Correlation methods in the statistical analysis of financial trading data

Yang Wu Azzollini

Lady Margaret Hall

Hilary Term, 2016

Thesis submitted for the degree of Doctor of Philosophy at the University of Oxford

Correlation methods in the statistical analysis of financial trading data

Yang Wu Azzollini

Lady Margaret Hall, Oxford

This thesis considers problems associated with the statistical analysis of correlation in financial trading data. Sources of data are identified and their characteristics are described. Nowadays most financial transactions are carried out electronically on automated exchanges or electronic communication networks and active participants in the market require sophisticated computing infrastructure to compete effectively. These data are shown to present novel statistical challenges both in retrospectively analysing the vast stores of accumulated historical data and also in online processing of high-bandwidth multiple data streams arriving on millisecond time-scales. We show that computational speed dictates the range of statistical tools that are available for high-speed calculation.

The measurement and interpretation of correlation is a dominant concern in the analysis of high-frequency financial data. We compare and develop methods for assessing volatility, cross-asset correlation and lead-lag effects. For volatility estimation we consider a class of estimates that can accommodate noisy irregularly spaced data. We derive explicit expressions for the variance of these estimators and show how the estimators can be modified to obtain infill consistency. We then consider the problem of covariance estimation and develop a new estimator, demonstrating its superior performance. We explore the problems of quantifying lead-lag relationships and show that our new covariance estimator provides a sharper estimate of lead-lag delay. We then develop a method of exploring lead-lag structure in depth and demonstrate how to obtain a maximum likelihood estimator of the delay structure. The final chapter briefly describes ongoing research questions relating to the design of hedging strategies at times of market disruption.

Dedication

To my parents, Wu Feng Zhen and Yang Huan Li, who have been there for me from day one, taking the blows and giving me a chance to thrive.

To my husband Domenico, who took my parents' baton and who gives me his merits of loyalty, friendship and passion for life.

To my daughter Julia, who has been the joy of my life.

This is a tribute to the four of you. Thank you for all of your encouragement, love and dedication.

Acknowledgments

Firstly, I would like to express my deep gratitude to my outstanding supervisor Dr. Peter Clifford for his clear and thorough teaching style, for his patient guidance and valuable suggestions during the development of this research, and for his assistance in keeping my progress on schedule. I could not have imagined having a better supervisor for my DPhil. Peter was the most influential person and mentor to me for over a decade of study in Oxford since I was an undergraduate. One day, I hope to be just like him, and to have his infinite wisdom and immense knowledge in statistics.

Besides Dr. Peter Clifford, I would like to thank to my other supervisor Prof. Brian D. Ripley, who supervised me for two and half years before his early retirement, for his office door was always open for me and for his valuable guidance. Brian encouraged me to reach out and learn from various branches of statistics and broadened my statistical techniques.

I would like to express my sincere gratitude to the industry experts who were involved in the validation for this research, Tim Sillitoe and Peter Rading, for their extraordinary support. Without their participation and countless input, this research could not have been successfully conducted. I would also like to thank Domenico Azzollini who gave expert advice and made sure that all the financial trading examples given in the thesis were accurate and precise.

This research would have been impossible without the support of the Royal Bank of Scotland, for the scholarship, the data, the trading infrastructure, and the working opportunities.

Finally, I must express my very profound gratitude to my friends, Reg Mehta, Fiona, Russell and Sophie Bickerton for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Contents

Chapter 1

Introduction

Nowadays most financial transactions are carried out electronically on automated exchanges or electronic communication networks (ECNs) and active participants in the market require sophisticated computing infrastructure to compete effectively. This is particularly the case when dealing with the most heavily traded financial instruments such as government bond, foreign exchange, equity index and interest rate futures. *High-frequency data*, known also as *tick-by-tick data*, are the transaction-by-transaction or quote-by-quote events that take place in a time interval from a few milliseconds to a few seconds. These data present novel statistical challenges both in retrospectively analysing the vast stores of accumulated historical data and also in online processing of high-bandwidth multiple data streams arriving on millisecond time-scales.

The measurement and interpretation of correlation is a dominant concern in the analysis of high-frequency financial data. There is substantial research activity in this area covering topics such as correlation between non-synchronous trade data; correlation within order book data – from the perspective of a market marker, hedger, etc. ; the effect of feedback

from algorithmic trading strategies; changes in correlation and liquidity as a result of large price movements; the classification and detection of autocorrelation and mean reversion; correlation and volatility term structure, the Epps effect; canonical correlation; vector-valued autoregression; cointegration; construction of sparse portfolios, e.g. calendar and butterfly spreads, and a great deal more.

The contributions of this thesis are in three areas: the comparison and development of methods for assessing volatility, cross-asset correlation and lead-lag effects. In Chapters 2 and 3 we set the background, describing the sources of data and the types of statistical tools available for high-speed trading. We recognize that high-frequency analysis means high-speed analysis of financial data and this places severe constraints on the type of estimators that are feasible. The need for fast calculation in real time leads to recursively defined estimators and we show how these can be built for volatility and covolatility estimation in Chapters 4 and 5. Throughout we emphasise the importance of working on a variety of different time scales rather than simple 'wall clock' time.

In Chapter 4 we focus on the estimation of volatility, looking particular at a class of estimates that can accommodate irregularly spaced data. We derive new expressions for the variance of these estimators and show how they can be modified to obtain infill consistency. We provide explicit expressions for the variance of a basic volatility estimator proposed by Zhou (1996) and report progress on the development of a symbolic algebra package to handle more general volatility estimators in an extended class. In Chapter 5 we move on to consider the problem of covariance estimation and show that an early contribution by Zhou (1995) contains many of the ideas that have recently become popular. We develop a new covariance estimator and demonstrate its superior performance.

Chapter 6 explores the problems of quantifying lead-lag relationships. We show that

the extended covariance estimator developed in Chapter 5 provides a sharper estimate of lead-lag delay. We then develop a method of exploring lead-lag structure in depth and demonstrate how to obtain a maximum likelihood estimator of the delay structure. Chapter 7 briefly describes ongoing research questions relating to the design of hedging strategies at times of market disruption.

Chapter 2

Financial background

Successive improvements in the availability of high quality data have led to increasingly detailed pictures of the trade and order placement activity in financial markets around the world. Although historical datasets exist dating back to the nineteenth century, these are of low resolution: for example, the daily values of the Dow Jones Industrial Average (DJIA) at the end of each trading day. The frequency of data collection has shifted from daily to minute, second and currently microsecond resolution – all in the context of the multiplication of trading venues both nationally within the developed economies and internationally as new exchanges open. The integration of such vast streams of high-frequency data into effective trading strategies has driven developments in computing infrastructure, statistical analysis, algorithm design and high-speed electronic communication.

2.1 Historical volatility

The assessment of market volatility is important for investors, not least in providing an indication of the chance that the value of an asset will drop below an unacceptable level at some point in the future. Financial historians have devoted considerable effort to understanding the causes of volatility, focussing particularly on events surrounding dramatic falls in the stock market. Although the primary focus of this thesis is on high-frequency data, it is of interest to describe some of the historical events that have shaped modern developments in volatility research.

Rank	Date	% Chg	Rank	Date	% Chg	Rank	Date	% Chg
1	1987-10-19	-22.61	7	1907-03-14	-8.29	13	2008-10-09	-7.33
2	1929-10-28	-12.82	8	1987-10-26	-8.04	14	1917-02-01	-7.24
3	1929-10-29	-11.73	9	2008-10-15	-7.87	15	1997-10-27	-7.18
4	1929-11-06	-9.92	10	1933-07-21	-7.84	16	1932-10-05	-7.15
5	1899-12-18	-8.72	11	1937-10-18	-7.75	17	2001-09-17	-7.13
6	1932-08-12	-8.40	12	2008-12-01	-7.70	18	1931-09-24	-7.07

Table 2.1: Dow Jones Industrial Average: largest daily percentage losses. Source: *The Wall Street Journal*.

Table 2.1 shows some of the largest daily percentage losses in the past 114 years. The most notorious of these is, of course, the Wall Street crash that started on Monday, October 28, 1929. The first downward movement on Monday was followed immediately by a major drop on October 29, known as Black Tuesday. There was then an extended period of extreme price fluctuations over months and years as the market attempted to find a new equilibrium, dropping to its lowest level on July 8, 1932. So severe was the correction that it took until 1954 for the price of the DJIA to return to the 1929 peak. See Figure 2.1.

The bubble of protracted unsustainable speculation in the years preceding 1929 is generally held to be the immediate cause of the collapse (Galbraith, 2009).The stock market had advanced steadily for seven years and the purchase of shares with borrowed money

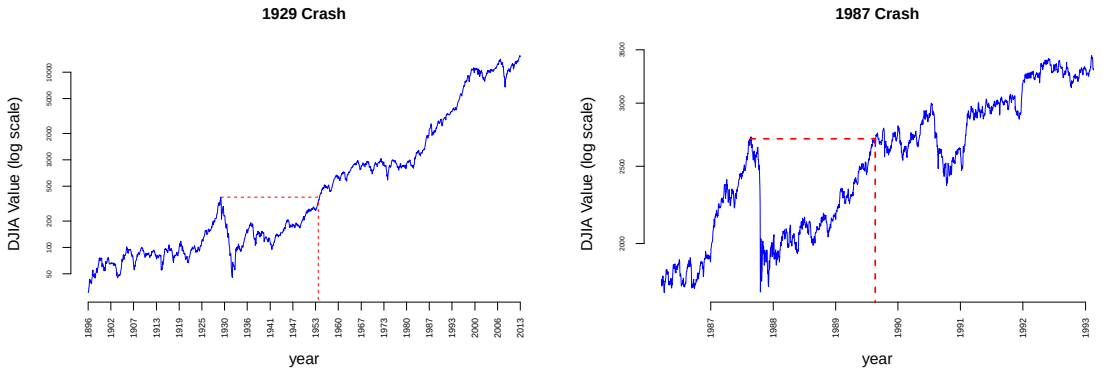


Figure 2.1: Dow Jones Industrial Average (logarithmic scale) before and after the 1929 and 1987 stock market crashes. Recovery from the 1929 crash took almost 25 years, whereas 1987 took 2 years. Data from `www.quandl.com`.

was seen to be a guaranteed way of making money. Speculation was fuelled by a general sense of improving living standards and excitement about media and new technology such as radio. Shares in companies such the Radio Corporation of America (RCA) had reached record prices and many private investors borrowed heavily to make further investments and take advantage of the apparently one-directional price movement. When the price started to slip there was panic selling among these investors as they realised they would not be able to cover their debts. Somewhat belatedly, regulators took action to limit trading on margin in the Securities Exchange Act of 1934.

