bootstrap. The reasoning behind this is that in addition to the QMLE both the variance estimates and the EM Algorithm rely on asymptotic theory and therefore using asymptotic variance estimates implies the application of asymptotic theory on asymptotic theory twice over. We could calculate the asymptotic variance estimates but we would not have an indication of their ability to actually reflect the true accuracy of parameter estimates.

3.4 Missing values and model identification

As stated earlier in the chapter we will regard a zero return as an unobservable state in the dynamic returns process. This leads us naturally to a discussion of methods for addressing missing values. There are two issues that require attention. The first one is the identification of models and the second is their fitting to data. We consider the former in this section, while a discussion on the latter will be deferred to the next section.

In general the autocorrelation function (ACF) and the spectrum of a time series process

are often used as indicators for model selection. We consider three methods for the computation of each in the presence of missing values. As there is a one-to-one relation between the ACF and the spectrum, for each method we might consider either but not necessarily both.

Parzen (1963) proposed the following estimator for the autocovariance function ($ACOF$) at lag s:

$$\hat{\gamma}_{\tilde{r}}(s) = \frac{\hat{\gamma}_r(s)}{\gamma_m(s)},$$

for all s such that $\gamma_m(s) \neq 0$. This method is not suited to our kind of analysis because in our case γ_m is not known and has to be approximated by $\hat{\gamma}_m$. This replacement of γ_m by $\hat{\gamma}_m$ sometimes causes erratic behavior in $\hat{\gamma}_{\tilde{r}}$ as small values in $\hat{\gamma}_m$ can imply values larger than one for $\hat{\gamma}_{\tilde{r}}$ purely due to sampling effects. It is know that a first order Markov chain is also an $AR(1)$ process, see Harvey (1993). Negative signs in $\hat{\gamma}_m$, due to negative parameter values for the AR process, would cause alternating signs in $\hat{\gamma}_{\tilde{r}}$ that does not necessarily reflect the characteristics of $\tilde{r}$.

The second method is a special case of tapering, see Priestley (1981, Chapter 7). Before calculating the spectrum, the "non-missing" values are centred, i.e. the sample mean of the "non-missing" observations are calculated and subtracted from them. Missing values are then replaced by zero. In this way missing values make no contribution to variance at different frequencies, but the order and spacing of observations are retained.

The final method uses the relationship between the $ACOFs$ of the two observed processes r_{t_i} and m_{t_i} and the partially observed process $\tilde{r}_{t_i}$ to derive an approximation to the $ACOF$ $\gamma_{\tilde{r}}$ and the spectral measure $f_{\tilde{r}}$. The method is described in the following Theorem.

Theorem 1 *Let* $\{X(t), t \in \mathcal{Z}\}$ *be a zero-mean second order stationary series and let* $\{m(t), t \in \mathcal{Z}\}$ *be a second order stationary series with* $m(t) \in \{0; 1\}$ *for all* t*, with* X *and* m *statistically independent. Define*

$$Y(t) = m(t)X(t).$$

Then

$$\gamma_X(s) = \frac{\gamma_Y(s)}{\gamma_{\tilde{m}}(s) + \mu_m^2}$$

and

$$f_X(\lambda) = \sum_{|s|<\infty} \frac{\gamma_Y(s)}{\gamma_{\tilde{m}}(s) + \mu_m^2} e^{2\pi i \lambda s}$$

where $\mu_m = E[m(t)]$ *and* $\tilde{m}(t) = m(t) - \mu_m$.

Proof. *Now,* $\{Y(t),\, t \in \mathcal{Z}\}$ *is also a zero-mean second order stationary process. We wish to find its spectrum. Let* Z_m, $Z_{\tilde{m}}$, *and* $Z_{\tilde{r}}$ *respectively denote the spectral measures of* $m, \tilde{m}$ *and* $\tilde{r}$, *see for example Priestley (1981). Using the Cramėr representation, we have*

$$\begin{aligned}
Y(t) &= \int_0^1 e^{-2\pi i\lambda t} Z_m(d\lambda) \cdot \int_0^1 e^{-2\pi i\lambda t} Z_X(d\lambda') \\
&= \int_0^1 e^{-2\pi i\lambda t} Z_{\tilde{m}}(d\lambda) \cdot \int_0^1 e^{-2\pi i\lambda' t} Z_X(d\lambda') + \mu_m \int_0^1 e^{-2\pi i\lambda' t} Z_X(d\lambda') \\
&= \int_0^1 e^{-2\pi i\lambda t} \left\{ \int_0^1 Z_{\tilde{m}}(d\lambda - \lambda') Z_X(d\lambda') + \mu_m Z_X(d\lambda) \right\}
\end{aligned}$$

where the last equality follows from the fact that the product of two Fourier transforms is the Fourier transform of their convolution. Hence, the spectral measure of $\{Y(t),\, t \in \mathcal{Z}\}$ *is*

$$Z_Y((a,b]) = \int_0^1 Z_{\tilde{m}}((a,b] - \lambda') Z_X(d\lambda') + \mu_m Z_X((a,b])$$

and the spectral distribution function is

$$\begin{aligned}
F_Y((a,b]) &= E|Z_Y((a,b])|^2 \\
&= \int_0^1 \int_0^1 E\left\{ Z_{\tilde{m}}((a,b] - \lambda') \cdot Z_{\tilde{m}}((a,b] - \lambda'')^* \right\} \cdot \\
&\quad E\left\{ Z_X(d\lambda') \cdot Z_X(d\lambda'')^* \right\} + \mu_m^2 E|Z_X((a,b])|^2 \\
&= \int_0^1 F_{\tilde{m}}((a,b] - \lambda') F_X(d\lambda') + \mu_m^2 F_X((a,b])
\end{aligned}$$

where $(\cdot)^*$ *denotes the complex conjugate. If both* $F_{\tilde{m}}$ *and* F_X *are absolutely continuous, then the spectral density of* Y *is*

$$\begin{aligned} f_Y(\lambda) &= \int_0^1 f_{\tilde{m}}(\lambda - \lambda') f_X(\lambda') d\lambda' + \mu_m^2 f_X(\lambda) \\ &= \int_0^1 f_X(\lambda - \lambda') f_{\tilde{m}}(\lambda') d\lambda' + \mu_m^2 f_X(\lambda). \end{aligned}$$

To recover f_X *from this relation we take the Fourier transform on both sides yielding*

$$\begin{aligned} \gamma_Y(s) &= \gamma_X(s)\gamma_{\tilde{m}}(s) + \mu_m^2 \gamma_X(s) \\ &= (\gamma_{\tilde{m}}(s) + \mu_m^2) \cdot \gamma_X(s), \end{aligned}$$

i.e.

$$\gamma_X(s) = \frac{\gamma_Y(s)}{(\gamma_{\tilde{m}}(s) + \mu_m^2)}.$$

It then follows directly that

$$f_X(\lambda) = \sum_{|s|<\infty} \frac{\gamma_Y(s)}{\gamma_{\tilde{m}}(s) + \mu_m^2} e^{2\pi i \lambda s}.$$

■

Thus, our estimate of $\gamma_{\tilde{r}}(s)$ is

$$\hat{\gamma}_{\tilde{r}}(s) = \frac{\hat{\gamma}_r(s)}{\hat{\gamma}_{\tilde{m}}(s) + \hat{\mu}_m^2}. \tag{6}$$

To assess its sampling properties, under the assumption of white noise, we conducted the following simulation: We generate 10000 Gaussian white noise realizations, each of length $N = 500$, and calculate the first order autocorrelation, $\hat{\rho}_1$. Figure 3.1 shows that the chosen sample size is sufficient for the asymptotic $N(0, N^{-1})$ distribution to hold.

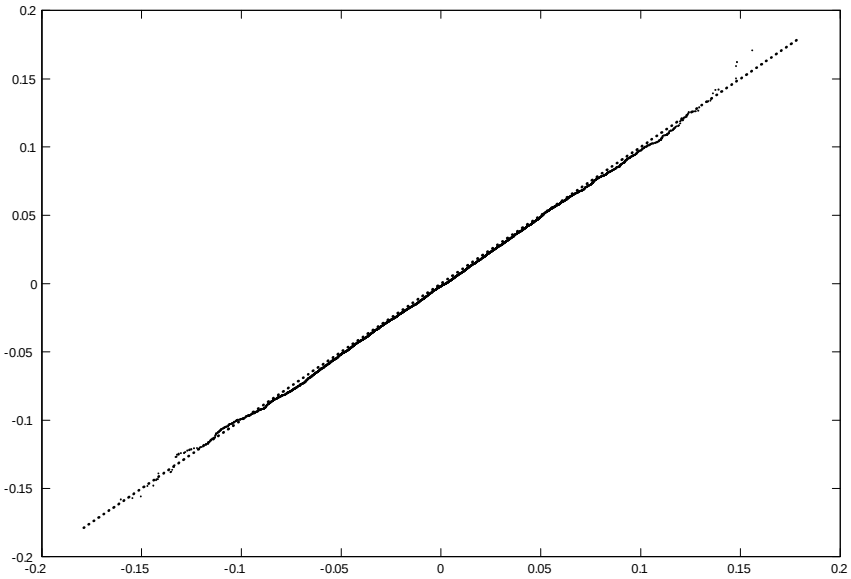


Figure 3.1. Q-Q plot for the first order sample autocorrelation of 10000 white noise realizations with $N = 500$. The dotted line represents a line through the origin with unit slope corresponding to a $N\left(0, N^{-1}\right)$ distribution.

Subsequently we investigate the impact that two scenarios have on the distribution of $\hat{\rho}_1$. The first scenario investigates the impact of a fixed proportion of missing values but the physical location of missing values differs, i.e. we use different censoring regimes. To this end we censor 50% of the observations in the 10000 realizations in two different ways. For the first we censor symmetrically around the middle of each series, i.e. we censor observations 126 to 375, and for the second we partition each series into a hundred contiguous subsets of five each and censor all the odd numbered subsets.

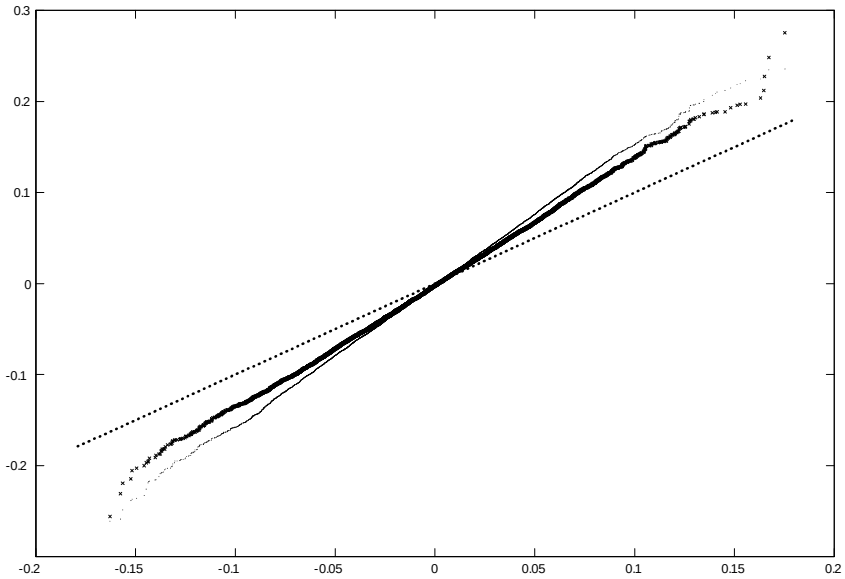


Figure 3.2. Q-Q plot for the first order sample autocorrelation of 10000 white noise realizations with a fixed proportion of missing values censored in different locations with $N = 500$. The heavier dots represent the realizations censored symmetrically around the middle, while the lighter dots represent the censoring of odd numbered subsets. The dotted line represents a line through the origin with unit slope corresponding to a $N\left(0, N^{-1}\right)$ distribution.

This simulation does not enable us to come to a more precise conclusion than that the way in which realizations are censored has implications for how the distribution of $\hat{\rho}_1$ differs from $N\left(0, N^{-1}\right)$. Both cases seem normally distributed, but the variance is not consistent with the asymptotic result in the absence of missing values. There are many possibly ways in which 50% of the observations can be censored and we merely selected two. Each possible way could potentially have a different effect on the distribution of the sample autocorrelations.

The second scenario we investigate is when we retain the same basic location of missing values, but we vary the proportion of missing values. To this end we censor 30% and 50% of the observations symmetrically around the middle of the series. A Q-Q plot of the results is shown in Figure 3.3.

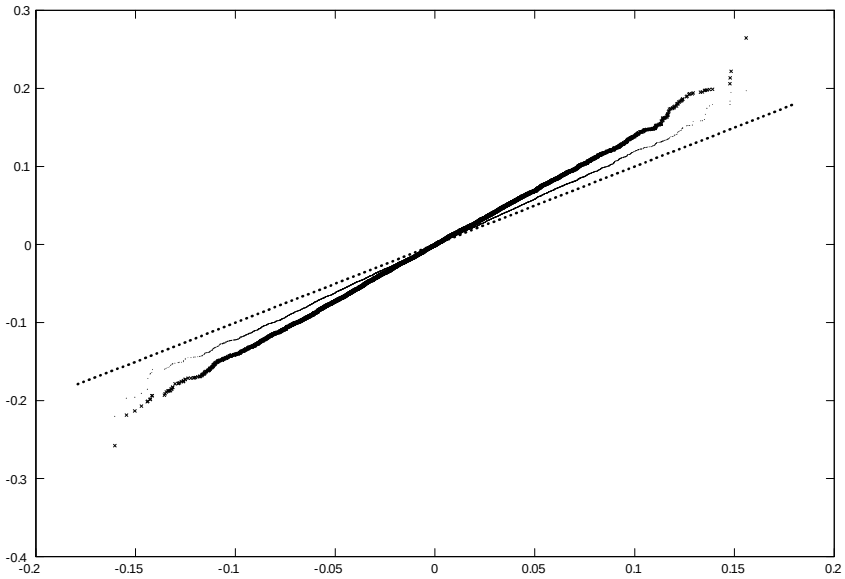


Figure 3.3. Q-Q plot for the first order sample autocorrelation of 10000 white noise realizations censored symmetrically around the middle of the series with $N = 500$. The heavier dots represent the case were 50% of observations were censored, while the lighter dots represents a 30% censoring. The dotted line represents a line through the origin with unit slope corresponding to a $N\left(0, N^{-1}\right)$ distribution.

Intuitively one would expect that the variance of autocorrelations would increase with more missing values. Figure 3.3 serves as confirmation since the higher proportion of missing values rotates the dots more counter clockwise with the implication that the variance of the sample autocorrelations increases for a higher proportion censored. Again, normality seems reasonable.