The 1929 crash can be contrasted with the collapse of October 19, 1987, known as Black Monday. Again there was a bubble preceding the event – from early in 1987 until August 25, the Dow rose by 44%. Various explanations have been advanced for the timing of the burst of sell orders that triggered the fall. For a brief account see Fortune (1993), and more extensively the references therein. It is worth noting that one of the aggravating factors was the inability of the New York Stock Echange to process the large volume of trades that day and the consequential uncertainty among traders about the speed and direction in which the market was moving. Important lessons were learnt that eventually

led to improved infrastructure and circuit breaking controls in the exchange. Nevertheless, despite the large fluctuation over the succeeding months in 1987, the DJIA recovered its previous value in August 1989, and the market returned to a period of growth. See Figure 2.1.

At the turn of the century there were further collapses when the dot-com bubble burst on March 10, 2000 and when the markets opened on September 17, 2001 following the 9/11 attacks on the World Trade Centre. Both led to bursts of volatility and then eventual stability until the global financial crisis of 2007-2008 that threatened the total collapse of large financial institutions, resulting in the bailout of banks by national governments, and downturns in stock markets around the world. This too was followed by a large but temporary increase in volatility.

In his analysis of a number of such stock market corrections, Harris (2002) observes that rapid downward movements in market valuation produce subsequent periods of high volatility. By contrast rapid upward movements tend not to have that effect. This can be illustrated with more recent data by plotting the daily values of the S&P 500 index and the VIX index, a implied measure of volatility derived from the Black-Scholes-Merton formula for option pricing of the S&P 500. See Figure 2.2. For further background on the VIX, see Whaley (2000). The *implied volatility* is the volatility obtained by using either the instantaneous or integrated version of the option-pricing equation to match the theoretical price with the market price. It is often said to be the market's estimate of volatility. Implied volatility reflects how the market is pricing the option currently (Giot, 2005). Since the market does not have perfect knowledge about the future, actual and implied volatility can and will be different. In practice, implied volatility seldom matches historical volatility and this may have significant implications for hedging strategies. See for example Dumas et al. (2002) and Mykland and Zhang (2008).

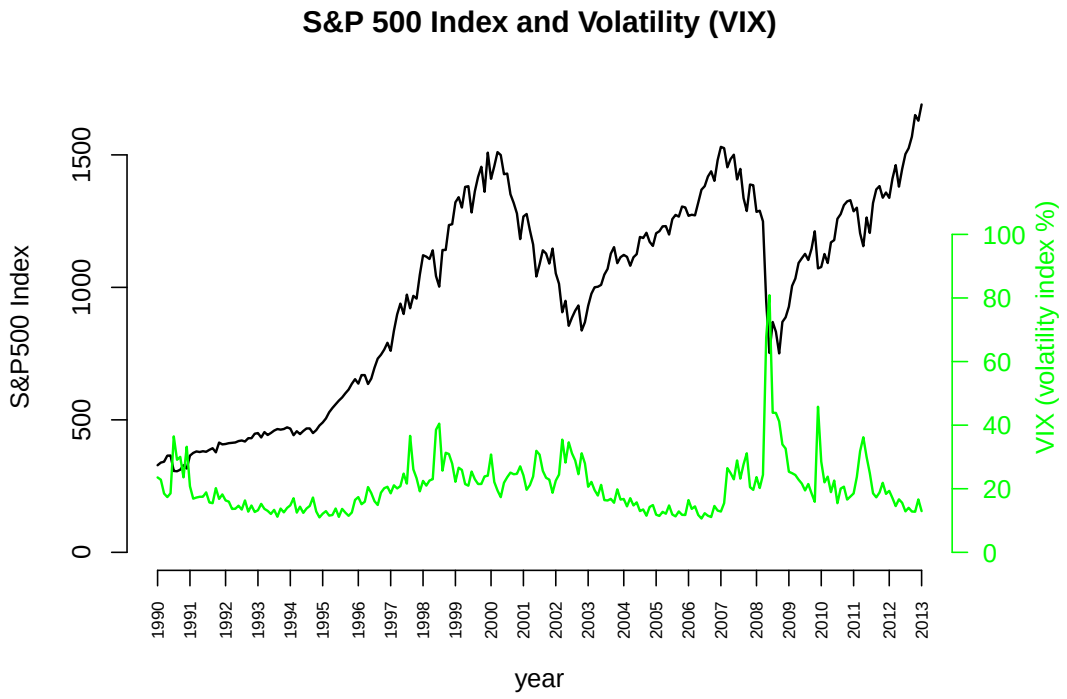


Figure 2.2: The volatility index (VIX) typically increases when there is a large fall in the S&P 500 index. The highest VIX value of 80.6% occurred on Nov 20, 2008. Data from `www.quandl.com`.

2.2 Consumers of volatility research

Volatility research is an extremely active area of research in econometrics. Building on the work of Engle (1982) on ARCH models, there is an extensive well developed econometric literature devoted to the problems of volatility measurement, modelling and forecasting. The Volatility Institute at New York University Stern School of Business founded in 2009 under Engle's direction is a focus of activity providing access to advanced techniques and online volatility analyses of several thousand economic series. Similarly the Oxford Man Institute provides online access to various non-parametric measures of historical volatility updated on a daily basis.

Econometricians are interested in establishing robust and informative summaries of price fluctuation. A wide range of such summaries have been proposed incorporating many of the well established methods of robust statistical estimation. Although the emphasis is primarily on historical analysis, the performance of these measures is often judged in a model framework, with specific models of stochastic volatility and covolatility. Descriptive measures are then scored and assessed with regard to their ability to estimate properties of the theoretical model. For a comprehensive review on forecasting volatility in financial markets, see Poon and Granger (2003).

The perspective of investment managers is somewhat different. Here the main focus is on prediction of the scale of future price fluctuation and the implication for risk assessment (Zimmermann et al., 2002). At a simple level, it may be of interest to predict the variance of the price of some asset at some future point in time. Similarly when a portfolio of assets is held, the correlation between assets become important since these will determine the optimal allocation of assets so as to minimise variability in the combined value of the set of assets. Since volatility and covolatility are known to be persistent in financial time series, model based predictions of volatility are clearly of interest and form an essential part of any investment strategy. These models are then tested and validated with historical data. Directional price movement on the other hand cannot persist for extended periods, since that would offer obvious *arbitrage* opportunities. This topic is developed in Chapters 4 and 5.

For both econometricians and investment managers, minor fluctuations on very short time scales are not of great importance and are treated as noise. However high-frequency traders focus on these and make their living by identifying and taking advantage of short lived opportunistic price fluctuations. They are sometimes referred to as *noise traders*. Their time scale is in fractions of seconds. Mid-frequency traders have perspectives of a

few minutes.

Although there is already a considerable body of research in this area, the need to predict volatility and quantify its effect on a wide range of time scales continues to be of importance. Modern approaches are outlined in Chapters 4 and 5.

2.3 The trading environment

To exploit opportunities presented by price movements in financial markets traders need access to liquidity. In financial terminology, *liquidity* describes the ability to trade *large quantities* of an asset *rapidly* and at *low cost*. Exchanges and other trading venues serve to provide liquidity by facilitating contact between buyers and sellers, either transparently by exposing buying and selling interest at specific prices and specific quantities in the case of traditional exchanges, the so-called "lit" venues, or in a guarded sense in the relatively unregulated world of dark pools where trades are executed on a private discretionary basis without publicly revealing the level of buy and sell interest.

The unlit venues emerged in the the early 2000s partly as a response to the concerns of traditional traders about the impact of computer driven algorithmic trading on the working of established exchanges (Scott-Quinn, 2012). Nevertheless, traditional exchanges continue to provide significant sources of liquidity and active participants continue to play traditional roles. Market makers post offers to buy and sell relatively small lots but adapt rapidly to demand. They are small scale liquidity providers, making money from the difference between their buy and sell prices and also from commission paid by equity exchanges on successful execution. Opportunistic fast traders look for small pricing anomalies and trade accordingly – they remove small scale liquidity. Block traders handle

large orders for specific clients. Their role is to split these blocks into smaller lots and execute the trades without significant market impact. Other traders analyse the fundamental value of the asset and trade when they believe it is mispriced. They provide *resiliency*. In principle their trades ultimately result in prices reverting to those that represent the fundamental value. For further background, see Madhavan (2000).

2.3.1 Latency

Speed has always been important in trading environments (Hasbrouck and Saar, 2013). Rapid information flow between national and international financial centres will determine trading decisions. Market data derived from different sources will have varying delays in arrival. Even the very best sources (perhaps co-located with the exchange) will have a delay (latency) measured in milliseconds (or possibly microseconds) between the event occurring and notification being received by the relevant system. Data transmission times become very important when the system is physically some distance from the data source. To minimise latency, high-frequency trading companies invest substantial sums in building fast communication networks between their installations. For example, if a trading event occurs on an exchange is in New York, latencies of around 30 milliseconds can be achieved for the data to be received in London via a transatlantic fibre-optic cable.

In general, it is desirable to know, to the nearest microsecond, the original time of a trading event. There are two reasons for this: to provide an accurate time series for the calculation of volatilities and correlations, and to determine the effective latency as it varies under specific market conditions. For this latter to be useful, the timestamps of the source must be related to a GPS time source, as must those of the receiving system (Korreng, 2010). Many modern exchanges provide this type of timestamp. However,

some do not, particularly for trade data, to prevent high-frequency algorithms taking too much advantage of the pattern of activity. Aggregated sources such as Reuters Selectfeed often do not provide this type of information.

Some Electronic Communication Networks (ECNs) such as EBS provide data on a sampled basis, rather than immediately as prices change. The ECN may send updated information at 100ms intervals for example, posting the price range of trades in the previous 100ms time slice and a snapshot of the order book. This may therefore hide details of multiple trades and other price movements in that period (Debelle, 2011).

In Chapter 6 we show how to use cross-asset measures of correlation to estimate transmission delays and quantify lead-lag response times. We show for example that the recent use by high-frequency traders of proprietary microwave links on the New York to Chicago route has reduced data transmission times to around 4.5 milliseconds, close to the light-speed determined minimum of 4 milliseconds.