As a result of our simulation study we can now assert that the presence of missing values also have implications for lack-of-fit tests based on sample autocorrelations. In Chapter 5 we mention four such tests more suited to large samples. However, none of these tests take the presence of missing values into account in the null distribution of the test statistic. To assess goodness-of-fit for $\tilde{r}$ in (3) we can replace the ACF in the Portmanteau statistic, Q, with (6), however, we cannot expect that the distribution of $Q\left(\hat{m}\right)$, for instance, will remain as specified in Section 5.1 since that result is based on the asymptotic $N\left(0, N^{-1}\right)$ distribution of the sample ACF of white noise. Therefore, the distribution of a test statistic

making use of (6) should be estimated through simulation. We will do so by generating white noise series and censoring the observations corresponding to the missing observation in our residuals of interest, before calculating a critical value. In this way we circumvent the possible implications of changing the location or proportion of missing values.

In further analysis we will make use of (6) to estimate the ACF were applicable.

3.5 Missing values and the EM algorithm

The EM algorithm by Dempster, Laird and Rubin (1977) is an approach to data augmentation in incomplete-data problems. It provides a way of doing maximum likelihood estimation for problems that would have been easy had there been a complete data set. We will apply this algorithm here and in subsequent chapters to estimate the parameters for the dynamic returns process $\tilde{r}_{t_i}$ in (3). Before we consider the EM Algorithm it is informative to consider the problem at hand through an example.

Example 1 *Consider parameter estimation for an $ARMA-GARCH$ model. In the absence of missing values parameter estimates can be found with relative ease utilizing the QMLE of Section 3.3 . In contrast, finding the QMLE is not so straightforward in the presence of missing values. If we were to replace the missing values and assume the augmented set of observations to be the full data set of interest, then fitting again becomes straightforward. However, augmenting the data set has implications for our parameter estimates, because if we knew the exact model to augment from we would not need to fit a model in the first place. It is at this point that the EM Algorithm contributes by providing a method for replacing the missing values without adversely affecting parameter estimates.*

We proceed with an overview of the EM Algorithm as it relates to our problem. For a more extensive overview of the Algorithm as well as modifications and alternative applications, see McLachlan and Krishnan (1997).

The algorithm is an iterative procedure for data augmentation that consists of two steps at each iteration. With an initial estimate (guess) of parameters the first step, the expectation

step, augments the data set. In this step each missing value is replaced by its expectation under the current estimates for parameters. This provides a data set to which maximum likelihood estimation can be applied without complication.

The second step, the maximization step, assumes the augmented data set from the first step to be the full set. Maximum likelihood estimation is now applied to calculate new parameter estimates. Dempster, Laird and Rubin (1977) showed that these new estimates are closer to a local maximum of the (full information) likelihood function than those started out with in the previous step to augment the data set. With the improved parameter estimates we can go back to the previous step to replace missing values with their (improved) expected values. By iterating on these two steps and continually improving our estimates, convergence is achieved. The EM Algorithm is shown to converge to a local maximum of the likelihood function, see McLachlan and Krishnan (1997, Chapter 3) for theorems on convergence of the algorithm in missing value problems.

To formalize, let $\phi^{(k)}$ denote the parameter estimates of the current value of $\phi \in \Phi$, the parameter space, after k iterations of the EM algorithm. The next iteration of the algorithm can then be described as follows:

- E-step: Estimate the complete data maximum log likelihood l by finding

$$l^{(k)} = E_{\phi^{(k)}}[l(\phi|\tilde{r})|r].$$

- M-step: Determine $\phi^{(k+1)} \in \mathbf{\Phi}$ as any value of ϕ that maximizes $E_{\phi^{(k)}}[l(\phi|\tilde{r})|r]$, hence

$$E_{\phi^{(k)}}[l(\phi^{(k+1)}|\tilde{r})|r] \geq l^{(k)}$$

  for all $\phi \in \Phi$.

The Algorithm is terminated once the difference between $l^{(k+1)}$ and $l^{(k)}$ is smaller than a predetermined distance. The remaining issue is to select a starting value for ϕ, say $\phi^{(0)}$. This choice is arbitrary and does not in general affect the value to which $\phi^{(k)}$ converges, see

McLachlan and Krishnan (1997). It does, however, affect the number of iterations required until convergence. For our purpose we will select $\phi^{(0)}$ to be the QMLE based on r_{t_i}.

It is worth mentioning that when the EM algorithm is applied to a $GARCH$ specification, care should be taken with the variance of residuals. In the likelihood function used, the variance of residuals is implicitly constrained to unity. However, as the algorithm replaces missing values by their expectation under a given parameter vector, the variance of residuals estimated for the expected values is relatively low, with the implication that the variance of residuals of non-censored observations is correspondingly inflated to satisfy the unity constraint. We circumvent this by penalizing the likelihood function. The penalty we use, and which is subtracted from the log-likelihood, is

$$Ind\left[\sum_{observed}\eta_{t_i}^2 > \chi^2_{\acute{n}-1}\left(1-\frac{\upsilon}{2}\right)\right] M \sum_{observed}\eta_{t_i}^2,$$

where $\acute{n}$ is the number of non-censored residuals, M is some arbitrary "large" constant and υ a predetermined significance level. The penalty can be explained and motivated as follows:

Firstly, we require a penalty that forces our optimization routine into the admissible parameter space without the need for the values with which we initialize the routine to actually comply, i.e. we wish to avoid the requirement of selecting values for the α's and β's that restrict the variance of the observed residuals, hence the need for a penalty with a non-trivial first derivative.

Secondly, the residual variance cannot be expected to be exactly one. We just want it to be "sufficiently" close to unity. Thus, to impose the penalty we need to formalize the concept, "sufficiently" close to unity, in some way. Consider an approximation to the residual variance, say $\hat{\sigma}^2$. For normally distributed residuals we have

$$P\left[\chi^2_{N-1}\left(\frac{\upsilon}{2}\right) < \frac{N\hat{\sigma}^2}{\sigma^2} < \chi^2_{N-1}\left(1-\frac{\upsilon}{2}\right)\right] = 1-\upsilon.$$

By construction, $\sigma^2 = 1$ in a $GARCH$ model. Hence, the preceding relation reduces to

$$P\left[\chi^2_{N-1}\left(\frac{\upsilon}{2}\right) < N\hat{\sigma}^2 < \chi^2_{N-1}\left(1-\frac{\upsilon}{2}\right)\right] = 1-\upsilon$$

An analogous approximation to the previous equality would be

$$P\left[\chi^2_{\acute{n}-1}\left(\frac{\upsilon}{2}\right) < \sum_{observed} \eta^2_{t_i} < \chi^2_{\acute{n}-1}\left(1-\frac{\upsilon}{2}\right)\right] \approx 1-\upsilon.$$

and our penalty follows. Thus, our penalty attempts to restrict the variance of the observed residuals to within a predetermined confidence interval. Since the observed residuals tend to have inflated variance we need not explicitly impose a lower bound.

To illustrate the use of the EM Algorithm we generate 1000 realizations from an $AR(1)$ model with sample size $N = 500$. We selected $\mu = 10$, $\phi = 0.5$ and $\sigma_\varepsilon = 2$. We fit an $AR(1)$ model to each of the realizations utilizing a conditional log likelihood. Then we randomly censor 50 observations in each realization and fit and $AR(1)$ model utilizing, in addition to the conditional log likelihood, the EM Algorithm. The results are illustrated by the following histograms.

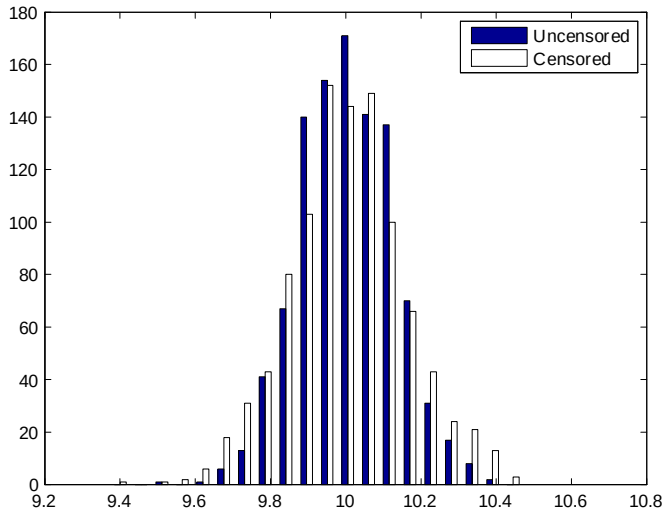


Figure 3.4.1. Histogram for the sample mean, $\hat{\mu}$, from simulation to compare estimates from a conditional log likelihood with results from the EM Algorithm (Censored).

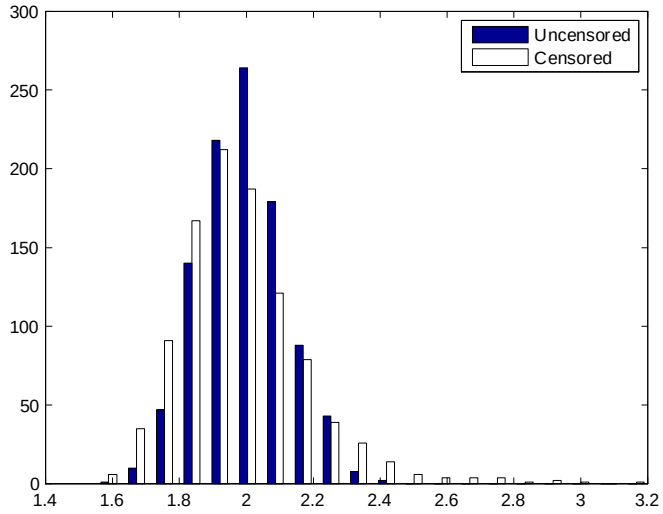


Figure 3.4.2. Histogram for the estimated innovation standard deviation, $\hat{\sigma}_{\varepsilon}$, from simulation to compare estimates from a conditional log likelihood with results from the EM Algorithm (Censored).

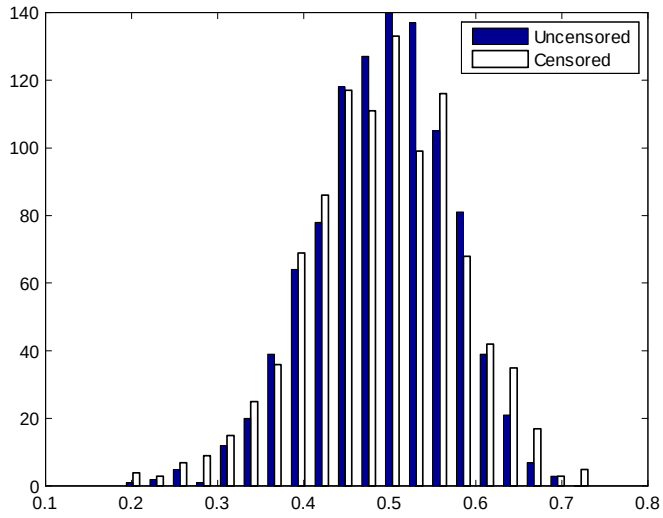


Figure 3.4.3. Histogram for the estimated AR parameter, $\hat{\phi}$, from simulation to compare estimates from a conditional log likelihood with results from the EM Algorithm (Censored).

For all three parameters, estimates utilizing the EM Algorithm show more variability than the estimates utilizing the conditional log likelihood. This is to be expected since the EM Algorithm estimates are based on 10% fewer observations and the correlation structure

is somewhat obscured by the censoring. The EM Algorithm estimates for μ and ϕ seem unbiased, however, the estimates for σ_ε seem to demonstrate some bias in that the median of the estimates fall below the choice of $\sigma_\varepsilon = 2$ from which the data was generated. This is consistent with the notion mentioned earlier in the section that, since missing values are replaced by the expectation under maximum likelihood, their variance is less than that of observed values. In general, the EM Algorithm seems to provide acceptable results.

3.6 The data

The four representative stocks we will use were mentioned in the previous chapter. The following rules were applied to the data prior to modeling, mostly to increase data integrity:

- We exclude all trades captured outside of standard trading hours.
- Simultaneous trades are replaced by a single trade with the volume weighted average price (VWAP) replacing the multiple prices[8]. It is noteworthy that often the prices of simultaneous trades are the same so that this common price is also the VWAP.
- No duration is associated with the first trade of each morning.
- Any return exceeding 50% on either side is deemed to be a data capturing error and hence such trades are omitted. This rule is designed to excluded trades where the decimal point seems to have been misplaced.

3.7 Approach to duration modelling

The following approach will be taken to the modelling of durations. It will be applied to both the raw and log durations to enable us to compare the implications of the log transformation.

[8] Advocates of discrete price models tend to select one of the prices to preserve the discrete nature of ticks while sacrificing on accuracy when all prices are not equal, see for example Hautsch (2004, Chapter 4).

1) *Model identification:* Without exception, duration series show the unmistakable characteristics of an $MA(1)$ process. This could be a reflection of the fact that series are not far removed from white noise to start off with. A white noise series differenced once would result in an $MA(1)$ process with a unit root. Hence, we test the differenced series for a unit root using Tanaka's test described in Section 5.2.

2) *Fitting:* Next we fit an $MA(1)$ model to each trading day and test the residuals for the presence of conditional heteroskedasticity. The latter is achieved with the Kuhn-Tucker test described in Section 5.3. Here we do not apply the FDR method (see Section 5.1.2) as our aim is model selection rather than simultaneous testing. If indicated, the $MA(1)$ model is refitted incorporating a $GARCH(1,1)$ specification for the residuals. We refrain from fitting other $GARCH$ specifications because $GARCH(1,1)$ is usually sufficient (see for example Bollerslev, et al. (1992) or Hansen and Lunde (2005)).

3) *Standard errors:* Standard errors are estimated with the bootstrap.

4) *Goodness-of-fit testing:* Goodness-of-fit is determined by the $Q(\hat{m})$ statistic applying the FDR method, see Section 5.2.

Results from the application of this approach to raw durations can be found in Tables A.9 – A.12 and the log durations in Tables A.13 – A.16 in the Appendix. We will discuss and compare these results in Section 4.2.

3.8 Approach to returns modelling

We outline the approach to be taken with regard to the modelling of log returns. The approach is the same for both non-standardized and standardized returns.

1) *Markov chain fitting:* We start by selecting the order of the Markov chain that adequately describes the process m_{t_i}, defined in (3). This is achieved by applying the logistic regression method described in Section 5.4. We then proceed with estimation of the transition matrix.

2) *Dynamic returns fitting:* An $ARMA$ model is fitted to the process $\tilde{r}_{t_i}$ using the EM algorithm. We select the order of the $ARMA$ process with Schwartz's Bayesian criterion. We consider models to a maximum of eight parameters, i.e. $p + q \leq 6$. Where there are less than twenty "observed" values we restrict the number of parameters to four, and for between twenty and forty values we restrict the number of parameters to six. Residuals are tested for the presence of conditional heteroskedasticity using the Kuhn-Tucker test, see Section 5.4. Any row with a missing value in either the observation vector or the covariate matrix is omitted from the estimation of the Kuhn-Tucker test statistic ξ_{KT}. If required, the $ARMA$ specification is refitted using a $GARCH(1,1)$ specification for the innovations.

3) *Standard errors:* Standard errors are estimated using the bootstrap. We use a smoothed bootstrap, see Davison and Hinkley (1997, Chapter 3). This is more sensible when our resamples are larger than our samples of estimated residuals, since the smoothed bootstrap resamples from a kernel denstity estimator of the residual distribution. In contrast, the standard bootstrap would imply that estimated residuals are repeated often when our resample is substantially larger than the estimated number of residuals. After a realization has been generated, we censor it with a Markov chain realization using the estimated transition matrix.

4) *Goodness-of-fit testing:* Finally, goodness-of-fit is assessed using the $Q(\hat{m})$ statistics and the FDR with the modifications required for the ACF, (6), as indicated earlier in Section 3.4.

Results for the first step can be found in Tables A.17 – A.20 in the Appendix. Results for the application of subsequent steps for non-standardized returns are in Tables A.21 – A.24 in the Appendix and Tables A.25 – A.28 contain similar results for standardized returns. We will discuss these results in Section 4.3.

Chapter 4

Application and classification

In this chapter we will discuss results from the application of our approaches to diurnal effect, duration and returns modelling. Firstly, we will discuss results in broad terms to assess the suitability of the proposed approaches. Subsequent to this initial discussion we will define characteristics of interest and evaluate each of our four shares in an attempt to identify relationships between characteristics. Through our classification we aim to add to the literature on market microstructure as our results could potentially contribute scientific evidence in favour of or against some of these propositions.

4.1 Results from diurnal curve modelling

To reiterate, our aim here was twofold: Firstly, to remove the diurnal effect from the data (see Section 2.2) and secondly, to model the evolution of the diurnal curve in an analytically tractable way (see Section 2.3). Following the guidelines of Cleveland (1979) we restrict our data span to between 20% and 80% of the data. The span selected by the "rule of thumb" introduced in Chapter 2, together with the estimates of the first six Chebyshev coefficients, can be found in Tables A.1 to A.4. Recall that we fit these polynomials to the log durations and the grid is mapped into the interval $[-1, 1]$. If we only consider the coefficients of the first six Chebyshev polynomials, this presents us with 82 vector valued observation (corresponding to the number of trading days) each with six variables (the first six coefficients associated with a specific trading day). We provide graphical illustration of both the LOESS fit and the Chebyshev basis fit (see Section 2.3) for the first and fifth trading days in November 2004 of AGL in the following figure. Since the polynomial fit and the basis fit are virtually indistinguishable in a figure including the original data, for both trading days, one panel contains scatter plots of the data with the LOESS fits and the other contains the LOESS fits and the polynomial basis fits. We specifically chose these to

trading days, because the depict the best and worst polynomial fit to the six trading days considered in Figure 2.2.

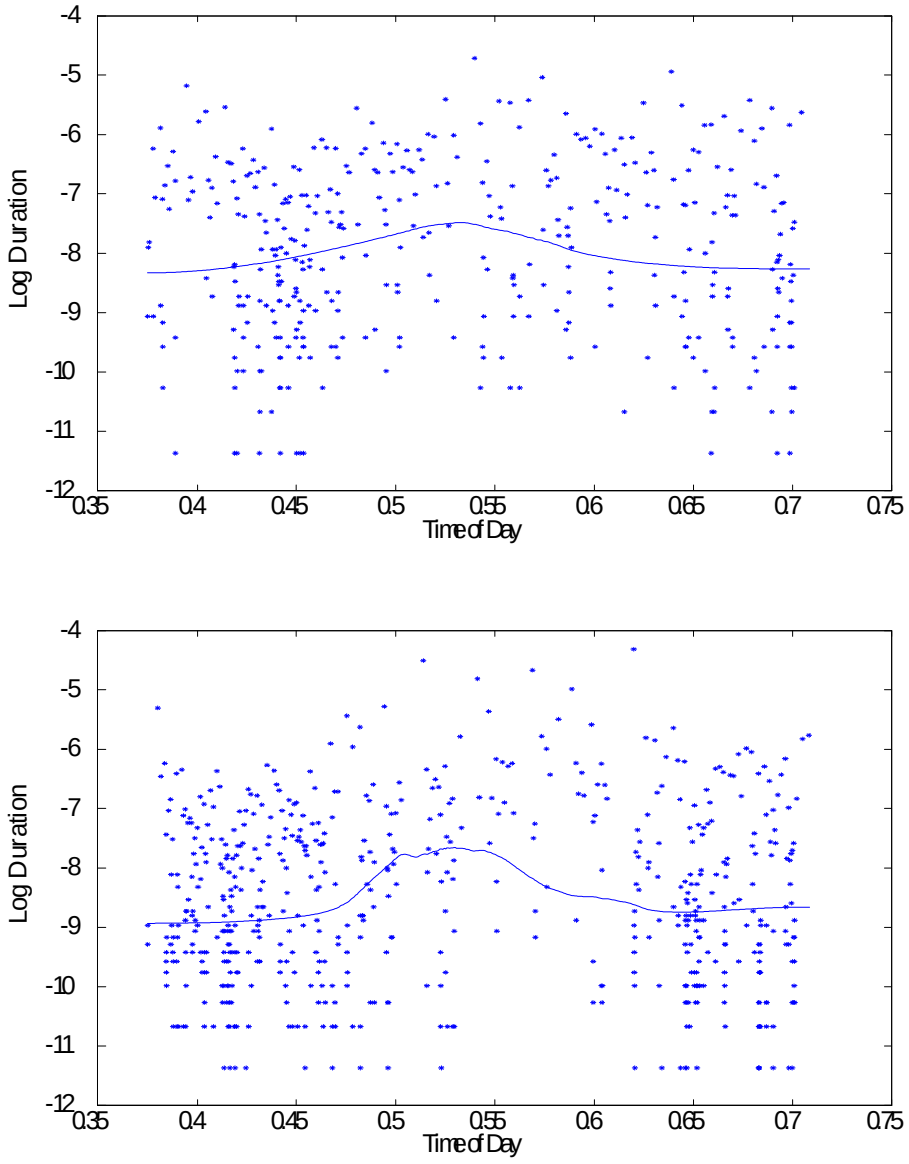


Figure 5.1.1 Scatter plots of log durations for the first (top) and fifth (bottom) trading days in November 2004 for AGL smoothed with the LOESS smoother of Cleveland (1979).

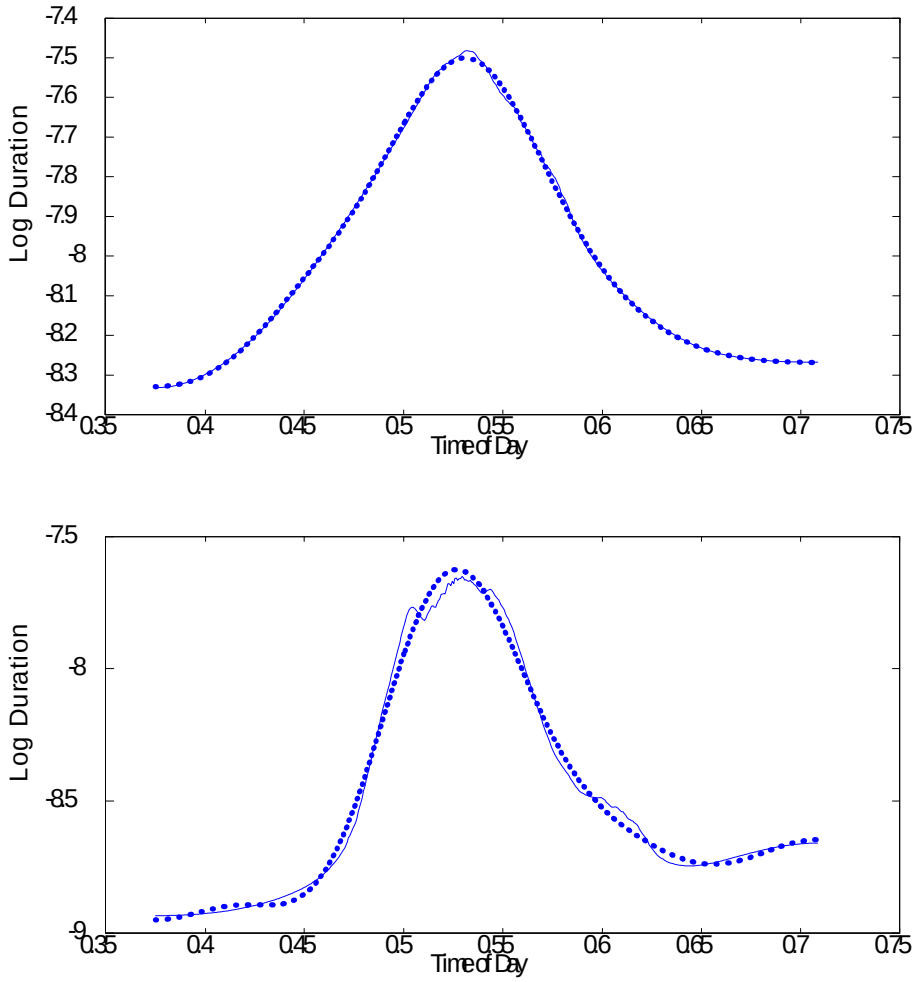


Figure 5.1.2 LOESS smooth (solid line) and polynomial basis fit (dotted line) for the first (top) and fifth (bottom) trading days in November 2004 for AGL.

We consider VAR[9] (vector autoregression) models up to a maximum of five lags. This implies that for each share we have to estimate up to 186 parameters and 21 nuisance parameters from 492 variables (six coefficients for each of the 82 trading days). The ratio of parameters to variables is potentially large, but we do not wish to discount the possibility of a weekly cycle from the outset. For each iteration (approach proposed in Section 2.4.2) for all four shares considered, the SBC selects a $VAR(1)$. The following table summarizes Granger causality at each step.

[9] We apply the standard specification for a VAR model, see for example Hamilton (1994, Chapter 11).

Table 4.1. Iterations for VAR modelling of diurnal curves. We use $A \to B$ to indicate A Granger causes B. The numbers used refers to the coefficient associated with the Chebyshev polynomial of that order.

Iteration	**AGL**	**MTN**	**HAR**	**TBS**
1	$0 \to 0$ $1 \to 2$ $4 \to 5$ $5 \to 5$ Exclude 3	$0 \to 0$ $0 \to 2$ Exclude 5	$0 \to 0$ $1 \to 0$ Exclude 5	$0 \to 0$ $1 \to 0$ $2 \to 1$ $2 \to 4$ $3 \to 2$ Exclude 5
2	$0 \to 0$ $1 \to 2$ $4 \to 5$ $5 \to 5$ Terminate	$0 \to 0$ $0 \to 2$ Exclude 4	$0 \to 0$ $1 \to 0$ Exclude 4	$0 \to 0$ $1 \to 0$ $2 \to 1$ $2 \to 4$ $3 \to 2$ Terminate
3		$0 \to 0$ $0 \to 2$ Exclude 3	$0 \to 0$ $1 \to 0$ Exclude 3	
4		$0 \to 0$ $0 \to 2$ Exclude 1	$0 \to 0$ $1 \to 0$ Exclude 2	
5		$0 \to 0$ $0 \to 2$ Terminate	$0 \to 0$ $1 \to 0$ Terminate	

Parameter estimates for the four VAR models settled on can be found in Tables A.5 to A.8. All four models contain significant parameters for the coefficients of the first Chebyshev polynomial T_0. This would suggest that the average duration for each trading day is (partially) predictable. None of the other coefficients appear in all models, suggesting that they are less predictable, i.e. random. This is especially true for MTN and HAR. The variation in the model specifications could perhaps in some sense be attributed to the fact that our four series only comprise of 82 observations each.

4.2 Results from duration modelling

Our goal was to model the times at which transactions occur. The use of differencing as a method of diurnal adjustment rendered the ACD model, Engle and Russell (1998), unsuitable as it does not allow for negative values. To decrease the likelihood of negative durations in our bootstrap simulations we proposed three approaches, see Section 3.2. We proposed to consider two of these in our application. The first was modelling of the raw differenced duration and adding the diurnal effect to partial sums. The second was modelling differenced log durations. The latter is intuitively more appealing as it guarantees positive durations when taking partial sums of a generated realization. Results can be found in Tables A.9 to A.16 in the Appendix. In each case the results for the raw durations (odd numbered tables) are followed by the results for the log durations of the same share (even numbered tables). For numerical convenience raw durations are multiplied by 1000 before fitting and tables should be interpreted as such.

Let us first consider the outcome of the moving average unit root testing. We wish to ensure that we are not over differencing the series. Table 4.2 contains the proportions of null hypotheses rejected using the FDR method. The null hypothesis of Tanaka's test is that a unit root is present.

Table 4.2. Proportions of null hypotheses rejected in MA unit root testing.

Ticker	**Raw**	**Log**
AGL	0.488	0.427
MTN	0.512	0.573
HAR	0.610	0.646
TBS	0.256	0.329

As the proportions of null hypotheses rejected exceed the 5% level in all cases we conclude that the taking of first order differences of either raw or log durations is not inappropriate, i.e. we are not differencing a white noise series. This being the case, we can now proceed with the remainder of our approach.

At a 5% significance level, the null hypothesis of a zero mean for the raw durations is only rejected once each for MTN and HAR[10]. Hence, we refit the $MA(1)$ processes with zero mean.

Table 4.3. Proportions of null hypotheses rejected in white noise testing of estimated duration residuals.

Ticker	**Raw**	**Log**
AGL	0.037	0.037
MTN	0.037	0.073
HAR	0.012	0
TBS	0	0

Table 4.3 refers to the proportions of rejected null hypotheses from white noise tests conducted on the estimated residuals from the two approaches mentioned, see Section 3.8. It is immediately apparent that the modelling of the differenced raw durations was more successful than that of the differenced log durations in the sense that we conclude simultaneously that residuals estimated from the modelling of differenced raw durations represent a white noise process. This is, however, not the case for the residuals of MTN estimated from modelling of the differenced logarithm of durations. One possible explanation for the comparative success of the differenced raw durations is related to the underlying assumptions with regard to the nature of the trend.