2.4 Financial datasets: terminology

Detailed financial data on a second-by-second basis started to be available in the 1980s. The foreign exchange market was structured as an electronic/telephone market at that time. Currencies were priced against the US dollar. The prices that banks and other major players were prepared to buy and sell various currencies were posted publically and updated on the pages of news services such as Reuters and Telerate. These prices would fluctuate competitively throughout the day – the so-called *fighting-screen* – partly with a view to keeping the names of the players visible as often as possible. For a review of the evolution of the foreign exchange market see King et al. (2011).

At any particular moment in time, a market trader could telephone any of these players to negotiate a price to buy or sell a specified amount of the currency. But there was no commitment to the price previously advertised on the news service. In the 1980s, data on this succession of posted prices were made available for statistical analysis, although the actual trades were private. The data consisted of a price to buy and sell a particular currency with a timestamp identifying when the price were offered. The timestamps were irregularly spaced – updating very rapidly during periods of financial uncertainty. These data formed the basis for early statistical analysis by Zhou (1996) where it was assumed that the posted prices reflected a theoretically correct value, the efficient price, but with an added noise term recognising uncertainty. This model (coinciding with a similar model proposed by Roll (1984) in another context) became the standard assumption in the statistical analysis of market data, giving rise to an extensive literature. See Chapter 4.

In the intervening years, markets have become progressively more open. Many financial assets, including foreign exchange, are now traded on ECNs and exchanges, where firm prices and quantities are quoted continuously and trades are reported with prices and amounts.

2.4.1 Exchange based trading

The *limit orderbook* is the basic framework within which modern high-frequency data are monitored and processed (Harris, 2002; Dacorogna et al., 2001; Hasbrouck, 2007). Price data for exchange-traded instruments fall broadly into two categories: quotes and trades done. In an orderbook-driven market, offers to buy and sell (*quotes*) are made by market-makers, and taken by institutions wishing to trade.

In statistical terms, the limit orderbook is essentially the realisation of a tightly coupled multivariate marked point process. Observational data for a given financial asset are provided by an exchange on which it is traded. The orderbook shows how many units of the asset are available for sale by market participants at various prices per unit. These are called *ask* or *offer quotes*. At time t we denote the cheapest ask price by $P_t(a_1)$ and the total quantity available at that price by $Q_t(a_1)$. There will also be amounts quoted at other (worse) prices. We denote these successively worse prices by $P_t(a_k)$ with associated available quantities $Q_t(a_k)$ for $k = 2, 3, \ldots$.

On the other side, there are market participants who want to buy the asset. They are said to be *bidding to buy* at various *bid prices*. At time t the highest bid price is denoted by $P_t(b_1)$, possibly quoted by several participants, and the combined quantity looking to be bought at that price is denoted by $Q_t(b_1)$. Of course, at the same time there will be others wishing to buy at cheaper prices. We denote these successively lower prices by $P_t(b_k)$ and the associated quantities by $Q_t(b_k)$, $k = 2, 3, \ldots$. For a particular exchange, only m values of the bid/ask quotes are published and viewable by subscribers to the data feed. The number of rungs in this ladder of quotes is typically $m = 3, 5$ and 10. All the while quotes remain on the orderbook we refer to them as *resting orders*.

At time t the published prices are always ordered by

$$P_t(b_m) < \cdots < P_t(b_1) < P_t(a_1) < \cdots < P_t(a_m). \tag{2.1}$$

The process $P_t(a_1) - P_t(b_1), t > 0$ is called the *spread*. Prices are discrete valued and for heavily traded assets usually recorded to at least 4 places of decimals. The minimum price movement is called the *tick size* or *pip*. The term 'tick' is derived from the tick of a ticker tape and the term *tick-by-tick data* denotes the sequence of data delivered by such a

device, or its modern electronic equivalent. For heavily traded assets, the spread is almost always equal to the tick size. The data associated with an orderbook are said to be high-frequency if there is a large number of changes in a brief period of time. For example, there are usually more than 3 million changes per day in the 5-level orderbook for futures trading of the S&P index. These include changes in prices, changes in amounts available at specific prices and records of completed trades. The interval between changes may be as small as a few microseconds.

During the day the orderbook changes from time to time as new orders arrive or existing ones are cancelled. The point processes associated with these changes are $\{N_t(a_k), t > 0\}$ on the ask side and $\{N_t(b_k), t > 0\}$ on the bid side, $k = 1, \ldots, m$, with associated marks corresponding to changes in the available quantities.

The orderbook is also affected by trading, in one of two ways. The first is when a matching quote arrives. For example a bid quote may arrive at a price that is matched by an ask price. When the amount in the arriving quote is matched completely, the ask price is consequently removed. Any unmatched units in the quote than become a bid quote for the residue. If on the other hand, only part of the order is matched (because of insufficient available quantity) the ask price is unchanged but the amount at that ask price is adjusted accordingly.

Alternatively, trades can be submitted to the exchange on an *immediate or cancel* (ioc) basis, meaning that if for example the trader sends an order to buy n units of the published amount offered at the best ask price, then the deal is done for the minimum of n units and the amount still available at the time the order arrives at the exchange. The ioc order is cancelled by the exchange if there are no units available at the target price (either because another trader has got there first or because the offered price has been withdrawn).

In both cases we can distinguish between trades that were completed at the existing ask price and those that were completed at the bid price. We denote the sequence of trade times by $\mathcal{T}^*(a)$ and $\mathcal{T}^*(b)$, respectively, with marks given by the associated price and size of the trade. Trade data (timestamp and price) are generally published by the exchange. The size of individual trades may not be given, although it can be inferred to some degree of accuracy from changes in the order book. Figure 2.3 is a small snapshot of the first

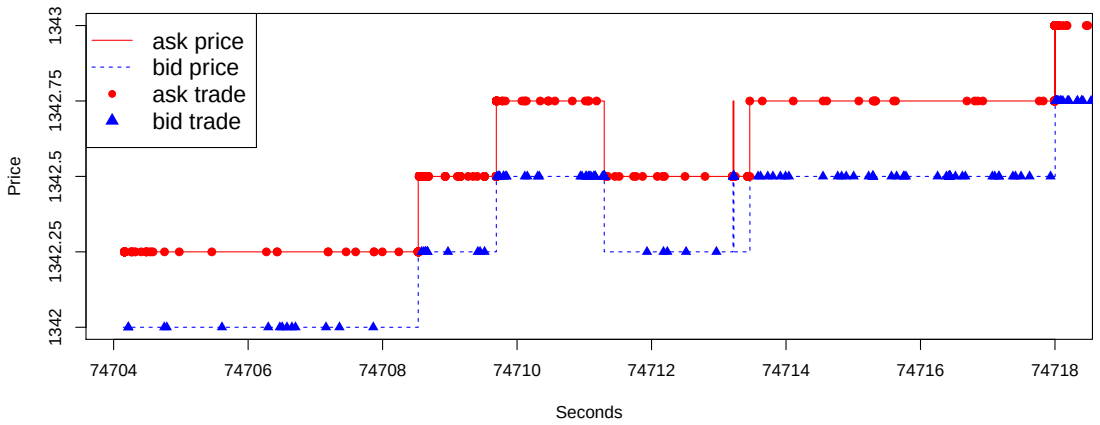


Figure 2.3: Prices and trades at the first level of a limit order book .

level of the S&P 500 index futures limit orderbook. The picture is typical, showing large numbers of trades in between relatively slowly changing prices.

This pattern is a comparatively recent development in orderbook trading. Twenty years ago trades were less frequent and individually much larger. The dominance of algorithmic trading in recent years has led to computer controlled strategies for the rapid submission of many small orders, in part with the aim of disguising the intention to buy and sell larger amounts of the asset (Hasbrouck and Saar, 2009).

For some purposes, the simple sequence of trade times $\mathcal{T}^*$ and associated prices P, without distinguishing between bid and ask, is taken to be a basic summary of market activity.

For some assets, $(\mathcal{T}^*, P)$ may be the only data available.

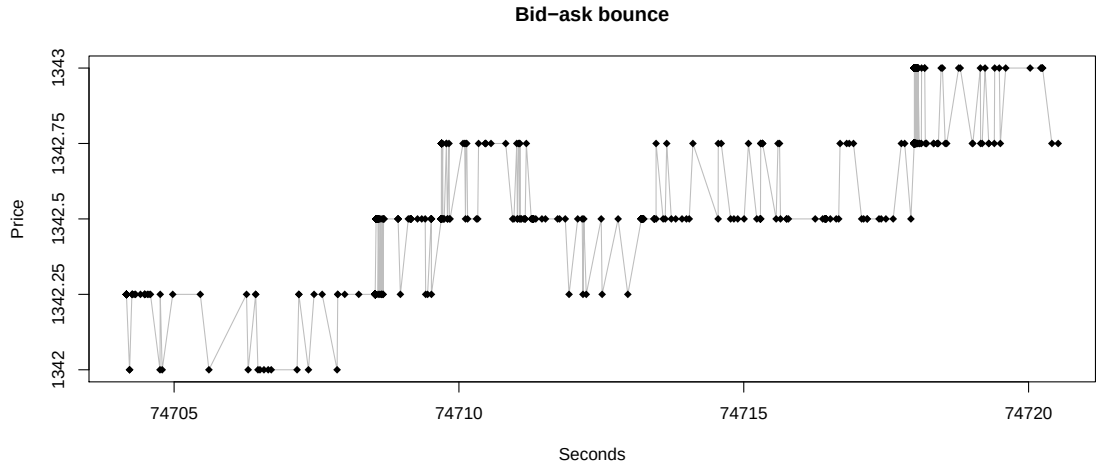


Figure 2.4: Trade price from a limit order book over a 15 second interval – illustrating bid/ask bounce .

Note that trade prices on either side of the book (bid and ask) will necessarily be separated by the spread. This explains the small rapid changes in trade prices seen in the simple sequence S, illustrated in Figure 2.4 – a phenomenon known as *bid/ask bounce*. It is one aspect of what is known as *microstructure noise*.

On a longer time scale, these changes are negligible and trade prices can move widely as in Figure 2.5.