The success of differencing raw durations as a method for diurnal adjustment rests on the assumption that (at least) locally the trend is approximately linear. The corresponding assumption for the differenced durations is that the trend is (at least) locally some fixed proportion. In revisiting Figure 3.2 it is apparent that the slope of the curve switched signs at least once a day, with the implication that the expected proportion between consecutive durations changes from larger than one to smaller than one or visa versa. This violates the assumption required for the success of differenced log duration.

[10]The hypotheses are rejected more often for the differenced durations, but not sufficiently often to contradict the overall conclusion.

To test the zero mean hypothesis we construct a 95% confidence interval around the estimated mean using the bootstrap standard error and the normal approximation. We can, of course, not apply the same approximation to the $GARCH$ parameters as zero lies on the boundary of the parameter space and hence the asymptotic normal approximation does not apply.

In any event, given the results of the white noise tests we are confident that our approach adequately describes intertrade durations.

4.3 Results from returns modelling

Our approach to the modelling of log returns was proposed in Chapter 3. Recall our model

$$r_{t_i} = m_{t_i}\tilde{r}_{t_i}$$

where m_{t_i} is a two state Markov process applying a censoring regime to an $ARMA-GARCH$ type process $\tilde{r}_{t_i}$. Let us first consider results for the modelling of m_{t_i}. We assessed the suitability of a first order Markov chain as a model for m_{t_i} by testing the hypotheses that a first order chain adequately describes a third (respectively tenth) order chain. Table 4.4 contains the proportions of null hypotheses rejected.

Table 4.4. Proportions of null hypotheses rejected in testing whether either a third or tenth order Markov chain could be adequately described by a first order Markov chain.

Ticker	**3 vs. 1**	**10 vs. 1**
AGL	0.073	0.073
MTN	0.098	0.085
HAR	0.049	0.049
TBS	0.061	0.024

When interpreting these proportions it is important to note that these were not calculated using the FDR method[11], rather these represent rejections at the standard 5% level. Had this been calculated using the FDR method, in all probability the proportions would have been less than the required 5% level. To see this, consider MTN where 8 null hypotheses that a third order chain could adequately be described by a first order chain were rejected at the 5% level. This implies that 8 p-values were less than the 5% level. To be rejected under

[11] Given the available CPU time it was not practically feasible to estimate the test statistic's distribution under the null hypothesis accurately enough to apply the FDR method.

the FDR method these p-values, when ordered, would in addition have to smaller than $\frac{8}{82}$ of 5% down to $\frac{1}{82}$ of 5%, which is substantially lower than the levels currently guaranteed.

The estimated transition matrices and corresponding parameter variances can be found in Tables A.17 - A.20. The fact that more than half of trades do not represent a change in price is immediately apparent.

Next, we proceed with our approach to modelling the dynamic returns process, $\tilde{r}_{t_i}$. Our parameter estimates for the non-standardized returns can be found in Tables A.21 - A.24 and for standardized returns in Tables A.25 - A.28. For numerical convenience non-standardized returns are multiplied by 1000 before fitting and tables should be interpreted as such. Table 4.5 contain the FDR values for white noise testing of the shares.

Table 4.5. Proportions of null hypotheses rejected in white noise testing of estimated residuals from non-standardized (non-std) and standardized (std) dynamic returns modelling.

Ticker	**non-std**	**std**
AGL	0.098	0.073
MTN	0	0.037
HAR	0.012	0
TBS	0.024	0.024

Our approach here was not quite as successful as was the case with duration modelling. While our approach seems to do an adequate job with the less traded stocks, there is room for improvement with AGL. We mention a few possible reasons:

- The non-standardized approach to duration modelling is naive in the sense that it does not take the influence of durations on price into account. The standardized approach was able to improve, but still not to below the 5% level. We will mention a few additional extensions to the durations model in the Epilogue.

- Our white noise test (see Section 5.2) is based on the dynamic returns process, $\tilde{r}_{t_i}$. In reality a test based on r_{t_i} would be more appropriate since the standard definition we applied throughout does not reflect the efficacy of the Markov chain, m_{t_i}, in capturing

the atom in the returns distribution. However, this would entail a more general definition of residuals.

- It is generally accepted that the innovations process for log returns is not normally distributed, see for example Audrino (2005), Gourieroux and Jasiak (2001, Chapter 6) or Nelson (1990). The normality assumption implicit in the use of quasi-maximum likelihood could adversely effect parameter estimation as the success of the estimation depends on an asymptotic result. Samples might not be sufficiently large for asymptotics to apply. It is, however, counter intuitive that the approaches are less successful at modelling the more frequently traded stocks which, by implication, would have larger samples.

- The sheer volume of zero returns in relation to non-zero returns could influence the efficacy of the EM algorithm. How effectively the full information maximum likelihood is approximated by the incomplete data maximum likelihood in an $ARMA-GARCH$ setting with such a high proportion of missing values is, to our knowledge, unknown. This is underlined by the fact that standard errors are so large that a substantial proportion of the $ARMA$ parameters could be considered insignificant[12].

- The number of parameters we allow for in the approach is somewhat arbitrary. It is quite possible that allowing no more than six $ARMA$ parameters for some of the trading days of AGL and MTN, which contain in excess of a thousand trades, is too restrictive.

- Some trading days show signs of $IGARCH$ (Integrated $GARCH$) effects, i.e. $\alpha_1 + \beta_1 \approx 1$. The unconditional variance of innovations in $IGARCH$ are infinite. The implication is that the unconditional variance of innovations are large in these instances, also influencing parameter estimation accuracy. Examples include 18 and 22 February

[12] Another possible reason for the large number of parameter estimates that are not significantly different from zero is that we do not iterate through all possible scenarios. Rather, when fitting for example an AR(6) model, we only fit the whole model and not all possible permutations with a significant lag six variable. Hence, it is quite possible that some of the parameters at lags smaller that the highest lag selected could be insignificant. The benefit from this simplification is that we only have to fit a third of the models, reducing runtime without adversely affecting goodness of fit assessment.

2005 for HAR (non-standardized returns) and 18 November and 15 December 2004 for AGL, amongst others. Interestingly, three of the cases mentioned convinced the goodness-of-fit test, i.e. the three null hypotheses of white noise were not rejected, 15 December 2004 for AGL being the exception.

It is inconclusive whether the modelling of standardized residuals is superior to the modelling of non-standardized residuals. The former performed (marginally) better in describing AGL, but (marginally) worse in describing MTN. What is of interest is a comparison between the parameter specifications selected by the SBC and the Kuhn-Tucker test. Firstly, we consider the proportions of trading days that required the addition of a $GARCH$ specification.

Table 4.6. Proportions of trading days that required a $GARCH$ specification for non-standardized (non-std) and standardized (std) dynamic returns modelling.

Ticker	**non-std**	**std**
AGL	0.823	0.427
MTN	0.793	0.415
HAR	0.537	0.207
TBS	0.293	0.207

Table 4.6 reflects the proportions of trading days that required a $GARCH$ specification. It is striking that, except for TBS, roughly double the number of trading days for the non-standardized returns, viz á viz the standardized returns, required the addition of a $GARCH$ component. When we consider that the performance in terms of goodness-of-fit is so similar, it seems that $GARCH$ is to a large extent able to compensate for the information contained in durations.

Table 4.7.1. Proportions of days more parsimoniously described by each returns model, i.e. modelling required fewer parameters.

Ticker	**non-std**	**tie**	**std**
AGL	0.146	0.207	0.646
MTN	0.098	0.354	0.549
HAR	0.159	0.171	0.671
TBS	0.159	0.378	0.463

Table 4.7.2. Ratio of number of parameters required for modelling standardized vs. non-standardized returns.

Ticker	
AGL	0.752
MTN	0.728
HAR	0.714
TBS	0.79

While we were not able to distinguish clearly between modelling of the two types of return on the grounds of goodness-of-fit, we are definitely able to do so on the ground of parsimony. In three of the four cases standardized returns required fewer parameters in more than half the trading days while in less than 16% of trading days did non-standardized returns require fewer parameters, see Table 4.7.1. Table 4.7.2 tells us that standardized residuals required between 20% and 30% fewer parameters in total. Our approach to fitting standardized residuals was thus able to achieve similar results in white noise testing, i.e. comparable proportions of null hypotheses rejected, by using substantially fewer parameters. In the light of Occam's razor[13] one would opt for the most parsimonious model, i.e. the modelling of standardized returns.

As an aside, one could argue that the inclusion of durations changes the nature of the returns process since it influences both the conditional variance and the autocorrelation structure of the latter process.

[13]William of Occam (circa 1284-1343 A.D.), English philosopher

4.4 Characteristics of interest

We proceed with definitions of characteristics to aid us in recognizing patterns in the trading days of our four shares. Our list is by no means exhaustive. We will consider two possible methods to aid us in the identification and interpretation of interactions between characteristics. Firstly, we can subdivide each characteristic into categories. To simplify the analysis we restrict the number of categories in each characteristic to three. In essence, this reduces to identifying whether a trading day classifies as "high", "medium" or "low" in a given category. Where the interpretation of "high", "medium" and "low" associated with a specified characteristic is not self explanatory, we will clarify by definition. Secondly, we will consider the covariance matrices associated with the 82 observed combinations of the characteristics.

We will consider the following seven characteristics:

1) A representative of the *mean duration* associated with a specific trading day: To this end we will utilize the intercept of the Chebyshev polynomials as represented by the coefficient of T_0. This should give us an indication of how frequently a specific share was traded on a specific trading day in comparison to trade frequency realized on other days. High mean duration would imply low trade frequency and vice versa.

2) The *concavity* of the *diurnal curve*: We will use the coefficient of T_2 as indicator. With the classes "high", "medium" and "low" we will associate convex, flat and concave respectively. By implication the subcategories represent the position of the global optimum in relation to the mean duration.

3) The (unconditional) *variance* of differenced *duration*: We are interested in the dispersion of our point process. In a sense we are trying to classify the tendency of trades on a specific trading day to cluster, i.e. trades go through in batches separated by periods of relative quiet.

4) *Memory* of the *returns* process: An $MA(q)$ process has memory of length q, while in contrast an AR process has infinite memory. We will classify trading days with

a predominantly AR structure as high memory, mixed $ARMA$ processes as medium memory and predominantly MA processes as low memory. As measurement unit for process memory we will use aggregated squared autocorrelation up to ten lags.

5) *Direction* of the *price* process: The mean of the returns process signifies whether on average prices are increasing, constant or decreasing. We will define these to be high, medium and low respectively.

6) The (unconditional) *variance* of dynamic *returns*: Classification of volatility, hence risk, associated with the dynamic returns process.

7) Proportion of *zero returns*: The volatility of the realized returns process cannot solely be characterized by that of the dynamic returns process. Hence, we include the proportion of zero returns to capture some of the influence of the Markov chain m_{t_i} on risk. Intuitively a high proportion zero returns would imply lower volatility and vice versa.

We mentioned a few other characteristics that could have been considered, together with motivation for not doing so.

- One could get an indication of the signal to noise ratio of durations by considering the data spans selected. However, in using the "rule of thumb" for span selection, the variance of the data plays a central role. This information is already captured by our third characteristic, the (unconditional) variance of differenced duration.

- Another possibility is the number of transactions that took place on a specific trading day. However, there is a direct relationship between the number of trades on a specific day and the expected duration between trades and the latter is our first characteristic.

4.4.1 Exploring the characteristics via subcategories

The results from the classification of our four data sets can be found in Tables A.29 to A.32. In essence we ranked the trading days according to each characteristic and classified the top

third as "high" the middle third as "medium" and the bottom third as "low". Since we have characteristics 4, 5 and 6 for both non-standardized and standardized returns, we will denote the former by a and the latter by b. We do not consider intersections between a and b measurements of different characteristics.

Initially, we consider pairs of characteristics both for the sake of recognizing patterns and to ensure that some of our characteristics do not implicitly convey similar information. The first pair to consider is mean duration (1) and variance of differenced durations (3). The following table highlights three possible combinations and their counts.

Table 4.8. Cell counts of coinciding categories for characteristics (1) and (3).

	AGL	**MTN**	**HAR**	**TBS**
H	15	13	14	5
M	13	12	10	7
L	19	18	14	10
Total	47	43	38	22

For the first two shares more than half of the 82 trading days are explained by coinciding categories of the two characteristics. This would suggest that days with more transactions have smaller variance in duration. The implication would be that more clustering occurs on days with fewer trades, i.e. transactions are more likely to go through in batches separated by periods of quiet. To determine the amount of clustering in a point process it is customary to consider the index of dispersion, see for example Cox and Isham (1980, Chapter 1). For durations the index is the ratio between the variance and the squared mean. However, because we model the change in duration, rather than the duration itself, this is not applicable. The variance of our $MA(1)$ process can be used as a measure of clustering for the following reason: We have seen in Section 4.2 that the (unconditional) expected change in duration is zero, i.e. we expect constant duration. The variance therefore gives an indication of the spread around the mean. A higher (lower) variance would then indicate larger (smaller)

deviations from constant duration. Larger deviations from a constant duration would imply more short and long durations in comparison, in other words, clustering. As mentioned in Section 1.3 the Admati and Pfleiderer (1988) model suggests the slow trading means informed trading. The Easley and O'Hara (1992) model predicts that informed trades will cluster. Our findings for AGL and MTN we supports these two models. No other pair has two shares with three out of the nine possible combinations accounting for more than half the trading days.

It turns out that combinations of more than two characteristics at a time does not return any obvious patterns. One possible explanation for this could be the vastness of what we are implicitly trying to accomplish. One could argue that we are trying to estimate the cell probabilities of a multinomial distribution. Between our four data sets we have 328 trading days at our disposal. The following table illustrates the number of cell probabilities and corresponding effective number of observations available for the estimation.

Table 4.9. Number of cells and available observation for estimation of cell probabilities. p-uniform represents the probability one would be trying to estimate under the assumption of uniform cell probabilities.

Combination of	**2**	**3**	**4**	**5**	**6**	**7**
Cells	189	945	2835	5103	5103	2187
Effective Observations	6888	11480	11480	6888	2296	328
p**-uniform**	1/9	1/27	1/81	1/243	1/729	1/2187

To estimate one in nine for pairs of characteristics we have about 36 observations available which, while not ideal, is not unreasonable. However, to estimate one in 27 with about 12 observations entails a flight of fancy and for more characteristics considered simultaneously the scenario is worse still. The implication is that, unless a specific combination of categories for a certain set of characteristics is blatantly obvious, i.e. convey the same information, it would require vast amounts of data to estimate cell probabilities to satisfaction.