2.4.2 Choice of observables

When working with orderbook data, there are a number of options in deciding which variables to use and when to observe them (Engle and Russell, 2002). Nowadays exchanges provide detailed price information in a more or less continuous stream. The first option is not to subsample at all, i.e. use all the data – but this may present substantial computational costs when considering a time sequence that ticks whenever there is any change in

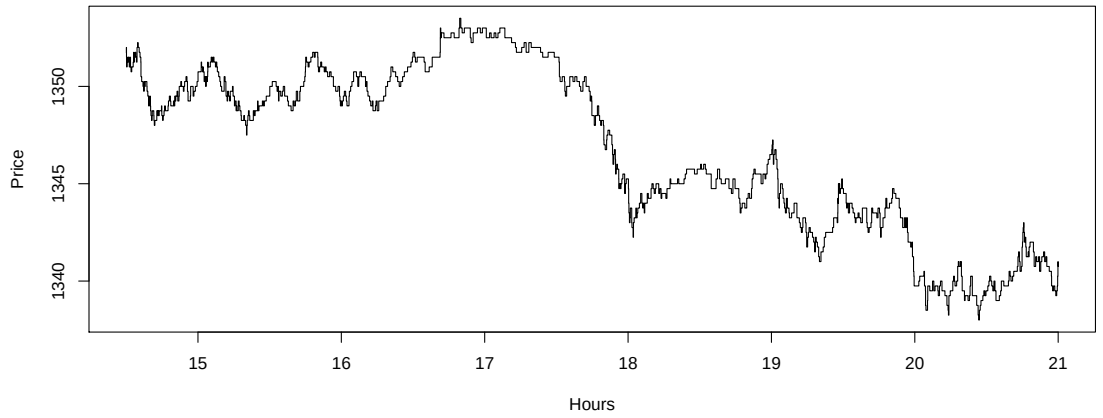


Figure 2.5: Trade price data over the course of the trading day (CME S&P 500 index futures) .

the orderbook at all. Another option is to time slice at regular intervals: per second, per minute, etc, but this risks missing important localised activity within particular time slices. Other options are slicing at trade times or changes of the best bid and ask prices $P_t(b_1)$ and $P_t(a_1)$ and there are further options of working with a time scale on which volatility is approximately constant. We return to this topic at several points in later Chapters. For further background see Pigorsch et al. (2012).

The mid-price $m(t)$ is a useful summary of an order book. In the simplest form it is the unweighted average of the best bid and ask prices at each time t, i.e.

$$m(t) = \frac{P_t(a_1) + P_t(b_1)}{2}.$$

A weighted version w_t is often used too, taking account of the amounts available at the best price, $Q_t(a_1)$ and $Q_t(b_1)$, inversely weighting so that

$$w_t = \frac{P_t(a_1)Q_t(b_1) + P_t(b_1)Q_t(a_1)}{Q_t(a_1) + Q_t(a_1)}.$$

More elaborate measures of the central price on the orderbook take account of further rungs in the orderbook ladder.

Figure 2.6 shows some tick-by-tick data in greater detail. These are data that "tick" whenever a trade is reported or the quoted amount changes at any of the 5 levels on either side of the order book. For S&P 500 index futures there may be around 3 million ticks per day. Each tick is time stamped by the exchange to microsecond accuracy.

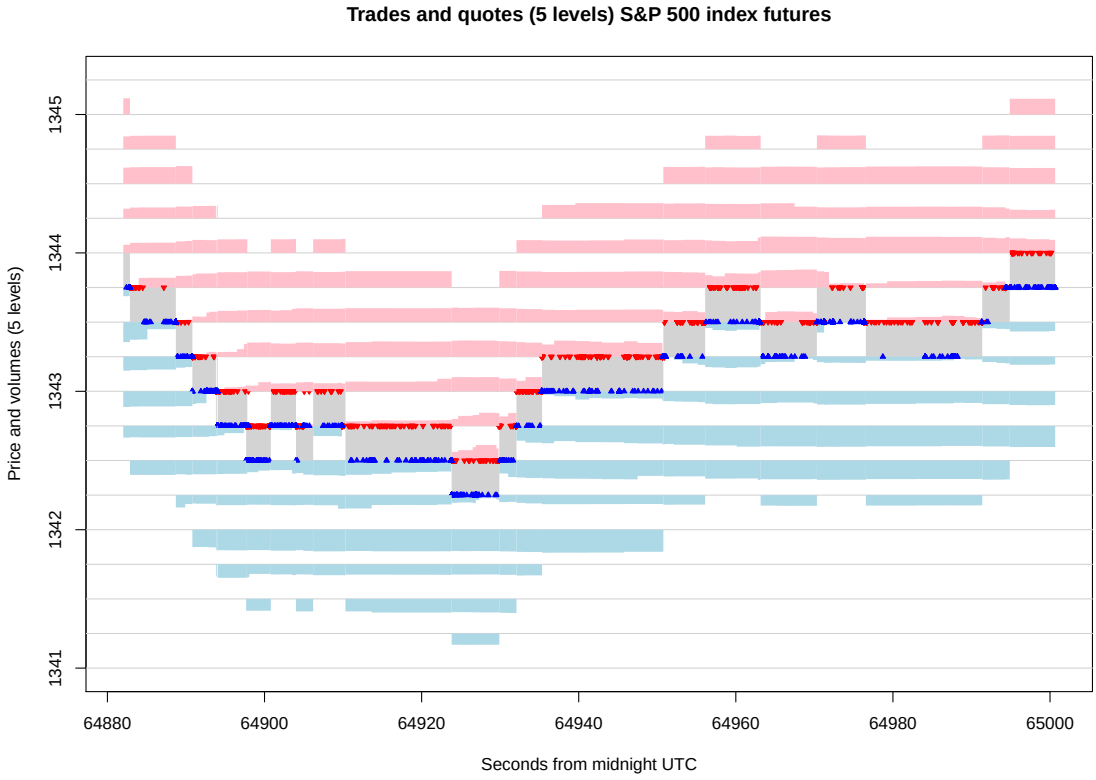


Figure 2.6: Example of order book and trades data, tick-by-tick, for S&P 500 index futures: 120 second interval, around 25000 ticks, 5 levels. The volumes quoted at each price are illustrated by coloured shading. The spread region is shaded grey. Trades are marked by triangular symbols. Trade sizes range between 1 and 5000 contracts.

In principle, volatility estimates can be obtained for any such sequence of exchange data. In one sense trade data is a more solid reflection of market activity since it reflects active

decisions, but high-frequency trading requires knowledge of the *immediate future* volatility and for this purpose the quoted prices and the general structure of the order book may be more informative.

2.5 Sources of high-frequency data

High-frequency computerised trading systems are geared towards markets with substantial trading activity. Traders tends to gravitate towards markets that are highly liquid. Liquid markets are characterised by an ability to absorb trading activity without substantial price movement. In other words, there are substantial volumes available to buy and sell and these amounts are rapidly renewed by a broadly based cohort of market participants whenever trades take place. High-frequency market-making firms play an important role in the provision of liquidity, thereby ensuring that prices are stabilised with the minimum difference between buying and selling prices.

Equity dealing on the major electronic exchanges around the world is an important focus as is foreign exchange trading. Alongside these cash markets, a wide variety of futures contracts are available and traded heavily on futures exchanges such as the Chicago Mercantile Exchange, Eurex in Frankfurt and Atlanta-based Intercontinental Exchange (ICE).

2.5.1 Futures

Futures are contracts to buy or deliver a certain quantity of a specific asset at a fixed point of time in the future, the expiry time of the contract (Hull, 2009). The buyer of such a contract is said to be *long* and the party who promises to deliver is said to be *short*. In their original form, futures were contracts to deliver physical quantities such as

foodstuffs at a specified price. Gradually in recent years, exchanges have extended the idea to less tangible assets where delivery is either impossible or rare. For these cases, delivery is eventually settled by cash equivalence with daily margin calls on the holders of the contract to reflect the inherent risk and value of the contractual obligation. For further details see Miller (1986).

2.5.2 Eurodollar and other interest rate futures

It may be helpful to illustrate the mechanics of Eurodollar futures trading. The 3-month London Interbank Offered Rate (LIBOR) is a pooled estimate of the cost of borrowing US dollars in the financial markets outside the United States for a 3-month period. Eurodollar futures contracts depend on the future value of LIBOR at specified expiry dates (CME, 2013). The contracts are traded heavily on the Chicago Mercantile Exchange (CME). Clearly there is no tangible asset here, although the current 3-month LIBOR is published daily and will eventually be known at the specified expiry date of the contract. Meanwhile an assessment of the rate can be made.

By convention the contract trades with a 'price' that is 100 minus the annualised interest rate. So if there is a consensus that the annualised 3-month LIBOR will be 1.01% at the expiry date of the contract, the notional 'price' will be approximately 100 minus the rate, i.e. 98.99. Some traders will assess the price higher and some lower.

The Eurodollar futures contract is based on lending a notional \$1 million. A trader who is short the contract agrees to deliver a synthetic product created at the expiry date with a value of $\$1 \text{ million} \times (100 - x)$, where x is the annualised 3-month LIBOR on that date. Similarly, a trader who is long agrees to buy the product at this value.

As the expiry approaches, the exchange transfers money between the traders with short and long positions, adjusting for their respective anticipated financial commitments at expiry. A movement of one basis point, one hundredth of a percentage point, equates to a value of $25, by the calculation

$$\$25 = (\$1 \text{ million}) \times (0.01 \text{ percent per year}) \times (90 \text{ days}/360 \text{ days per year}).$$

Money is transferred between those with short and long positions on this basis. In fact, the number of days in a 3 month period is not exactly 90, and the notional $1 million is re-interpreted by the exchange and varied from quarter to quarter, to preserve the simple value of $25 per basis point movement.

For Eurodollar, there are individual contracts for each of the quarterly expiry months: March, June, September and December for 10 years into the future along with 4 monthly expiry dates for the nearest months not in the quarterly cycle. In addition there are spread contracts where the trader takes a long position in the contract for one expiry date and a short position in a subsequent expiry, a *calendar spread* and more complicated combinations involving contracts with 3 or more expiry dates. The contracts are highly correlated with strong linear relationships (Jegadeesh and Pennacchi, 1996).

2.5.3 Government bond futures

Government bond futures operate in a similar manner to interest rate futures such as Eurodollar. The most frequently traded are US Treasury and German bond futures. Contracts are defined by the exchange which is used to trade them, and specify various parameters of the trade being done, including the quantity of the underlying instrument (e.g. $100,000

bond) and the delivery date, which is usually set to be one of 4 specific times each year (March, June, October, December). Again futures trading does not necessarily result in delivery of the underlying instrument - a contract may be bought and subsequently sold to result in no outstanding position (liability to deliver or receive). Since futures contracts are specific to a single exchange, they are identified by a code defined by the exchange. For example, 10-year US treasury futures contracts on the CBOT Globex electronic exchange are identified by the code ZN. The specific maturity date is then appended as a further two character code - for example, ZNZ9 refers to the contract for delivery in December 2009. For further details see Burghardt et al. (1994).

2.5.4 Other futures contracts

Equity Index futures contracts are based on the projected values of equity indices such as the S&P 500 index of share prices on the New York Stock Exchange, the FTSE index in London and the equivalent indices on the Frankfurt exchange. Foreign Exchange (FX) futures contract follows the exchange rates between major currencies. The CME exchange is the most active market for FX futures.

2.5.5 Cash FX

Cash FX data relate to the current spot rate between two currencies. So for example, GBPUSD spot refers to the number of US dollars that one pound sterling will buy for spot delivery (2 days from today). In our context the trading ECNs include Reuters and Electronic Broking Services (EBS) that is an electronic inter-dealer broker and provider of post-trade services.

time	reg_b1	reg_a1	b1	qb1	a1	qa1	given	paid
53731.41	90.12	90.21	90.17	4	90.18	5	0	0
53731.513	90.12	90.21	90.17	4	90.18	6	0	0
53731.61	90.12	90.21	90.17	3	90.18	6	90.17	0
53731.711	90.09	90.2	90.15	15	90.16	10	90.16	0
53731.811	90.1	90.2	90.15	11	90.16	4	0	90.16
53731.91	90.09	90.2	90.15	5	90.16	3	0	90.16
53732.017	90.09	90.2	90.15	5	90.17	9	0	0
53732.109	90.09	90.2	90.15	5	90.16	1	0	0
53732.208	90.09	90.2	90.15	4	90.16	1	0	0
53732.309	90.09	90.2	90.15	3	90.16	1	0	0
53732.415	90.08	90.2	90.15	2	90.16	1	0	0
53732.708	90.08	90.2	90.15	2	90.16	1	0	0
53733.306	90.08	90.2	90.15	2	90.16	1	0	0

time	- time, in seconds of day
reg_b1	- regular bid (i.e. price you can sell a regular amount)
reg_a1	- regular offer (i.e. price you can buy a regular amount)
qb1	- volume at bid on the top of the book (i.e. volume at b1)
qa1	- volume at offer on the top of the book (i.e. volume at a1)
b1	- bid at the top of the mkt orderbook
a1	- ask at the top of the mkt orderbook
given	- last price given (sold)
paid	- last price paid (bought)

Table 2.2: Example of foreign exchange data from EBS

The FX orderbook as with other limit orderbooks contains a 'ladder' of prices, with the bid-offer price spreads varying according to the quantity on offer. For example, 1 million USDJPY might be offered at 90.48/90.50 whereas 20 million could be offered at 90.47/90.51. The top 'rung' of the orderbook - i.e. that with the narrowest and therefore best spreads - is quoted as the best bid/ask data and generally available via electronic quoting services such as Reuters IDN Selectfeed. In the case of FX this may be a blended, or 'best' rate spanning multiple exchanges.

An example of high-frequency FX data from EBS is given in Table 2.2.

Note that, these EBS records are time-sliced at approximately 100 millisecond intervals. The quote data, price and volume, are correct at each time interval but the trade data only provide the best price of a trade in the time interval and do not show how many trades occurred or the corresponding trade volumes (Chaboud et al., 2014). On the EBS system

in operation when these data were collected, FX prices could only change by multiples of 1% of a basic unit which is a 1/100 dollar for the US currency. The terms 'regular bid' and 'regular offer' above refer to the price the trader would have to pay to sell or buy a fixed amount of the currency, e.g. 50 million US dollars when exchanging Japanese Yen. For EBS, traders are additionally able to view the volume of currency available at certain prices above and below the bid and ask prices. Underlying all of this is a complicated automated process managed by the ECN, dealing with high-frequency submissions and deletions of quotes – but that is hidden from the trader.

The difference between the data formats of FX and bond futures presents a difficulty in estimating cross-correlations before between these two assets. This is discussed further in Chapter 5.

Chapter 3

Exploiting mean reversion

> "*I can calculate the motions of the heavenly bodies but not the movements of the stock market*". – Sir Isaac Newton in 1768, after being wiped out in one of the many stock market crashes of his era.

Although financial markets are subject to discontinuities as described in Chapter 2, in some cases close interaction between financial instruments can serve to stabilise price movement over a period of time. In foreign exchange markets, for example, rates are expected to move in a self-consistent manner. If the Dollar-Yen and Euro-Dollar exchange rates change then the Euro-Yen rate should change accordingly. Delayed information flow and price inertia open up arbitrage opportunities. When the Dollar-Yen and Euro-Dollar markets are highly active the Euro-Yen rate may get out of line, but the expectation is that in due course it will revert to a value consistent with the other rates. This is a so-called *triangular arbitrage*. By buying and selling small lots throughout the day, high-frequency traders are able to exploit the mean-reverting property of exchange rates.

For traders who trade larger lots with a longer time horizon, mean reversion cannot be

exploited in this way. The profit would have been 'locked in' after the 3rd currency pair is traded, but the trades will only be settled on two days later. So by the time the trade is settled, it doesn't matter whether the price has reverted or not, the profit has already been made. In my experience, a trader would do the reverse trades afterwards so that they don't need to commit the capital which they used for leveraging the trade before the settlement, due to credit risk, capital deposit or regulatory requirements such as BASEL II.

More generally and in a less mechanical manner when the prices of two positively correlated assets diverge temporarily, we will expect one of the prices to revert to a price consistent with the other, over some period of time. A classical strategy is *pairs trading* where a portfolio consisting of a weighted combination of two assets is traded. Profit is determined by the choice of weights in the linear combination as well as the market timing to enter and close a position. The choice of weights involves the measurement and interpretation of correlation and volatility. Adaptive estimates of these quantities can be used with the dynamic linear model of Section 3.4 to predict the return for an asset that has not recently traded. The predicted value is then compared with the quoted price in the order book to take advantage of any temporary anomaly and to trade profitably relying on subsequent mean reversion to remove the anomaly and the cost of the trade. See Chapters 4 and 5, for further discussion of volatility and correlation estimation in financial data.

Hunting for mean reversion lies at the heart of successful high frequency trading. Proprietary noise traders exploit pricing inefficiencies across securities and other asset classes, both locally and globally. In local trading on a specific exchange, transient price anomalies can be expected to disappear rapidly as the force of market opinion returns the price to its former level. If the price is low, for example, this is an opportunity to buy with the prospect of selling later as the price reverts. By trading rapidly, small gains can be accrued on such events.

At a global level, high-frequency traders exploit low-latency computer networks to access geographically separated markets. They determine the price they are prepared to quote to other market participants bearing in mind lead-lag delays as pricing information flows around the world. Stale prices on exchanges that are remote from the principal sources of price formation can be expected to revert to a more generally accepted norm over a short period of time. This is studied in Chapter 6.

The high-frequency trading arms race is a consequence of continuous-time trading on electronic financial exchanges. That is, under the current continuous limit order book market system, it is possible to buy, sell, and post orders for exchange-traded financial instruments extremely rapidly at any instant during the trading day. The appearance of multiple electronic trading venues has led to an increase in frequency and volume of transactions and order flow. The objective of all traders, whether an algorithmic or a human trader, remains the same: to buy low and sell high.

A typical high-frequency trading algorithm runs on a sub-millisecond time scale and has very short portfolio holding periods, usually well under five seconds and frequently less than one second. Large amounts of data are processed every millisecond and there is a requirement is to act rapidly to these data as the profitability of the signals in the market decays very quickly. Substantial investments have been made by the high-frequency trading community utilising custom built fibre-optic, microwave and millimetre-wave networks and dedicated hardware based on field programmable gate arrays (Leber et al., 2011). Even with these investments, the need for speed prioritises statistics that can be computed in linear time and updated recursively.

In this chapter we focus on recursive methods for rapid trading and the exploitation of mean reversion. The applied forecasting world has a long history of using recursive

methods to predict future outcomes. Basic kernel-based methods are described in Section 3.3 and the extension to recursive linear regression in Section 3.3.5. The classic filtering methods of the dynamic linear model (DLM) and Kalman filter are motivated in a financial context in Section 3.4. The close connection between the Kalman filter and the exponentially weighted moving average forecast is illustrated in Section 3.4.4 for completeness.

News announcements can produce major disruptions to the market, as described in Section 3.5. A fundamental question is whether prices will revert to their earlier values, either in real or absolute terms, and whether historical correlations under benign conditions can be relied upon to persist in a new and unexplored trading environment. Modelling of these special events needs to take account of the influence of external factors as well as the sensitivity of dynamic models to initial conditions. This is briefly discussed in Section 3.5 – although it is not the main focus of the high-frequency trading community.

3.1 Forecasting, filtering and time change

A basic concept in finance is that of the underlying fair value of an asset. In practical terms, when looking for profitable opportunities, traders compare the currently available price with their estimate of the fair value at some future time point. When the current price is lower than their estimate they buy and when it is higher they sell. The underlying belief is that the price will revert to their estimated fair value.

In estimating future values based on the past, practitioners can call on a range of forecasting methods drawn from a wide variety of disciplines. Broadly speaking the methods are of two types: those based on mathematical modelling often with an underlying Bayesian

philosophy and those based on practical experience in Operations Research and other applied fields. We will refer to them as *model-based* and *traditional* forecasting methods, respectively. Shephard (2015) uses the term 'non-model' to describe traditional methods.

3.1.1 Traditional forecasting

Traditional forecasting has a long history going back to the work of Brown and Holt in the 1950s. Holt's 1957 report to the Office of Naval Research is reprinted in Holt (2004) and Brown's contributions are recounted in his historical review (Brown, 2004). The exponentially weighted moving average (EWMA) emerged at that time as an expedient way of computing a moving average providing 'savings in data storage over moving averages' owing to the recursive nature of its calculation (Gardner, 2006). The EWMA can be used as a forecast of future values when no change is expected.

In general, the quality of a traditional forecast is judged by empirical performance and any parameters involved optimised accordingly. Similarly, extensions aimed at forecasting trends can be assessed by their performance in practice. In other words, model assumptions are not emphasised and judgements about the adequacy of the method are entirely empirically based. One of the benefits of this practical approach is that considerable flexibility can be introduced into forecasting formulas. For example, Trigg and Leach (1995) developed a popular adaptive version of the EWMA. For other practical modifications, extensions and variants of the EWMA, see Brown (2004) and Gardner (2006). The disadvantage of traditional methods is that it is difficult to find guidance on which method to use, so each new application has to be approached on a trial and error basis.