4.4.2 Exploring the characteristics via correlation matrices

An alternative to the categorizing of characteristics is the correlation matrix. We can estimate the matrix for each of the four shares considered using the 82 measurements for each of the seven characteristics, i.e. we use 574 values to estimate 21 parameters. Before we attempt to interpret the off-diagonal elements of the correlation, we use Bartlett's test for sphericity (see for example Green (1978, pp. 361–362)) to test the hypothesis that the correlation matrix is in fact the identity matrix. Asymptotically, the test statistic for a 7×7 correlation matrix would have a χ^2 distribution with 21 degrees of freedom. Results from the four tests are summarized in the following table.

Table 4.10. Results from Bartlett's test for sphericity in the correlation of the 7 characteristics. We use (a) to indicate non-standardized returns and (b) for standardized returns.

	AGL	**MTN**	**HAR**	**TBS**
Test statistic (a)	160.511	60.251	59.464	25.705
p**-value** (a)	<0.0001	<0.0001	<0.0001	0.2179
Test statistic (b)	116.753	71.844	44.1337	40.4472
p**-value** (b)	<0.0001	<0.0001	0.0022	0.0066

The null hypotheses is rejected for all shares but TBS (using measurements from non-standardized residuals). We conclude that for three of the four shares there is significant interaction between characteristics. The correlation matrices can be found in Tables A.33 to A.36 in the Appendix. We proceed with a discussion of the largest (in absolute terms) off-diagonal correlations. As the largest absolute off-diagonal correlations for TBS are about 0.3 and the null hypothesis was rejected on one of the two occasions, we will attempt to interpret all correlations in absolute value terms larger than 0.3. Before we proceed with the interpretation, it is instructive to consider the correlations between corresponding characteristics for non-standardized and standardized returns. One would expect, seeing they describe the same characteristic in different models, that the correlation would be significant. We apply Bartlett's test for sphericity to the correlation matrix consisting of characteristics 4 through 6, both a and b, to establish to what extent our assertion is true.

Asymptotically, the test statistic for a 6×6 correlation matrix would have a χ^2 distribution with 15 degrees of freedom.

Table 4.11. Results from Bartlett's test for sphericity in characteristics 4 through 7 measure for both non-standardized and standardized returns.

	AGL	**MTN**	**HAR**	**TBS**
Test statistic	133.033	62.474	23.097	31.81
p**-value**	<0.0001	<0.0001	0.0821	0.0068

As expected the null hypothesis is rejected for AGL, MTN and TBS. However, this is not the case for HAR. In addition, when we use our rule of thumb of considering correlations exceeding 0.3 (in absolute value terms) significant, the correlations we expect to see are not so obvious. The two measurements of characteristic 4, memory of the returns process, is significant for AGL, MTN and HAR, but not for TBS. For MTN the measurements of characteristic 5, price direction, is significant, but insignificant for the other three. The correlation between the two measurements of characteristic 6, (unconditional) variance of dynamic returns, is not significant for any of the four shares. In addition, curious relationships, like the one between $5b$ and $6a$, contribute to the rejection of the null hypotheses. The fact that measurements of the same characteristic through different models are not so obviously correlated supports the assertion that the inclusion of durations changes the nature of the returns process (made in relation to the autocorrelation and conditional variance structures of data series, Section 4.3).

The positive relationship between characteristics (1) and (3), identified in the previous section, is also identified by this method for AGL and MTN. For an econometric interpretation we refer the reader to the previous section. However, it does not feature significantly in the two less frequently traded stocks, HAR and TBS. This would tend to suggest that the relationship is more significant for, frequently traded stocks, at least in a South African context.

AGL is the only stock for which the relationship between characteristics (4) and (6) is significant, and then only for non-standardized returns. This suggests that the more volatile

the stock price is, the longer an impulse / shock will affect the price in transaction time. As the relationship between mean duration and memory is significant for none of the stocks, we could safely assume that the relationship between (4) and (6) also holds in calender time.

The relationship between (3) and (5a) in AGL to some extent contrasts the relationship between (1) and (3) found for AGL and MTN as the former suggests that trades occur at less regular intervals when the price is increasing. This contradicts the proposition that prices on the up are associated with a higher trade frequency, which is in turn associated with lower variance in duration. Since for AGL the correlation between (1) and (3) is more than double that between (3) and (5a) we can assume that the relationship between (1) and (3) is dominant. This conclusion is reinforced by the fact that the correlation between (3) and (5a) is insignificant for MTN.

The Easley and O'Hara (1992) model predicts clustering of informed trades. The relationship between 3 and 4b, significant for AGL, supports this notion. If an experienced observer gradually learns the information of informed traders through the order flow, the memory of the price process should be longer (measured in number of transactions) in the presence of informed trading. The latter being a consequence of clustering in the arrivals of transactions under the Easley and O'Hara (1992) model.

The relationship between 4b and 5b could be reflective of the difference between calendar and transaction time. The relationship implies that in transaction time the price rate process has a longer memory when the price is increasing. As we have seen that transactions occur more frequently when the price is increasing, the relationship between 4b and 5b would counteract the decay in memory due to more transactions in the same period of time, i.e. time interval. An implication could be that length of memory is associated with calendar time, rather than transaction time.

The final relationship (arguably) significant for AGL is the one between (1) and (5a). The positive relationship between the two characteristics suggests that stocks are more frequently traded when the price is increasing. This is consistent with the Diamond and Verrecchia (1987) model introduced in Section 1.3. They argued that informed traders will act on good news, but can only act on bad news if they already own the stock or are allowed to sell short,

otherwise they will merely refrain from transacting in the stock. Therefore, an increasing price should be associated with higher trade frequency.

The positive relationship between $(5b)$ and $(6b)$ features in MTN. A possible implication of this relationship is that an increasing price rate is more volatile than a decreasing price rate. This is consistent with the proposition of an asymmetry in the relationship between stock returns and volatility, see Hamilton (1994, Chapter 21) or Nelson (1990). This does not necessarily imply that an increasing price is more volatile than a decreasing price as the effect of duration (relationship between (1) and (3)) and the proportion of zero returns also affect price volatility.

For both HAR and TBS the relationship between $(4a)$ and (7), proportion of zero returns, is (arguably) significant. In Theorem 1, Chapter 3, we derived an approximation to the autocorrelation function of the dynamic returns process. In using this approximation for measurement of the memory of the dynamic returns it is quite conceivable that the autoregressive structure induced by the Markov component of returns is removed to a lesser extent in smaller samples, in which case these two shares would be more susceptible to "residual" autocorrelation induced by the Markov process not removed by our ACF estimator, see Section 3.5. However, since the two correlations in question barely reach the 0.3 mark in absolute terms, this is of a lesser concern. On the contrary, the fact that the other shares do not exhibit this feature is encouraging. In addition, it is to be expected that a higher proportion of missing values would induce more autocorrelation between those actually observed since our estimate of autocorrelation should improve with sample size.

The remaining relationship of possible significance for HAR is that between (1) and (7), which is negative. This suggests the possibility that a higher proportion of zero returns occur on trading days when the stock is traded more often. Both MTN and TBS show the same sign for this relationship, although the correlations are of little significance. We have seen that higher trade frequency is associated with an increase in the price, which in turn is associated with lower volatility of dynamic returns. As the other component of variance in returns, a higher proportion of zero returns for a higher trade frequency would support the proposition of less risk.

It is worth mentioning that none of the microstructure models discussed in Section 1.3 finds support in all four shares considered here. This provides additional support for our argument against the homogeneity of UHF data sets. It would be interesting to repeat the same analysis for more stocks over a longer period of time to assess whether some microstructure hypotheses can be associated with specific sectors or trade frequency categories.

Chapter 5

Technical detail on hypothesis tests

To facilitate the use of a variety of tests applied throughout this thesis we provide a discussion of tests that are not frequently applied and raise issues with regards to the use of standard tests in our scenario. Where issues are raised, alternatives are mentioned or proposed as remedy. Specifically, we consider Lack-of-Fit, Moving Average unit root and $GARCH$ tests and Markov chain order selection.

5.1 Large sample size and goodness-of-fit tests

Due to the size of UHF data sets, where tens of thousands of observations are not uncommon, an approach to white noise testing that is sensitive to large sample size will reject all models with few exceptions. This issue is also mentioned with reference to the Ljung-Box test in Engle (2000): "The $LB(15)$ test for autocorrelation ... is 32.8 which exceeds the 5% point of 25 but is quite reasonable considering the large sample size", but he does not consider the issue in any detail. We explore this characteristic of the Lung-Box test in some detail and consider four alternative methods.

For an autoregressive-moving average model

$$\begin{aligned} \boldsymbol{\phi}(L)x_t &= \boldsymbol{\theta}(L)\varepsilon_t \\ \text{where } \boldsymbol{\phi}(L) &= 1 - \phi_1 L - \ldots - \phi_p L^p, \\ \boldsymbol{\theta}(L) &= 1 - \theta_1 L - \ldots - \theta_q L^q \text{ and} \\ L^k x_t &= x_{t-k}, \end{aligned}$$

where $\{x_t, t \in \mathcal{Z}\}$ and $\{\varepsilon_t, t \in \mathcal{Z}\}$ respectively denote the observation and innovation series, Ljung and Box (1978) proposed the test statistic

$$Q(m) = N(N+2)\sum_{k=1}^{m} \frac{r_k^2}{N-k},$$

where r_k is the autocorrelation at lag k of the residuals from the fit of the autoregressive-moving average model and N denotes sample size. Q has a χ^2 distribution with $m-p-q$ degrees of freedom.

For large N Q can be approximated by

$$Q(m) \approx N \sum_{k=1}^{m} r_k^2,$$

which, incidentally, is the original test statistic proposed by Box and Pierce (1970).

Now, let c be a critical value taken from the χ^2_{m-p-q} distribution, then for the hypothesis of white noise not to be rejected, i.e. $Q(m) < c$, it must hold for large N that

$$\sum_{k=1}^{m} r_k^2 < \frac{c}{N}.$$

Thus, as N increases linearly the tolerance allowed for in the Ljung-Box test statistic in terms of deviation from zero autocorrelation decreases at a rate proportional to $N^{-1/2}$. Consider as an example the $LB(15)$ test statistic referred to above. At a 5% significance level, $c = 25$. The table below gives an indication of the deviation allowed for $\sum_{k=1}^{m} r_k^2$:

Table 5.1. Deviations allowed for the sample ACF under the Ljung-Box test. c indicates the critical value.

$\boldsymbol{N}$	$\boldsymbol{c/N}$	$\boldsymbol{(c/N)^{1/2}}$
50	0.5	0.707
100	0.25	0.5
1000	0.025	0.158
10000	0.0025	0.05
100000	0.00025	0.016

If $\sum_{k=1}^{m} r_k^2$ must be smaller than c/N, then by implication

$$\max_{1 \le i \le k} r_i^2 < \frac{c}{N}.$$

Now for a sample size of $N = 10000$ the table implies that the maximum absolute deviation tolerated on the autocorrelation at any lag is 0.05, and then only if all other lags considered

were identically zero. While $|r_i| = 0.05$, for $1 \leq i \leq k$, would thus be considered statistically significant, the practical significance thereof is questionable. The paradox that arises is whether we are interested in statistical significance or practical significance. For a smaller sample, these concepts seem to coincide, but for a sample of the size in our example one could argue that they imply different hypotheses. For this reason we will consider four alternative methods of lack-of-fit assessment in order to test the hypothesis of practical significance.

Before we proceed with the alternatives we wish to raise another issue with regards to the Ljung-Box test. The choice of m has a big influence on the outcome of the test, but is often made arbitrarily. To a certain degree it gives the analyst the ability to manipulate the outcome of the hypothesis test to advance a favoured point of view. To illustrate the possible manipulation of results due to the choice of m we compare the power of $Q(m)$ in a simulation study for different choices of m. We will use the choices $m = 1$, 5, 10 and 15 for the orders of the Ljung-Box test. We choose $N = 500$ and generate 10000 realizations from an $MA(1)$ model for each value of θ over a grid covering the interval $[0,\ 0.4]$ in increments of 0.1. The results are shown in Table 5.2.

Table 5.2. Results from a power comparison between different order choices for the Ljung-Box test. Entries reflect the proportion of null hypotheses rejected at a given parameter level.

$\boldsymbol{\theta}$	**0**	**0.1**	**0.2**	**0.3**	**0.4**
m = 1	0.053	0.6155	0.9932	1	1
5	0.0473	0.3634	0.9471	1	1
10	0.0542	0.2718	0.8766	0.9993	1
15	0.0534	0.2302	0.8116	0.9981	1

Since the first order autocorrelation is the only order that is theoretically significant for an $MA(1)$ model it is to be expected that $m = 1$ would represent the most powerful test. Indeed, for $\theta = 0.1$ at a sample size as large as 500, the choice of $m = 1$ gives almost three times more power than $m = 15$. If we favoured a white noise conclusion as the outcome of the test and we suspected the presence of only first order autocorrelation, i.e. an $MA(1)$ type dependence, we could choose m larger to benefit our preconceived notion. To circumvent this difficulty we opt for a data driven choice of m proposed by Hart (1997, pp. 185–186).

Let Q be defined as above, then we choose $m = \hat{m}$ in the Ljung-Box test

$$\hat{m} = \arg\max \left[\hat{R}(m) \right]$$

and

$$\hat{R}(m) = \begin{cases} 0, & m = 0 \\ \sum_{j=1}^{m} \left(N r_j^2 \right) - 2m, & m = 1, 2, \ldots, N-1. \end{cases} \tag{7}$$

This choice, however, has implications for the distribution of Q. Making use of the fact that under the null hypothesis the distribution of r_j is approximately $N\left(0, N^{-1/2}\right)$ we can find the asymptotic distribution of $Q\left(\hat{m}\right)$. Let $Z_1, Z_2, \ldots$ be i.i.d. standard normal variates and define

$$S(k) = \sum_{j=1}^{k} \left(Z_j^2 - 2 \right), \ k = 1, 2, \ldots$$

Further, let $E_j(x)$ be defined to be the event

$$\left\{ 0 < S(j) < x - 2j, S(k) \leq S(j); k = 1, \ldots, j-1 \right\}, \ \ j = 1, 2, \ldots$$

and let

$$n_x = \left\lfloor \frac{x}{2} \right\rfloor.$$

Then under appropriate regularity conditions the distribution of $Q(\hat{m})$ asymptotically converges to that of the random variable τ where

$$P(\tau \leq x) = \begin{cases} 0, & x < 0 \\ 0.71, & 0 \leq x \leq 2 \\ 0.71 \left(1 + \sum_{j=1}^{n_x} P[E_j(x)] \right), & x > 2. \end{cases}$$

Critical values can be estimated by simulation. We now discuss four alternative approaches.