3.1.2 Model-based forecasting

By contrast, in this formulation a model is assumed for an underlying unobserved process and a specific model is introduced for the observation process. The objective is to recover the underlying process after filtering out the noise. Basic filtering theory with emphasis on stochastic differential models is covered in Davis (1997) and Øksendal (2003), for example. In finance the simplest model of price movement on a logarithmic scale has Brownian motion as the underlying process with independent observational noise. We will assume that price is always measured on a logarithmic scale, use 'log-price' to denote the logarithm of the price from now on. In practice the model is a poor fit to observations particularly with high-frequency data. To remedy this the model has been extended to allow stochastic evolution of the infinitesimal variance in the underlying diffusion, i.e. the introduction of *stochastic volatility*. Shephard (2015) provides historical background with further extensions, and modern developments. For a comprehensive review of filtering and forecasting in finance with particular focus on statistical techniques that have been used to accommodate stochastic volatility, see the compendious Bauwens et al. (2012).

3.1.3 Time change

Before discussing forecasting methods, it is important to consider the nature of time in a high-frequency trading environment. We cannot emphasis more strongly the importance of this topic in determining the fundamental time-scale for the analysis of high-frequency trading data. It is well known that price fluctuations occur in bursts, as a result of highly competitive trading interaction. These changes in price volatility can be explained by stochastic volatility models. In the simplest case, log-price is modelled as Brownian motion with stochastic infinitesimal variance. Since the volatility is not constant, price

changes over a fixed period of time are no longer normally distributed and may be heavy-tailed, as observed empirically.

In an alternative formulation, as has been pointed out by Clark (1973), the price process can be thought of as evolving on a non-uniform time scale. With a suitable time change, the volatility is constant and standard Brownian motion can be reconstructed. In other words, if the underlying price $X(t)$ is modelled as

$$dX(t) = \sigma(t)dB(t),$$

where $B(t)$ is standard Brownian motion, then

$$U(t) = X(T(t)), \quad t \geqslant 0 \tag{3.1}$$

is standard Brownian motion when $T(t) = \int_0^t 1/\sigma^2(u)du$. For a discussion of the necessary conditions for this to hold, see for example Veraart and Winkel (2010).

When only the price of the asset is available and the underlying volatility is unknown, the representation via time-change is of limited practical use. However Ané and Geman (2000) have shown that approximations to $T(t)$ can be obtained by using other information provided in modern high-frequency market data. In particular they show that an *activity time-scale* can be constructed as an affine function of the cumulative transaction count and that with this time change, log-price increments are no longer heavy tailed and are well approximated by Gaussian distributions. See also Easley and O'hara (1992). To a certain level of approximation they argue importantly that on the activity time-scale, volatility can be considered to be constant. In the same paper they consider an earlier proposal by Clark (1973) to use the cumulative *volume* of trade as a time scale but show

that this has inferior properties.

High-frequency traders can choose to construct their forecasts on physical time, trading activity time or more elaborate time scales involving cumulative activity as measured by cancellations, modifications and the arrival of new orders. The choice will be determined by the empirical performance of the forecasting or filtering methods they adopt.

3.2 Marked point datasets

Electronic exchanges produce streams of data in packets tagged by timestamps and values. The timestamps and values are both discrete. If we consider the arrival of such a packet as an event, the data have the structure of observations from a marked point process. There is an extensive literature on the modelling of marked point process with early applications to the statistical analysis of earthquake data where events correspond to earthquakes at specific times and the marks describe characteristics of the seismic activity. See Cox and Isham (1980); Ogata (1988); Daley and Vere-Jones (2007) and the references therein.

A marked point dataset is a sequence of events at distinct times $t_1, t_2, \ldots, t_n$ with associated values $Z_1, Z_2, \ldots, Z_n$ at these times. The data can be represented by the step function

$$S(t) = \sum_{i:t_i \leqslant t} Z_i, \quad t \in R, \tag{3.2}$$

the cumulative sum of the values at or before time t.

The simplest example is the counting function $N(t), t \in R$, which counts the number of events that have occurred at or before t.

In financial application, we may be interested in comparing two counting functions, for

example the number of *given* trades and the number of *paid* trades over a period $(0, t)$. If $N_p(t)$ is the number paid and $N_g(t)$ is the number given then $N_p(t) - N_g(t)$ is a marked point dataset where $t_1, t_2, \ldots, t_n$ are the trade times and $Z_i = 1$ at time t_i if the trade is paid and $Z_i = -1$ if it is given, $i = 1, 2, \ldots, n$.

Similarly if $V_p(t)$ is the cumulative volume of paid trades over the period $(0, t)$ and $V_g(t)$ is the cumulative volume of given trades, then the net volume of trade, $V_p(t) - V_g(t)$, is a marked point dataset with Z_i equal to V_i or $-V_i$ depending on whether the traded volume, V_i, is paid or given.

Finally, the bid, ask or mid log-price of an asset can be thought of as marked point dataset where $t_1, t_2, \ldots$ are the times at which prices have changed and $Z_1, Z_2, \ldots$ are the log-price changes, positive or negative.

Trades are the simplest events in data from an electronic exchange. When a trade occurs, the exchange publishes the time of the trade along with the price and quantity trade. The price and quantity as a pair are the mark associated with the trade. Events involving new quotations to buy or sell, modifications of existing orders and cancellations produce more complicated marks. These form the fundamental 'ticks' of market data. In some cases the events will result in changes to the best prices offered to buy and sell. When a quoted price changes it persists until altered by subsequent events. An order associated with the price is said to be resting and is available for trade until removed. Of course, alterations to prices deeper in the order book do not affect the best prices, but nevertheless this deeper information is published by the exchange and may indicate changes in market sentiment and the direction in which the market will move.

3.3 Recursively calculated forecasts

A first step in the exploitation of mean reversion is to establish a starting point. In practice this involves smoothing recent values using an appropriate moving average. As a specific example we suppose the data are observations, $X(t) = x(t)$ of the exchange quoted log-price – taking 'price' to be mid-price, the average of the bid and ask prices. Let $\{t_k\}$ be the times at which prices are updated and write $x_k = x(t_k)$ for notational convenience.

We consider a class of kernel-based smoothers that include the exponentially weighted moving average as a special case. Choosing the kernel is one of the main difficulties in statistical analysis. In practice, high frequency traders try out different kernels. See Brown (2004) for an introduction to the extensive traditional forecasting literature.

3.3.1 Kernel based moving averages

Throughout this Chapter, we will use the word 'kernel' to refers to any integrable function K on $[0, \infty)$ such that

$$\int_0^\infty K(u)du = 1. \tag{3.3}$$

Note that K is not required to be non-negative.

The kernel can be used to produce a moving average of the quoted log-price $x(t), t \in R$, as follows:

$$\hat{\nu}(t) = \int_{-\infty}^{t} K(t-u)x(u)du.$$

For the exponential kernel, $K(x) = \tau^{-1}e^{-x/\tau}, x \geqslant 0$, we define the *exponentially weighted*

moving average (EWMA) of $x(t), t \in R$ to be

$$\hat{\nu}_1(t) = \int_{-\infty}^{t} \tau^{-1} e^{-(t-u)/\tau} x(u) du.$$

Note in the traditional forecasting there are other variants of the EWMA (Gardner, 2006). Since $x(t)$ is a step function with value x_k on the interval $[t_k, t_{k+1})$ we have

$$\begin{aligned}\hat{\nu}_1(t) &= x_k \int_{t_k}^{t} \tau^{-1} e^{-(t-u)/\tau} du + \sum_{i \leqslant k} x_{i-1} \int_{t_{i-1}}^{t_i} \tau^{-1} e^{-(t-u)/\tau} du \\ &= x_k \left(1 - e^{-(t-t_k)/\tau}\right) + \sum_{i \leqslant k} x_{i-1} \left(e^{-(t-t_i)/\tau} - e^{-(t-t_{i-1})/\tau}\right), \end{aligned} \tag{3.4}$$

where $t_k < t \leqslant t_{k+1}$.

At event times we get the EWMA recurrence,

$$\begin{aligned}\hat{\nu}_1(t_{k+1}) &= \int_{-\infty}^{t_{k+1}} \tau^{-1} e^{-(t_{k+1}-u)/\tau} x(u) du \\ &= e^{-\Delta_k/\tau} \int_{-\infty}^{t_k} \tau^{-1} e^{-(t_k-u)/\tau} x(u) du + x_k \int_{t_k}^{t_{k+1}} \tau^{-1} e^{-(t_{k+1}-u)/\tau} du \\ &= e^{-\Delta_k/\tau} \hat{\nu}_1(t_k) + (1 - e^{-\Delta_k/\tau}) x_k, \end{aligned} \tag{3.5}$$

where $\Delta_k = t_{k+1} - t_k$, and we have used the fact that $x(u)$ is a step function with constant value x_k for $t_k \leqslant u < t_{k+1}$.

For an intermediate time t between t_k and t_{k+1} we have

$$\hat{\nu}_1(t) = e^{-(t-t_k)/\tau} \hat{\nu}_1(t_k) + (1 - e^{-(t-t_k)/\tau}) x_k. \tag{3.6}$$

Note that when $\{t_k\}$ are equally spaced, i.e. $\Delta_k = \Delta$, from expression (3.4) we have the

simplest form of the EWMA

$$\begin{aligned}\hat{\nu}_1(t_{k+1}) &= (1-e^{-\Delta/\tau})\sum_{i\leqslant k} x(t_i)e^{-\Delta(k-i)/\tau} \\ &= (1-\lambda)\sum_{i\leqslant k} x_i\lambda^{(k-i)}, \quad \text{where } \lambda = e^{-\Delta/\tau}. \end{aligned} \tag{3.7}$$

More general kernels are provide by the gamma family. Gamma kernels are non-negative with flexible shape depending on a parameter α. The gamma kernel is

$$K(x) = x^{\alpha-1}e^{-x/\tau}\tau^{-\alpha}/\Gamma(\alpha), \quad \tau > 0, \quad \alpha > 0,$$

giving the moving average

$$\hat{\nu}_\alpha(t) = \int_{-\infty}^{t} (t-u)^{\alpha-1}e^{-(t-u)/\tau}\tau^{-\alpha}/\Gamma(\alpha)x(u)du.$$

The case $\alpha = 1$ this is the exponential kernel. When $\alpha > 1$ the moving average puts less emphasis on values immediately prior to t.

Moving averages based on the gamma kernel can be updated recursively. For example, when $\alpha = 2$ we have

$$\begin{aligned}\hat{\nu}_2(t_{k+1}) &= \int_{-\infty}^{t_{k+1}} \tau^{-2}(t_{k+1}-u)e^{-(t_{k+1}-u)/\tau}x(u)du \\ &= \int_{-\infty}^{t_k} \tau^{-2}(\Delta_k + t_k - u)e^{-(t_{k+1}-u)/\tau}x(u)du + x_k\int_{t_k}^{t_{k+1}} \tau^{-2}e^{-(t_{k+1}-u)/\tau}du \\ &= e^{-\Delta_k/\tau}\left(\frac{\Delta_k}{\tau}\hat{\nu}_1(t_k) + \hat{\nu}_2(t_k)\right) + x_k\left(1 - \frac{\Delta_k}{\tau}e^{-\Delta_k/\tau} - e^{-\Delta_k/\tau}\right),\end{aligned}$$

a recursion in terms of $\hat{\nu}_1$ and $\hat{\nu}_2$.

What does $\hat{\nu}_\alpha$ estimate?

Up until now, as in much of the traditional forecasting literature, we have made no distributional assumptions about $x(t)$. Nevertheless, if we assume that $x(t)$ it is a realisation of a random process $X(t)$ then by linearity

$$E[\hat{\nu}_\alpha(t)] = \int_{-\infty}^{t} \frac{\tau^{-\alpha}(t-u)^{\alpha-1}}{\Gamma(\alpha)} e^{-(t-u)/\tau} \nu(u) du, \tag{3.8}$$

where $\nu(u) = E[X(u)]$.

If we suppose that $\nu(u)$ is locally linear around time t, i.e. $\nu(u) \approx b_0 + (u-t)b_1$ then from (3.8)

$$E[\hat{\nu}_1(t)] \approx b_0 - b_1\tau \quad \text{and} \quad E[\hat{\nu}_2(t)] \approx b_0 - 2b_1\tau.$$

In other words, $\hat{\nu}_1(t)$ estimates the expected value of X at time $t-\tau$ and $\hat{\nu}_2(t)$ at $t-2\tau$.

Putting these together, we see that the expected value at time t is estimated by $2\hat{\nu}_1(t) - \hat{\nu}_2(t)$, the so-called *zero lag forecast* (Brown, 2004).

The zero lag forecast is also a kernel estimate with kernel

$$K(u) = \frac{e^{-u/\tau}}{\tau^2}(2\tau - u), \quad u \geqslant 0.$$

Note that $K(u)$ is negative when $u > 2\tau$ but tends to zero as $u \to \infty$. Furthermore

$$[2 + \tau^{-1}(s-t)]\hat{\nu}_1(t) - [1 + \tau^{-1}(s-t)]\hat{\nu}_2(t)$$

estimates the expected value at some time s in the future, $s > t$ and the local slope, b_1, is estimated by $\tau^{-1}(\hat{\nu}_1(t) - \hat{\nu}_2(t))$.

3.3.2 Detecting changes with an EWMA

Suppose as previously that $x(t)$ is the quoted log-price of an asset. To assess the difference between the current log-price and earlier log-prices, and hence detect changes and exploit mean reversion, we can look at the current log-price minus a kernel-based moving average of the log-price, for example

$$x(t) - \hat{\nu}_1(t).$$

Alternatively, in this case, we can look at the Stieltjes integral

$$\hat{\mu}_1(t) = \int_{-\infty}^{t} \tau^{-1} e^{-(t-u)/\tau} dx(u),$$

where exponential weights are applied to the *increments in the log-price*. It turns out that the two approaches are essentially the same since, integrating by parts, we have

$$\begin{aligned}
\hat{\mu}_1(t) &= \int_{-\infty}^{t} \tau^{-1} e^{-(t-u)/\tau} dx(u) \\
&= -[x(t) - x(u)]\tau^{-1} e^{-(t-u)/\tau}|_{-\infty}^{u=t} + \int_{-\infty}^{t} [x(t) - x(u)]\tau^{-2} e^{-(t-u)/\tau} du \\
&= \tau^{-1}[x(t) - \int_{-\infty}^{t} x(u)\tau^{-1} e^{-(t-u)/\tau} du] \\
&= \tau^{-1}[x(t) - \hat{\nu}_1(t)].
\end{aligned}$$

The only difference being the term τ^{-1} that converts the log-price change into change per unit time.

3.3.3 Iterating EWMAs

EWMAs can be iterated, i.e. the EWMA of an EWMA can be calculated. We can show that this does no more than produce a linear combination of two EWMAs. Write $\hat{\nu}_1(t, \tau_1)$ for the EWMA of X using half-life parameter τ_1. The EWMA of this EWMA using parameter $\tau_2 \neq \tau_1$ is then

$$\begin{aligned}\int_{-\infty}^{t} \tau_2^{-1} e^{-(t-u)/\tau_2} \hat{\nu}_1(u, \tau_1) du &= \int_{-\infty}^{t} \tau_2^{-1} e^{-(t-u)/\tau_2} \int_{-\infty}^{u} \tau_1^{-1} e^{-(u-w)/\tau_1} x(w) dw du \\ &= \int_{-\infty}^{t} (\tau_1 \tau_2)^{-1} e^{-t/\tau_2 + w/\tau_1} \int_{w}^{t} e^{-u(1/\tau_1 - 1/\tau_2)} du dw \\ &= \frac{\tau_2}{\tau_2 - \tau_1} \hat{\nu}_1(t, \tau_2) - \frac{\tau_1}{\tau_2 - \tau_1} \hat{\nu}_1(t, \tau_1).\end{aligned}$$

What does the difference between two EWMAs estimate?

Suppose that $E[X(u)] = \nu(u)$ is locally linear around time t so that $\nu(u) \approx b_0 + (u - t)b_1$ as before, then

$$E[\hat{\nu}_t(t, \tau_1)] \approx b_0 - b_1 \tau_1 \quad \text{and} \quad E[\hat{\nu}_t(t, \tau_2)] \approx b_0 - b_1 \tau_2.$$

It follows that b_1, the slope of the local trend, is approximated by

$$b_1 \approx \frac{\hat{\nu}_t(t, \tau_1) - \hat{\nu}_t(t, \tau_2)}{\tau_2 - \tau_1}$$

and the zero-lag approximation to b_0 is

$$b_0 \approx \frac{\tau_2 \hat{\nu}_t(t, \tau_1) - \tau_1 \hat{\nu}_t(t, \tau_2)}{\tau_2 - \tau_1}.$$

Combinations of these quantities give projections to future time points.

3.3.4 Estimation of parameters

In practice, parameters in a forecasting formula are determined by assessing the performance of the forecast. If the forecast aims to predict the price at some future point in time, then the quality of the forecast can be tested with historical data and the parameters adjusted accordingly. Taylor (2004) notes that 'The literature generally recommends that the smoothing parameters should be estimated from the data, usually by minimising the sum of *ex post* 1-step-ahead forecast errors'. These are the recommendations in Gardner (2006). In practice historical data is split into two sets: the 'in-sample' and the 'out-sample'. The method is optimised using the in-sample data and then performance is checked with the out-sample.

A typical experiment would be to take several days of market data and at each time point compare the EWMA smoothed price with the current price of the asset. This difference is then compared with the future return of the asset over say 30 seconds, i.e. the difference between the price 30 seconds ahead and the current price. This is repeated for each time point. The objective is to predict the future return from the EWMA difference. When the predicted future return is positive, this is an opportunity to buy and when it is negative an opportunity to sell. There are parameters in the gamma-kernel smoother, α and τ and parameters in the prediction procedure. All of these are then optimised to achieve the maximum empirical profit in the historical data. The approach is model-free and the optimal parameters are determined empirically.

3.3.5 Exponentially weighted least squares

Recursive smoothing and forecasting methods can be extended to include external factors. By formulating the problems as exponentially weighted least squares, recursions can be established for the regression parameters. There is a long history of this algorithm in tracking time-varying parameters (Brown, 2004).

Again using quoted log-price $y(t)$ as an example, let $y_j = y(t_j)$ where t_j are the times of log-price increments. Suppose that at time t_j we have available a vector of explanatory variables $\mathbf{a}_j$, $j = 1, 2, \ldots, k$. The objective is to predict y_k based on $\{y_i, i \leqslant k-1\}$ and $\mathbf{a}_k$. In high-frequency trading applications, the elements of $\mathbf{a}_k$ might include measurements such as bid-ask spread, bid-ask volume imbalance, number of quotes, trades during the immediate past and other characteristics of the order book, i.e. the difference between the amount available to buy and to sell at the best price. For brevity we will only consider the case where $\{t_j\}$ are equally spaced.

The exponentially weighted least squares estimate of the regression coefficients β_k at the time of the kth event is defined to be

$$\hat{\beta}_k = \underset{\beta}{\operatorname{argmin}} \sum_{j=0}^{k} (y_j - \mathbf{a}_j^T \beta)^2 \phi^{k-j}. \quad \text{where } 0 < \phi < 1. \tag{3.9}$$

From (3.9), differentiating the right hand side w.r.t. β gives $\hat{\beta}_k$,

$$\hat{\beta}_k = \left(\sum_{j=0}^{k} \mathbf{a}_j \mathbf{a}_j^T \phi^{k-j} \right)^{-1} \sum_{j=0}^{k} \mathbf{a}_j y_j \phi^{k-j} = R_k^{-1} P_k,$$

where R_k and P_k can be calculated recursively by

$$\begin{aligned} R_k &= \phi R_{k-1} + \mathbf{a}_k \mathbf{a}_k^T, \qquad R_0 = I \\ P_k &= \phi P_{k-1} + \mathbf{a}_k y_k, \qquad P_0 = \mathbf{a}_0 y_0. \end{aligned} \tag{3.10}$$

If R_k is nonsingular, then $\hat{\beta}_k$ is the unique solution for predicting $\tilde{y}_k$. Using a matrix inversion formula Rao (2009), the inverse of R_k can also be calculated recursively as

$$R_k^{-1} = \frac{1}{\phi} \left\{ I - \mathbf{v}_k \mathbf{a}_k^T \right\} R_{k-1}^{-1},$$

where

$$\mathbf{v}_k = R_{k-1}^{-1} \mathbf{a}_k / (\phi + \mathbf{a}_k^T R_{k-1}^{-1} \mathbf{a}_k).$$

In practice, there are difficulties with the R_k^{-1} recursion since R_k may be near-singular during patches of data where the feature vectors change very little and are highly correlated. The method can be stabilised by the addition of ridge terms λ_k on the diagonal of R_k from time to time before inversion, for example when indicated by the condition number of the matrix R_k.

$$R_k^{ridge} = R_k + \lambda_k I.$$

This makes R_k^{ridge} nonsingular, even if R_k is not of full rank, and was the main motivation for ridge regression when it was first introduced by Hoerl and Kennard (1970).