5.1.1 Parzen's Test

One alternative approach to lack-of-fit assessment is the use of an order selection criterion. Shibata (1976) found that when a process is white noise the Akaike Information Criterion (AIC) is minimized at 0 about 71% of the time , i.e. the AR model that minimizes the criterion would be a white noise, i.e. $AR(0)$. Parzen (1977) seems to have been the first one to propose the use of an order selection criterion as a lack-of-fit test by utilizing his criterion autoregressive transfer (CAT) function. Hart (1997) proposed a modification that allows the achievement of a specified significance level and it is this version of Parzen's test that we mention. Let $\hat{\sigma}^2(p)$ be the maximum likelihood estimate of the variance of the residuals from an AR(p) model. Then the modified $CAT^*(k)$ function is given by:

$$\begin{aligned} CAT^*(k) &= \begin{cases} -\left(\frac{1+\frac{q_\alpha}{N}}{\tilde{\sigma}^2(0)}\right), & k=0 \\ \frac{1}{N}\sum_{j=1}^{k}\left(\frac{1}{\tilde{\sigma}^2(j)}\right) - \frac{1}{\tilde{\sigma}^2(k)}, & k=1,2,...,K_N \end{cases} \\ \text{where } \tilde{\sigma}^2(i) &= \frac{N}{N-i}\hat{\sigma}^2(i),\ i=0,1,\ldots,N-1, \\ \frac{K_N^2}{N} &\to 0 \text{ as } N\to\infty \text{ and} \\ q_\alpha &= \text{ a coefficient that controls the significance level } \alpha. \end{aligned}$$

Values for q_α can be found in table 9.1 of Hart (1997). We reject the null hypothesis of white noise if and only if the value of k that minimizes $CAT^*(k)$ is larger than zero.

5.1.2 Natural partitions and simultaneous testing

A second approach to overcoming the large sample sensitivity of the Lung-Box test is to partition the data set into smaller subsets and do simultaneous white noise testing on the subsets. This raises two issues. The first one is the possibly arbitrary nature of the partitioning and the effect of possible dependence between partitions. The second one is the effective significance realized when performing simultaneous testing.

To address the issue of partitioning it is worth mentioning that the daily subsets of a UHF data set forms a "natural" partition because they are determined by the underlying process and not by the observer. The heterogeneity between trading days discussed in Chapter 2 also justifies the use of daily subsets as partitions of choice because the method we employ for simultaneous testing requires independence of the test statistics and this definitely cannot be achieved if a specific trading day contributes to more that one test statistic.

It is well known that in order to obtain a type I error level of α in simultaneous testing, we cannot merely test all individual hypotheses at the level α as the level obtained in this way is in fact substantially larger. One way of controlling the type I error is with a Bonferoni procedure. For g simultaneous hypotheses this entails testing each hypothesis at level α/g. There are a few difficulties associated with this procedure, foremost that this procedure tends to have less power, see Benjamini and Hochberg (1995). We opt for controlling the false discovery rate (FDR) as proposed by Benjamini and Hochberg (1995). The FDR is defined as

$$FDR = E[V/R]$$

where V is the number of true null hypotheses rejected and R is the total number of null hypotheses rejected.

The procedure is implemented as follows: Consider g null hypotheses $H_1, H_1, \ldots, H_g$ with corresponding p-values $P_1, P_2, \ldots, P_g$. Let $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(g)}$ be the ordered p-values let $H_{(i)}$ denote the null hypothesis corresponding to $P_{(i)}$. Let k be given the largest index, i, for which $P_{(i)} \leq \frac{i}{g}q^*$, where $i = 1, 2, \ldots, g$ and q^* denotes the desired FDR level, then we reject $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$. If $P_{(i)} > \frac{i}{g}q^*$, for all i, then $k = 0$.

We can now partition our UHF data set into daily subsets, and determine a p-value for each day using for example the Ljung-Box test. As the sample size for each subset is significantly reduced by partitioning, so is the effect of large sample size on the outcome of the test. If the proportion of null hypotheses rejected exceeds a chosen level q^*, we reject the simultaneous hypothesis of white noise. This method has the added benefit that we are not bound by the same model specification for different days. It allows us to choose a model

parameterization best suited to the characteristics of each partition rather than enforcing a single specification on a data set consisting of heterogeneous subsets. There are at least two reasons why our approach to the analysis of UHF data is suited to this approach to testing. We did not induce interdependence between the durations over different trading days with our method of diurnal adjustment (see Chapter 2) and we possibly use different parameter specifications for the returns of different trading days (see Chapter 3). For these reasons, amongst others, we employed the FDR method in conjunction with the $Q(\hat{m})$ statistic in our approaches, see Section 3.8 and Section 3.9.

5.1.3 Test based on discretionary choice of practical significance

A third method of accommodating the effect of excessively large sample sizes is to make a discretionary choice of practical significance limit. The idea is to specifically set up the goodness-of-fit hypothesis to answer the question of practical significance. If we only consider lag 1 first order autocorrelation, the standard null hypothesis for white noise would be

$$H_0 : \rho_1 = 0.$$

As previously discussed, the effect of very large sample size is that values of r_1 that are of little practical significance would cause us to reject the null hypothesis. If we were, however, to change the simple null hypothesis to a compound null hypothesis, we could overcome the difficulty. For this, we choose an interval with which we associate practical significance, i.e.

$$H_0 : \rho_1 \in [-c, c]$$

where $c > 0$ is the practical significance limit. We can now, for example, decide that an absolute autocorrelation of less that 0.05 is not to be considered practically significant and rather use the $\rho_1 \in [-0.05, 0.05]$ as our null hypothesis.

We now proceed with the derivation of a test statistic for this compound null hypothesis. Rao (1968) shows that the sample estimate r of the Pearson correlation coefficient ρ asymp-

totically has a normal distribution under the hyperbolic tangent transformation. Formally, let

$$z = \tanh^{-1}(r), \tag{8}$$

then asymptotically z has a normal distribution with

$$\begin{aligned} E(z) &= \tanh^{-1}(\rho) + \frac{\rho}{2(N-1)} \text{ and} \\ Var(z) &= \frac{1}{N-1} + \frac{4-\rho^2}{2(N-1)^2}. \end{aligned}$$

The result would not in general be directly be applicable to an autocorrelation as it is not calculated from independent pairs. We can, however, artificially construct such pairs under the null hypothesis. The first subset would consist of

$$\begin{aligned} &(x_1, x_2) \\ &(x_3, x_4) \\ &\vdots \\ &(x_{j-1}, xj) \end{aligned}$$

and the second subset would consist of

$$\begin{aligned} &(x_2, x_3) \\ &(x_4, x_5) \\ &\vdots \\ &(x_{k-1}, x_k) \end{aligned}$$

where

$$\begin{aligned} j &= N \text{ and } k = N-1, \text{ if } N \text{ is even and} \\ j &= N-1 \text{ and } k = N, \text{ if } N \text{ is odd.} \end{aligned}$$

Under the null hypothesis the pairs in each subset would be independent and we can proceed with the application of (8) to derive a test. For a specific choice of c the null hypothesis and alternative can be formulated as

$$\begin{aligned} H_0 &: \rho = c \\ H_A &: \rho > c. \end{aligned}$$

Let $i = 1, 2$ label the subsets. Then

$$\begin{aligned} 1-\alpha &= P(r_i < c_\alpha | \rho = c) \\ &= P(\tanh^{-1}(r_i) < \tanh^{-1}(c_\alpha) | \rho = c) \\ &\approx P\left(\frac{\tanh^{-1}(r_i) - \left(\tanh^{-1}(c) + \frac{c}{2(N_i-1)}\right)}{\sqrt{\frac{1}{N_i-1} + \frac{4-c^2}{2(N_i-1)^2}}} \right. \\ &\quad \left. < \frac{\tanh^{-1}(c_\alpha) - \left(\tanh^{-1}(c) + \frac{c}{2(N_i-1)}\right)}{\sqrt{\frac{1}{N_i-1} + \frac{4-c^2}{2(N_i-1)^2}}} \right). \end{aligned}$$

With the knowledge that under the null hypothesis

$$\frac{\tanh^{-1}(r_i) - \left(\tanh^{-1}(c) + \frac{c}{2(N_i-1)}\right)}{\sqrt{\frac{1}{N_i-1} + \frac{4-c^2}{2(N_i-1)^2}}} \xrightarrow{D} N(0,1)$$

we can solve c_α to find

$$c_\alpha = \tanh\left[\Phi(1-\alpha) \left(\frac{1}{N_i - 1} + \frac{4-c^2}{2(N_i-1)^2} \right)^{1/2} + \left(\tanh^{-1}(c) + \frac{c}{2(N_i-1)} \right) \right].$$

We thus reject H_0 if $r_i > c_\alpha$. This can be repeated for both subsets with additional consideration required if the conclusion for both subsets is not the same. If $r_i < 0$, we would apply the specified test to $|r_i|$.

The difficulty with employing this test lies in the arbitrary nature of the choice of c. The next test tries to address this by making the choice implicitly.

5.1.4 Test with significance level dependent on sample size

Finally, we consider a test that does not allow for the choice of the significance level, rather a significance level is implied as a function of sample size. The penalty, $2m$, employed in (7) is similar to the one used in the AIC. By replacing the AIC type penalty with a Schwartz Bayesian type penalty, see Hart (1997, p. 185), we obtain

$$B(m) = \begin{cases} 0, & m = 0 \\ \sum_{j=1}^{m} \left(N r_j^2\right) - m \log\left(N\right), & m = 1, 2, \ldots, N-1 \end{cases}$$

and similar to $\hat{m}$ we define

$$\tilde{m} = \arg\max\left[B(m)\right].$$

If $\tilde{m} \geq 1$ we reject the null hypothesis that the noise is white. It is apparent on inspection that the penalty imposed on $B\left(m\right)$ is now dependent on sample size. The advantage of a test based on $\tilde{m}$, compared to (7), is that critical values need not be simulated, with a corresponding reduction in runtime. We proceed with the derivation of an upper bound for the level of significance of the test.

Lemma 1 *The level of significance of the test which rejects when $\tilde{m} \geq 1$ is bounded from above by*

$$\sqrt{\frac{e \log(N)}{N}}.$$

Proof.

$$\begin{aligned} P\left(\textit{reject } H_0 \middle|\ H_0\right) &= P\left(\tilde{m} \geq 1 \middle|\ H_0\right) \\ &= P\left(\max_{1 \leq\ m \leq N-1} \frac{1}{m} \sum_{j=1}^{m} Z_j^2 > \log\left(N\right)\right) \\ &= P\left(\max_{1 \leq\ m \leq N-1} \frac{1}{m} \sum_{j=1}^{m} \left(Z_j^2 - 1\right) > \log\left(N\right) - 1\right) \end{aligned}$$

where Z_j are iid $N(0,1)$ under the null hypothesis. Now, the series

$$Y_m = \frac{1}{m}\sum_{j=1}^{m}\left(Z_j^2 - 1\right), \quad m = N-1, N-2, \ldots, 1$$

is a martingale, see Chow and Teicher (1988, p. 247). Hence, for $t > 0$, we have

$$\begin{aligned} & P\left(\max_{1\leq\, m\leq N-1} Y_m > \log(N) - 1\right) \\ = \;& P\left(\max_{1\leq\, m\leq N-1} \exp\left[tY_m\right] > \exp\left[t\left(\log(N) - 1\right)\right]\right) \\ \leq \;& \exp\left[-t\left(\log(N) - 1\right)\right] E\left(\exp\left[t\left(Z_1^2 - 1\right)\right]\right) \\ = \;& N^{-t}\left(1 - 2t\right)^{-1/2}, \end{aligned}$$

where the last inequality follows from the Doob inequalities. The right hand side above is minimized at $t = \frac{1}{2}\left(1 - [\log(N)]^{-1}\right)$. Thus,

$$\begin{aligned} P\left(\textit{reject } H_0 |\; H_0\right) &\leq N^{-\frac{1}{2}\left(1-[\log(N)]^{-1}\right)}\left[\log(N)\right]^{1/2} \\ &= \left[\frac{\log(N)}{N}\right]^{1/2} e^{1/2} \end{aligned}$$

and the result follows. ■

The upper bound cannot necessarily be expected to be sharp. The value of the bound lies in the fact that it converges to zero in the limit as sample size tends to infinity, i.e. for larger samples a smaller type I error is made. Of course, $P\left(\max_{1\leq\, m\leq N-1} Y_m > \log(N) - 1\right)$ could equally be estimated by simulation. As an illustration of the implied significance and power of the test we generated realizations from both $AR(1)$ and $MA(1)$ processes with zero mean and unit variance for the innovations. We used $N = 50$, 100, 100 and 10000 to coincide with the first four rows in Table 5.1 and selected ϕ and θ from a grid covering the interval $[0, 0.1]$ in increments of 0.01. For each grid point we generated 1000 realizations and applied the test. The results are summarized in Table 5.3.

Table 5.3.1. Results from simulation study on the power of a test based on $\tilde{m}$ for an $AR(1)$ model. Entries reflect proportions of null hypotheses rejected.

ϕ	**0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**	**0.1**
N = 50	0.053	0.048	0.044	0.051	0.052	0.05	0.051	0.058	0.063	0.07	0.079
100	0.034	0.035	0.036	0.035	0.038	0.048	0.056	0.065	0.068	0.081	0.104
1000	0.009	0.011	0.022	0.044	0.082	0.142	0.213	0.32	0.443	0.576	0.697
10000	0.003	0.024	0.145	0.484	0.83	0.977	0.998	1	1	1	1

Table 5.3.2. Results from simulation study on the power of a test based on $\tilde{m}$ for an $MA(1)$ model. Entries reflect proportions of null hypotheses rejected.

θ	**0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**	**0.1**
N = 50	0.048	0.051	0.056	0.061	0.064	0.069	0.078	0.086	0.096	0.11	0.115
100	0.033	0.032	0.038	0.044	0.05	0.063	0.074	0.086	0.098	0.113	0.129
1000	0.008	0.014	0.025	0.05	0.088	0.154	0.236	0.341	0.462	0.585	0.703
10000	0.002	0.021	0.155	0.483	0.838	0.975	0.999	1	1	1	1

Since power increases with both sample size and (absolute) autocorrelation the test is consistent. There are two interesting insights to be gained from Table 5.3. Firstly, for sample sizes larger than fifty the significance level is substantially lower than the standard 5%. If we compare the results from our simulation to the compound null hypothesis in the previous section, we are implicitly willing to accept first order correlation of between 0.04 and 0.06 for the standard 5% significance level at a sample size of 100. For the two larger sample sizes we are willing to accept autocorrelation of between 0.03 and 0.04. The test thus gives us an objective way of determining an amount of autocorrelation considered of practical significance rather that the arbitrary way in which it is determined in the discretionary significance limit test.

The second interesting insight is that for the larger parameter values in the table the test is more powerful for AR alternatives than for MA alternatives. Intuitively this makes

sense. If we consider the respective theoretical first order autocorrelations for $\phi = \theta$ we find

$$|\phi| \geq \left| \frac{-\theta}{1+\theta^2} \right|,$$

i.e. the (absolute) autocorrelation of the AR process is larger than that of the MA process for non-zero values of the respective parameters. Therefore a test based on autocorrelation should be able to identify the AR process with greater ease. The benefit does not seem to be noteworthy for small parameter values (in absolute value terms) when the differences between the two (absolute) autocorrelations become negligible. We would expect the phenomenon to continue outside the scope of Table 5.3 as the parameter values moves further away from the origin.

5.2 Moving Average unit root tests

As an alternative to the multiplicative diurnal effect model proposed by Engle and Russell (1998) we proposed differencing as a method for de-trending in Chapter 2. The fact that all trading days showed the characteristics of an $MA(1)$ process might be an indication that we were differencing a white noise series. This naturally leads us to consider tests for over-differencing, i.e. MA unit root tests. In particular, we will consider the work of Tanaka (1990) and Breitung (2002). For a more general discussion on MA unit root tests, see Breitung (1994).

For an $MA(1)$ model

$$x_t = \epsilon_t - \theta\varepsilon_{t-1}$$

where $|\theta| \leq 1$ and $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, 2, ...$ Tanaka (1990) tests

$$H_0 \ : \ \theta = 1 \text{ vs.}$$

$$H_A \ : \ \theta < 1.$$

The test statistic, S_N, for a time series containing N observations is given by

$$S_N = \frac{1}{N} \frac{\sum_{t=1}^{N} \left((t-1)x_1 + (t-2)x_2 + \ldots + x_{t-1} - \frac{t}{N+1} \sum_{s=1}^{N} (N-s+1)\, x_s \right)^2}{\sum_{t=1}^{N} \frac{1}{t(t+1)} (x_1 + 2x_2 + \ldots + tx_t)^2}.$$

Tanaka (1990) goes on to show that the test statistic has the following distribution: Let $\{Z_t\} \sim N(0,1)$, then

$$P(S_N < y) = P\left(\sum_{t=1}^{N} \xi_{t,N}(\theta, y) Z_t^2 > 0 \right),$$

$$\xi_{t,N}(\theta, y) = ((1-\theta)^2 + \theta\lambda_{t,N}) \left(\frac{y}{\lambda_{t,N}} - \frac{1}{N\lambda_{t,N}^2} \right)$$

$$\text{with } \lambda_{t,N} = 4\sin^2\left(\frac{t\pi}{2(N+1)} \right).$$

This probability can either be evaluated using a numerical approximation, see for example Imhoff (1961) or Martynov (1975), or by simulation. For large N, S_N can be approximated by

$$S_N \approx \frac{1}{N} \frac{\sum_{t=1}^{N} X_t^2}{\sum_{t=1}^{N} x_t^2}$$

where $X_t = \sum_{i=1}^{t} (x_t - \bar{x})$ and $\bar{x} = \frac{1}{N} \sum_{i=1}^{t} x_t$. The approximation is useful for large sample sizes because of its relative computational convenience.

Curiously, Breitung (2002) quotes Tanaka's test only in the latter form under zero mean, and makes the assertion that Tanaka formulates the null as

$$H_0 : \theta < 1.$$

To adjust for a non zero mean of the form $d_t = \delta' z_t$ Breitung proposes as test statistics

$$\hat{\varrho}_N = \frac{1}{N} \frac{\sum_{t=1}^{N} \hat{U}_t^2}{\sum_{t=1}^{N} \hat{u}_t^2}$$

where $\hat{u}_t = x_t - \hat{\delta}' z_t$, $\hat{U}_t = \sum_{i=1}^{t} \hat{u}_t$ and $d = 0$ if y_t has zero mean. Under certain assumptions proposition 3 Breitung (2002), claims the distribution of $\hat{\varrho}_N$ to be

$$N^{-1}\hat{\varrho}_N \Rightarrow \frac{\int_0^1 \left[\int_0^a \tilde{W}_j(s)ds\right]^2 da}{\int_0^1 \tilde{W}_j(a)^2 da}$$

where

$$\begin{aligned}
&\tilde{W}_0(s) \equiv W(s), && \text{for } d_t = 0, \\
&\tilde{W}_1(s) \equiv W(s) - \int_0^1 W(a)da, && \text{for } d_t = 1, \\
&\tilde{W}_2(s) \equiv W(s) - (4-6s)\int_0^1 W(a)da - (12s-6)\int_0^1 aW(a)da, && \text{for } d_t = [1, t]',
\end{aligned} \tag{9}$$

which can only be approximated by simulation. In (9), $W(.)$ denotes a standard Wiener process. Table 5 in Breitung (2002) supplies approximations to critical values for the test statistic. Breitung states: "The critical values are computed from the empirical distribution of 10000 realization of the limiting expressions of the test statistic, with Gaussian random walk sequences instead of Brownian motions (Tables 5 and 6)." This experiment is, however, implausible as the limiting distribution does not depend on sample size and would not explain why Breitung has critical values dependent on sample size. More significantly, the test statistic S_T as defined by Breitung (2002) does not have the indicated limiting distribution. If we were rather to set

$$\varepsilon_0 = 0$$

and define

$$\hat{u}_t = \sum_{i=1}^{t} (x_t - \hat{\delta}' z_t)$$

then under the null hypothesis for a zero mean process this renders

$$\begin{aligned} \hat{u}_t &= \sum_{i=1}^{t} (\varepsilon_i - \varepsilon_{i-1}) \\ &= \varepsilon_t. \end{aligned}$$

Under this alternative formulation for $\hat{u}_t$ the test statistic has the required limiting distribution indicated in (9). This would also explain why the asymptotic distributions differ for Breitung and Tanaka's test statistics even though Breitung claims the two test statistics to be equivalent for a process with zero mean. Tanaka's test statistic uses the partial sum while Breitung's test uses the partial sum of partial sums. A simulation study where the alternative formulation for Breitung's test statistic was computed under the null hypothesis for time series of lengths 100, 250 and 500 respectively delivers results similar to those shown by Breitung in his Table 5. We applied Tanaka's test for testing for the presence of an MA unit roots throughout this thesis.

5.3 $GARCH$ tests

A commonly used test for the presence of conditional heteroskedasticity in the innovations is the Lagrange multiplier (LM) test; see Engle (1982). Lee (1991) showed that the LM test for $ARCH(m_0)$, for some order $m_0 \geq 1$, is the same as the LM test for $GARCH(r, m_0)$, for any $r \geq 0$. Hence, without lack of generality, our discussion highlights a single deficiency in the former. For a discussion of other tests for $ARCH$ see for example Duchesne (2004) or Dufour, et al. (2004).

The LM test for $ARCH(m)$ uses linear regression to fit the model

$$\varepsilon_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_m \varepsilon_{t-m}^2 \tag{10}$$

to some uncorrelated series $\{\varepsilon_t,\, t \in \mathcal{Z}\}$. The hypothesis tested is then

$$\begin{aligned} H_0 \;&:\; \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0 \\ &\quad \text{against} \\ H_A \;&:\; \alpha_i \neq 0, \text{ for some } i \in \{1, 2, \ldots, m\} \end{aligned}$$

and the test statistic is NR^2, where R^2 denotes the multiple correlation coefficient and N denotes the sample size. Asymptotically $NR^2 \sim \chi^2$ with m degrees of freedom. The deficiency of the LM test lies in the fact that R^2 fails to recognize the one-sided nature of the alternative hypothesis, i.e. the α's are constrained to be non-negative for $ARCH$, with the implication that the alternative should rather be

$$H_A : \alpha_i > 0, \text{ for some } i \in \{1, 2, \ldots, m\}.$$

The scenario can be illustrated more clearly with the following example.

Example 1 *Let x be a N(μ,1) random variable and assume we want to test the hypothesis*

$$\begin{aligned} H_0 \quad &: \quad \mu = 0 \\ &\textit{against} \\ H_A \quad &: \quad \mu > 0 \end{aligned}$$

by making use of x^2 in some form. To merely conclude that under the null hypothesis $x^2 \sim \chi^2(1)$ ignores the one sided nature of the alternative and the type I error would be double the intended α. Rather, we should use the test statistic

$$u = \begin{cases} 0, & \textit{if } x < 0 \\ x^2, & \textit{if } x \geq 0 \end{cases},$$

and the distribution of u would have an atom at zero, i.e. $u \sim \frac{1}{2} + \frac{1}{2}\chi^2(1)$. With the test statistic u we would then obtain the intended type I error.

Following Gourieroux, et al. (1982) we will make use of the Kuhn-Tucker multiplier test as opposed to the LM test in model (10) to exploit the one-side nature of the alternative hypothesis. By a similar argument to the one used in the example the LM test also achieves a type I error substantially higher than intended. For a Kuhn-Tucker test of order q the test statistic has a mixture of χ^2-distributions of degrees of freedom 1 through q and an atom at the origin. The weights of the mixture depends on the order and becomes complicated to calculate for moderate orders, see Gourieroux, et al. (1982).

As the first order Kuhn-Tucker multiplier test is of specific importance in our application we explicitly mention it. We want to test

$$\begin{aligned} H_0 \quad &: \quad \alpha_1 = 0 \\ &\text{against} \\ H_A \quad &: \quad \alpha_1 > 0. \end{aligned}$$

α_1 is the estimated by a regression of ε_t^2 on its first lag. The test statistic is

$$\xi_{KT} = \begin{cases} 0 & \text{if } \hat{\alpha}_1 < 0 \\ \frac{\hat{\alpha}_1^2}{\hat{Var}(\hat{\alpha}_1)} & \text{otherwise.} \end{cases}$$

Asymptotically $\xi_{KT} \sim \frac{1}{2} + \frac{1}{2}\chi^2(1)$.

We compare the power of the two tests in a simulation study. We generated 10000 $ARCH(1)$ realizations for each α_1 on a grid over the interval $[0, 0.45]$ in increments of 0.05. We took $N = 50$, 100, 500 and 1000. Both tests were applied at a 5% level of significance. The results are reflected in Table 5.4.

Table 5.4.1. Results from simulation for the power of the LM test for $ARCH(1)$. Entries

reflect proportions of null hypotheses rejected.

α_1	**0**	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**
N = 50	0.108	0.132	0.164	0.207	0.252	0.306	0.345	0.395	0.438	0.474
100	0.112	0.157	0.222	0.314	0.408	0.501	0.574	0.64	0.706	0.758
500	0.117	0.276	0.576	0.807	0.929	0.978	0.995	0.998	1	1
1000	0.112	0.405	0.818	0.97	0.997	1	1	1	1	1

Table 5.4.2. Results from simulation for the power of the Kuhn-Tucker test for $ARCH(1)$. Entries reflect proportions of null hypotheses rejected.

α_1	**0**	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**
N = 50	0.043	0.075	0.128	0.186	0.24	0.306	0.356	0.411	0.463	0.5
100	0.048	0.118	0.213	0.326	0.424	0.527	0.606	0.674	0.742	0.789
500	0.051	0.285	0.611	0.838	0.945	0.984	0.997	0.999	1	1
1000	0.052	0.432	0.848	0.98	0.999	1	1	1	1	1

Once again, it is clear that the two tests are consistent. A series with $\alpha_1 = 0$ corresponds to a white noise series, hence, the associated proportions approximate the level of significance. As predicted by our example, the level of significance associated with the LM test is close to double the intended, while that of the Kuhn-Tucker test is as anticipated. The Kuhn-Tucker test is also more powerful as the expected rate at which the power increases with departure from the null hypothesis is higher. For instance, when considering the two larger sample sizes, the Kuhn-Tucker test already has more power by $\alpha_1 = 0.05$, not withstanding the "head start" of the LM test. For these reasons we applied the Kuhn-Tucker test in our analysis.

5.4 Markov chain order selection

As this was of particular interest to us in the formulation of our model for asset prices (see Chapter 3), we consider two methods for selection of the order of dependence in a Markov chain, namely selection by sequential elimination and selection by (auto)logistic regression.

5.4.1 Markov chain order selection by sequential elimination

Hoel (1954) and Lowry and Guthrie (1968) proposed tests based on sequential testing for the suitability of reducing the order of a Markov chain by one. The idea is to start with a high but manageable order and then sequentially reduce the order until either the test statistic is significant, or you reach an order where it is senseless to reduce any further. Both these tests require counts of all possible unique sequences of k transitions where k is the order of the chain. This requirement is cumbersome. As an illustration, consider a two state chain where if we want to test the null hypothesis that a fourth order chain can be represented as third order, we have to enumerate $2^5 = 32$ unique sequences of five states. In general if the higher of the two orders are k we need to enumerate 2^{k+1} sequences of $k + 1$ states. This complicates the application of these tests since we cannot in general expect to find all of the unique sequences at least once. Both tests are unable to handle a situation where any unique sequence never occurs. For this reason we opt for a second approach.

5.4.2 Markov chain order selection by logistic regression

Following Cox and Snell (1989) we will select the order of a binary Markov chain by fitting an (auto)logistic regression model to the chain using lagged observations as covariates. There are two ways in which we could use logistic regression to do the order selection. The first one is related to the test statistic, usually t or F, often quoted for parameters by statistical software packages. We would find confidence intervals for our parameter estimates, possibly using the bootstrap, and then test our estimates to assess whether they are significantly different from zero. Care should be taken when using this approach as the confidence intervals achieved in this fashion are only valid in the presence of all the other parameters included. Thus, if k is the largest lag that has parameter estimate significantly different from zero it cannot merely be concluded that the chain represents a k-th order Markov chain. It would be necessary to eliminate the highest order sequentially, akin to the method mentioned in the previous section.

The second method would be to test the hypothesis that higher order parameters in the logistic regression are simultaneously insignificant with the help of a likelihood ratio test. As with logistic regression the test statistic only asymptotically has a χ^2-distribution, hence we will bootstrap the percentiles of our test statistic. We proceed with a description of logistic regression model fitting and critical value estimation in our application.

Parameters will be estimated with maximum likelihood using the following design: Let $m_1, m_2, \ldots, m_N$ represent the binary chain of which we want to determine the order. We first have to decide on the maximum order that we want to consider. Let this order be represented by d. The choice is somewhat arbitrary, but if our initial choice of d were found to be too small, i.e. the order of the chain is close to or exactly d the analysis could always be repeated with a larger choice of d. For the response vector we use

$$\mathbf{X} = \begin{bmatrix} m_{d+1} \\ m_{d+2} \\ \vdots \\ m_N \end{bmatrix}$$

and for the covariate matrix we use

$$\mathbf{Y} = \begin{bmatrix} 1 & m_d - 1/2 & m_{d-1} - 1/2 & \cdots & m_1 - 1/2 \\ 1 & m_{d+1} - 1/2 & m_d - 1/2 & \cdots & m_2 - 1/2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & m_{N-1} - 1/2 & m_{N-2} - 1/2 & \cdots & m_{N-d} - 1/2 \end{bmatrix}.$$

where the adjustment (subtracting a half) ensures that parameters contribute for both states. Utilizing this design we can now fit the model

$$\log\left\{\frac{P(m_i = 1)}{P(m_i = 0)}\right\} = \beta_0 + \beta_1\left(m_{i-1} - \frac{1}{2}\right) + \ldots + \beta_l\left(m_{i-d} - \frac{1}{2}\right)$$

by maximum likelihood. The log likelihood function is

$$\begin{aligned} l(\boldsymbol{\beta}|\mathbf{X},\mathbf{Y}) &= \sum_{i=d+1}^{N} \left(m_i y_i - \log(1+e^{y_i})\right), \\ \text{where } y_i &= \beta_0 + \beta_1 \left(m_{i-1} - \frac{1}{2}\right) + \ldots + \beta_l \left(m_{i-d} - \frac{1}{2}\right). \end{aligned}$$

The variance estimator for $\hat{\beta}_j$, $j = 1, 2, \ldots, d$, is

$$\begin{aligned} S^2(\hat{\beta}_j) &= -\left(\frac{\partial^2 l}{\partial \beta_j^2}|_{\beta=\hat{\beta}}\right)^{-1} \\ &= \left\{\sum_{i=d+1}^{N} m_{i-j}^2 \frac{e^{y_i}}{1+e^{y_i}}\left(1 - \frac{e^{y_i}}{1+e^{y_i}}\right)\right\}^{-1}. \end{aligned}$$

To assess whether the Markov chain is of order k we fit a logistic regression model of order k and order d, say, with $d > k$. Let us for the moment call the former the reduced model and the latter the full model. Using the respective log likelihood function values obtained in the respective maxima we can compute a likelihood ratio statistic.

Asymptotically the likelihood ratio statistic would have a χ^2-distribution. However, since we partitioned data sets, our samples might not necessarily be sufficiently large to rely on asymptotic theory. We can, however, estimate quantiles for our likelihood ratio statistic using the bootstrap and in the following way: We generate B realizations from the reduced logistic regression model each of length N. For each realization we fit both the the full and the reduced model obtaining resampled maximized log likelihood function values l_i^{f*}, $i = 1, 2, \ldots B$ and l_i^{r*}, $i = 1, 2, \ldots B$ respectively. With the help of these maximized log likelihood values we can now estimate the quantiles of interest for our likelihood ratio. Finally, we compare computed likelihood ratio statistic with the quantile of interest to determine whether or not the chain of order d can be sufficiently explained by a chain of order k.

Epilogue

This study considered the modelling of UHF financial data. We found that the assumption of a homogeneous diurnal effect is contradicted by empirical evidence. To remedy we proposed the use of differencing as preferred method of diurnal adjustment. The implication of employing differencing is that duration models, which by implication require strictly non-negative observation, are not suitable. We explicitly model the evolution of the diurnal effect over time by approximating the diurnal curves of individual trading days, reducing their dimension and modelling the evolution of coefficients with a VAR model. To account for the heterogeneity found over different trading days we use a unique regime for each trading day for the modelling of both durations and return.

Of course, when modelling differenced durations we are not guaranteed that data generated from such a model will render positive durations, i.e. partial sums will not necessarily be positive. We proposed methods to address this. We found that an $MA(1)$ model, possibly including a $GARCH$ specification for residuals, adequately describes the differenced duration series.

For the modelling of log returns we introduced the idea of mixing a Markov chain with a continuous state space model in (3). We employed two parameterizations for dynamic returns. Firstly, we used a standard $ARMA-GARCH$ specification, (4), and secondly, we applied the same specification to returns standardized for duration, (5). In the application of these models we found the modelling of standardized residuals to be superior in the sense that it was able to describe the data as successfully as we could with the modelling of non-standardized residuals, but with substantially less parameters. For three of the four data sets considered these models were able to convince a goodness-of-fit test at a 5% significance level. However, for neither approaches were the results from modelling AGL able to do the same. Other parameterizations for the dynamic returns process therefore remains an avenue for future investigation. We mention a few:

We opted for the EM Algorithm as augmentation scheme to address unobserved dynamic returns. The Algorithm requires the use of maximum likelihood for parameter estimation.

This automatically excludes the use of the temporal aggregation formula, mentioned in Section 1.2, because the formula cannot be applied in conjunction with maximum likelihood estimation. Since standard errors were rather large in our approach other methods of parameter estimation in the presence of missing values should be considered. This could also potentially open the door to specifications using the temporal aggregation formula.

In the standardization of returns, (5), we assume a specific functional form that relates duration to variance in returns. Although the assumption of variance proportional the duration is acceptable practice, it is quite conceivable that the relationship is not quite that simple. The possibility that there is some form of decay in the factor of proportionality as duration increases could potentially improve fit.

In the exploration of interdependencies of characteristics associated with stock data (Section 4.4) we did not find conclusive evidence to support a single market microstructure model for all stocks considered. This raises the question whether some models apply to specific categories of stocks, e.g. a specific sector or trade frequency category. We also found that more data was required for an analysis of cell probabilities of the simultaneous realization of characteristics. Repeating our analysis for a period longer than the four months we considered and for a larger number of shares could potentially shed more light on both issues.

It is our hope that this thesis makes a significant contribution to the literature on UHF models by advocating, amongst others, the use of mixture state space models for the returns process and relaxing the assumption of homogeneous trading days.

References

[1] Admati, A.R. and Pfleiderer, P. (1988). A theory of intraday patterns: Volume and price variability. *The Review of Financial Studies.* **1**, 3-40.

[2] Andersen, T.G. and Bollerslev, T. (1998). Deutsche mark–dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer run dependencies. *Journal of Finance.* **53**, 219-265.

[3] Audrino, F. (2005). Local likelihood for non-parametric ARCH(1) models. *Journal of Time Series Analysis.* **26**, 251-278.

[4] Bacidore, J., Battalio, R.H. and Jennings, R.H. (2003). Order submission strategies, liquidity supply, and trading in pennies on the New York Stock Exchange. *Journal of Financial Markets.* **6**, 337-362.

[5] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B.* **57**, 289-300.

[6] Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy.* **81**, 637-654.

[7] Bollerslev, T., Chou, R.Y. and Kroner K.F. (1992). ARCH modeling in finance: A review of the theory and empirical evidence. *Journal of Econometrics.* **52**, 5-59.

[8] Box, G.E.P. and Pierce, D.A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association.* **65**, 1509-1526.

[9] Breitung, J. (1994). Some simple tests of the moving-average unit root hypothesis. *Journal of Time Series Analysis.* **15**, 351-370.

[10] Breitung, J. (2002). Nonparametric tests for unit roots and cointegration. *Journal of Econometrics.* **108**, 343-363.

[11] Brémaud, P. (1981). *Point processes and queues, martingale dynamics.* New York. Springer-Verlag.

[12] Chow, Y.S. and Teicher, H. (1988). *Probability theory: independence, interchangeability, martingales*, 2nd edition. New York. Springer-Verlag.

[13] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association.* **74**, 829-836.

[14] Cont, R. and Tankov, P. (2004). *Financial modeling with jump processes.* Boca Raton. Chapman and Hall / CRC.

[15] Cox, D.R. and Isham, V. (1980). *Point Processes.* London. Chapman.

[16] Cox, D.R. and Snell E.J. (1989). *Analysis of binary data*, 2nd edition. London. Chapman and Hall.

[17] Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap methods and their applications.* Cambridge. Cambridge University Press.

[18] Diamond, D.W. and Verrecchia, R.E. (1987). Constraints on short-selling and asset price adjustment to private information. *Journal of Financial Economics.* **18**, 277-311.

[19] Dempster, A.P., Laird, N.M. and Rubin D.B. (1977). Maximum likelihood of incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B.* **39**, 1-38.

[20] Duan, J.C. (1995). The GARCH option pricing model. *Mathematical Finance.* **5**, 13-32.

[21] Duchesne, P. (2004). On robust testing for conditional heteroscedasticity in time series models. *Computational Statistics and Data Analysis.* **46**, 227-256.

[22] Dufour, J.M., Khalaf, L., Bernard, J.T. and Genest, I. (2004). Simulation-based finite-sample tests for heteroskedasticity and ARCH effects. *Journal of Econometrics.* **122**, 317-347.

[23] Easley, D. and O'Hara, M. (1992). Time and the process of security price adjustment. *Journal of Finance.* **47**, 577-605.

[24] Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica.* **50**, 987-1008.

[25] Engle, R.F. (2000). The econometrics of ultra-high-frequency data. *Econometrica.* **68**, 1-22.

[26] Engle, R.F. and Dufour, A. (2000). Time and price impact of trade. *Journal of Finance.* **6**, 2467-2498.

[27] Engle, R.F. and Russell, J.R. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica.* **66**, 1127-1162.

[28] Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications.* London. Chapman and Hall.

[29] Francq, C. and Zakoïan, J.M. (2004). Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli.* **10**, 605-637.

[30] Ghysels, E. and Jasiak, J. (1997). GARCH for irregularly spaced financial data: The ACD-GARCH model. *Studies in Nonlinear Dynamics and Econometrics.* **2**, 133-149.

[31] Glosten, L.R. and Milgrom, P.R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics.* **14**, 71-100.

[32] Gouriéroux, C., Holly, A. and Monfort, A. (1982). Likelihood ration test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. *Econometrica.* **50**, 63-80.

[33] Gouriéroux, C. and Jasiak (2001). *Financial econometrics.* Princeton. Princeton University Press.

[34] Gramming, J. and Maurer, K. (2000). Non-monotonic hazard functions and the autoregressive conditional duration model. *Econometrics Journal.* **3**, 16-38.

[35] Green, P.E. (1978). *Analyzing multivariate data.* Hinsdale. Dryden Press.

[36] Hamilton, J.D. (1994). *Time series analysis.* Princeton. Princeton University Press.

[37] Hansen, P.R. and Lunde, A.L. (2005). A forecast comparison of volatility models: does anything beat GARCH(1,1)? *Journal of Applied Econometrics.* **20**, 873-889.

[38] Hart, J.D. (1997). *Nonparametric smoothing and lack-of-fit tests.* New York. Springer.

[39] Harvey, A.C. (1993). *Time series models.* 2nd edition. London. Prentice Hall.

[40] Hasbrouck, J. (1991). Measuring the information content of stock trades. *Journal of Finance.* **46**, 179-207.

[41] Hautsch, N. (2004). *Modelling irregularly spaced financial data.* Berlin. Springer-Verlag.

[42] Hoel, P.G. (1954). A test for Markoff chains. *Biometrika.* **41**, 420-433.

[43] Imhoff, J.P. (1961). Computing the distributions of quadratic forms in normal variables. *Biometrika.* **48**, 419-426.

[44] Lee, J.H.H. (1991). A Lagrange multiplier test for GARCH models. *Economics Letters.* **37**, 265-271.

[45] Lowry, W.P. and Guthrie, D. (1968). Markov chains of order greater that one. *Monthly Weather Review.* **96**, 798-801.

[46] Ljung, G.M. and Box, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika.* **65**, 297-303.

[47] Madhavan, A. (2000). Market microstructure: A survey. *Journal of Financial Markets.* **3**, 205-258.

[48] Martynov, G.V. (1975). Computation of distribution functions of quadratic forms of normally distributed random variables. *Theory of Probability and its Applications.* **20**, 782-793

[49] McLachlan, G.J. and Krishnan, T. (1996). *The EM algorithm and extensions.* New York. John Wiley and Sons.

[50] Nelson, D.B. (1990). Arch models as diffusion approximations. *Journal of Econometrics.* **45**, 7-38.

[51] O'Hara, M. (1995). *Market microstructure theory.* Cambridge. Blackwell Publishers Ltd.

[52] Parzen , E. (1963). On spectral analysis with missing observations and amplitude modulation. *Sankhyā, Series A.* **25**, 383-392.

[53] Parzen, E. (1977). Multiple time series: determining the order of approximating autoregressive schemes. In *Multivariate Analysis - IV* (P. Khrishnaiah, ed.). Amsterdam. North-Holland. 283-295.

[54] Priestley, M.B. (1981). *Spectral analysis and time series.* Volume 1. London. Academic Press.

[55] Rao, C.R. (1968). *Linear statistical inference and its applications.* New York. John Wiley and Sons.

[56] Russell, J.R. (1999). Econometric modeling of multivariate irregularly-spaced high-frequency data. Discussion paper. Graduate School of Business. University of Chicago.

[57] Russell, J.R. and Engle, R.F. (2002). Econometric analysis of discrete-valued irregularly-spaced financial transaction data. Unpublished manuscript.

[58] Russell, J.R. and Engle, R.F. (2005). A Discrete-state continuous-time model of financial transaction prices and times: The autoregressive conditional multinomial – autoregressive conditional duration model. *Journal of Business and Economic Statistics.* **23**, 166-180.

[59] Rydberg, T.H. (2000). Realistic Statistical Modelling of Financial Data. *International Statistical Review.* **68**, 233-258.

[60] Rydberg, T.H. and Shephard, N. (2000). A modelling framework for the prices and times of trades made on the New York Stock Exchange. In *Nonlinear and nonstationary signal processing* (eds. Fitzgerald, W.J., Smith, R.L., Walden, A.T. and Young, P. C.). Cambridge. Cambridge University Press. 217-246.

[61] Rydberg, T.H. and Shephard, N. (2003). Dynamics of trade-by-trade price movements: decomposition and models. *Journal of Financial Econometrics.* **1**, 2-25.

[62] Schoutens, W. (2003). *Lévy processes in finance.* London. John Wiley and Sons.

[63] Shibata, R. (1976). Selection of the order of an autoregressive model by Akaiki's information criterion. *Biometrika.* **63**, 117-126.

[64] Tanaka, K. (1990). Testing for a moving average root. Econometric Theory. **6**, 433-444.

[65] Tay, A., Ting, C., Tse, Y.K. and Warachka, M. (2004). Transaction-data analysis of marked durations and their implications for market microstructure. Working paper. Singapore Management University.

[66] Tsay, R.S. (2002). *Analysis of financial time series.* New York. John Wiley and Sons.

[67] Veredas, D., Rodriguez-Poo, J. and Espasa, A. (2002). On the (intradaily) seasonality and dynamics of a financial point process: A semiparametric approach. CORE Discussion Paper 2002/23. Université Catholique de Louvain.

[68] Weisstein, E.W. (1999). Chebyshev polynomial of the first kind. From *MathWorld – A Wolfram Web Resource.* (http://mathworld.wolfram.com/ChebyshevPolynomialofthe FirstKind.html).