Ridge regression shrinks the estimated coefficients by imposing a penalty on the residual sum of squares. Here $\lambda_k \geq 0$ is a tuning parameter which controls the strength of the

penalty term. As such, it has the affect of shrinking the regression coefficients towards zero. This introduces some bias, but can greatly reduce the variance, resulting in a better mean-squared error. Large λ_k means more shrinkage, and so we get different $\hat{\beta}_k$ for different values of λ_k. Choosing an appropriate value of λ_k is hard but normally is done by 5-fold cross-validation in high-frequency trading analysis.

Ridge regression performs particularly well when there is a subset of true coefficients that are small or even zero. It does not do as well when all of the true coefficients are moderately large; however, in this case it can still outperform linear regression. See Hastie et al. (2001).

The prediction of the next value y_k, given $\mathbf{a}_1, \ldots, \mathbf{a}_k$ and $y_1, \ldots, y_{k-1}$, is then given by

$$\tilde{y}_k = \mathbf{a}_k^T \hat{\beta}_{k-1}.$$

The estimated coefficients $\{\hat{\beta}_k\}$ are updated by

$$\hat{\beta}_k = \hat{\beta}_{k-1} + R_k^{-1} \mathbf{a}_k (y_k - \tilde{y}_k).$$

Again it is worth emphasising that estimates obtained by exponentially weighting do not in general coincide with the optimal values derived from the Kalman filter of Section 3.4.3. They are convenient approximations that can be rapidly calculated. For further details and extensions to forward forecasting, see Brown (2004).

3.3.6 Moving averages for volatility and covolatility

In the interest of speed, it may be possible more generally to replace a locally defined linear statistic by a recursively defined approximation. As we shall see in Chapters 4 and 5, the realised volatility and the ZHY estimate of covolatility both have a linear form. These quantities can be expected to vary throughout the trading day. As a consequence local estimates defined in a window of time may be appropriate.

In essence, the idea is to replace the windowed sum $\sum_{i=n-m+1}^{n} a_i x_i$ (m fixed) for arbitrary sequences (x_i) and (a_i) with

$$\hat{h}(n) = \sum_{i \leqslant n} \phi^{n-i} a_i x_i, \quad 0 < \phi < 1$$

which can be calculated with the recursion

$$\hat{h}(n+1) = \phi \hat{h}(n) + a_{n+1} x_{n+1}. \tag{3.11}$$

3.4 Dynamic linear models

A simple strategy to profit from mean-reversion is to place thresholds at one standard deviation above and below the estimated mean-reverting level. Buy one unit of the portfolio if the low threshold is hit or sell one unit of the portfolio if the high threshold is hit. To take profit from the strategy is to close the position when the value of the portfolio is within a small interval of the mean-reverting level. This naive strategy is not guaranteed to make profit and in fact the trader may enter a long or short position which never reverted back to the inner band by the end of the trading horizon. This will induce potential losses into the

strategy as the trader tries to flatten his or her position. The wider the trigger thresholds, the less likely it is to end up flat.

To give some idea of the principles applied in successful trading we start by looking at an elegant technical paper by Thompson (2002) who develops optimal strategies for trading an asset driven by a hidden Markov process. He presents an intuitive derivation by viewing the optimal trading problem as a pair of simultaneous optimal stopping problems. We then go on to consider how to turn such ideas into practical trading strategies in terms of dynamic linear models. In this context, the trader will make a single round trip trade and there is no terminal time horizon.

In the simplest case, Thompson assumes the efficient log-price $\breve{X}(t)$ is an Ornstein-Uhlenbeck process with

$$d\breve{X}(t) = -\gamma\breve{X}(t) + \sigma dB(t),$$

where $B(t)$ is standard Brownian motion and $\gamma > 0$ is the reversion rate. We annotate X as $\breve{X}$ to distinguish the observable log-price from the hypothetical efficient log-price.

To take account of the cost of *paying the spread* in real life aggressive trading, each trade incurs a transaction cost of $c/2 > 0$. The efficient log-price is assumed to be observable without error. The trader can only hold either 1 or 0 units of the asset.

3.4.1 Optimal trading

Thompson shows that the optimal strategy to maximise long-term profit is to buy when $\breve{X}(t)$ hits a low threshold and sell when it hits a high threshold, This is perhaps not sur-

prising but Thompson gives the optimal thresholds precisely as $\pm b/\gamma$ where

$$2b - \gamma c = 2e^{-b^2/(\gamma\sigma^2)} \int_0^b e^{u^2/(\gamma\sigma^2)} du.$$

The equation has a unique solution in b for each pair (γ, σ).

Thompson's solution accords with intuition even though the precise form of the threshold is not immediately obvious. To translate Thompson's result into practice, the log-price process $\breve{X}(t)$ would need to be modelled in discrete time, e.g.

$$\breve{X}_i = a_i \breve{X}_{i-1} + \epsilon_i$$

with $\breve{X}_i = \breve{X}(t_i)$ and suitable choice of $\{a_i\}$ and $\{\epsilon_i\}$. As a consequence of this discrete form $\breve{X}_i$ can be expected to overshoot any fixed threshold before a decision is made. Nevertheless the intuition gained is helpful.

Of course it is impossible that a price will move persistently is such an artificial mean reverting manner. Thompson goes on to look at models where the log-price reverts to a moving level. The simplest of these is where $\theta(t)$ is the 'efficient log-price' process and

$$d\breve{X}(t) = dB(t) + L(\theta(t)), \tag{3.12}$$

with $L(x) = -\gamma x$ and where

$$\theta(t) = \breve{X}(t) - \sigma B'(t), \tag{3.13}$$

with $B'(t)$ a second independent Brownian motion. Since the process is linear and Gaussian the conditional distribution of $\hat{h}_t = \mathbb{E}\left[L(\theta(t))|\breve{X}(s), s \leqslant t\right]$ is normal. Furthermore

Thompson shows that in the steady state $\hat{h}_t$ proceeds as an Ornstein-Uhlenbeck process, for which the optimal trading solution has been previous been established.

Rewriting (3.12) and (3.13) as

$$d\breve{X}(t) = -\gamma\left(\breve{X}(t) - \sigma B'(t)\right) + dB(t),$$

we see that $\breve{X}(t)$ is a process that is trying to revert to the moving target $\sigma B'(t)$.

Again translating this in terms of discrete time we are led to consider a system of the following type

$$\begin{aligned}\breve{X}_i &= a_i\left(\breve{X}_{i-1} - \theta_i\right) + \epsilon_i \\ \theta_i &= \theta_{i-1} + \eta_i,\end{aligned} \tag{3.14}$$

where the $\{\theta_i\}$ are unobserved latent variables. Following Thomson we would then be led to calculate $\hat{\theta}_i = \mathbb{E}\left(\theta_i|\breve{X}_j, j \leqslant i\right)$ and to trade when this value exceeds thresholds symmetrically placed about zero.

3.4.2 Linear filtering

The system (3.14) is an example of a dynamic linear model (DLM). Such systems have been studied extensively in Bayesian system modelling. Software is available in R and other computing environments.

The general multivariate DLM is a highly flexible modelling framework. It encompasses many of the common statistical models as special cases such as multiple regression, exponential smoothing and mixed autoregressive-moving average processes (West and Har-

rison, 1989). The material below is taken from their book.

The general multivariate formulation of the DLM is

$$\begin{aligned} X_i &= F_i\theta_i + \epsilon_i, & \epsilon_i &\sim N_q(0, Q_i) \\ \theta_i &= G_i\theta_{i-1} + w_i, & w_i &\sim N_p(0, W_i), \end{aligned} \tag{3.15}$$

together with a specified initial prior distribution $\theta_0 \sim N_p(m_0, C_0)$. The terms F_i, G_i are specified matrices in the model formulation; Q_i and W_i are covariance matrices.

The process $\theta_i, i = 1, 2, \ldots$ is unobservable, but evolves with normally distributed increments. Observations, subject to normally distributed errors, are made at times $t_i, i = 1, 2, \ldots$.

Example: Trend models

In the simple one-dimensional model $\theta_i = \theta_{i-1} + w_i$ the forecast for $\theta_{i+\Delta}$ is constant given $x_{1:i}$, when $\Delta > 0$.

When local trends are suspected the following model may be more appropriate

$$\begin{aligned} X_i &= \eta_i + \epsilon_i, & \epsilon_i &\sim N(0, Q_i) \\ \eta_i &= \eta_{i-1} + \beta_{i-1} + w_i^{[1]}, & w_i^{[1]} &\sim N(0, W_i^{[1]}) \\ \beta_i &= \beta_{i-1} + w_i^{[2]}, & w_i^{[2]} &\sim N(0, W_i^{[2]}). \end{aligned}$$

In terms of the DLM, we have $\theta_i = (\eta_i, \beta_i)^\mathsf{T}$, $F_i = (1, 0)$,

$$G_i = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad \text{and} \quad W_i = \begin{bmatrix} W_i^{[1]} & 0 \\ 0 & W_i^{[2]} \end{bmatrix}.$$

With this model the underlying expectation increases linearly.

3.4.3 Updating the DLM – Kalman filter

Suppose that at stage k,

$$\theta_k | x_{1:(k)} \sim N_p(m_k, C_k).$$

The mean and variance-covariance of θ_{k+1} given $x_{1:k}$ are then

$$a = Gm_k \quad \text{and} \quad R = GC_kG^\mathsf{T} + W_{k+1},$$

(dropping subscripts for ease of notation), so that

$$\begin{bmatrix} X_{k+1} \\ \theta_{k+1} \end{bmatrix} \sim N_{p+q} \left(\begin{bmatrix} Fa \\ a \end{bmatrix}, \begin{bmatrix} FRF^\mathsf{T} + Q & FR \\ RF^\mathsf{T} & R \end{bmatrix} \right).$$

Applying the conditional multivariate normal formula, we conclude that

$$\theta_{k+1} | x_{1:k+1} \sim N_p(m_{k+1}, C_{k+1}),$$

where

$$m_{k+1} = a + RF^{\mathsf{T}}\left(FRF^{\mathsf{T}} + Q\right)^{-1}(X_{k+1} - Fa)$$
$$C_{k+1} = R - RF^{\mathsf{T}}\left(FRF^{\mathsf{T}} + Q\right)^{-1}FR.$$

This is the *Kalman filter* and $A = RF^{\mathsf{T}}\left(FRF^{\mathsf{T}} + Q\right)^{-1}$ is called the *Kalman gain*.

3.4.4 Comparison between EWMA and Kalman filter

For comparison of the EWMA and Kalman filter, let $X_i = x_i$ be the log-price of an asset observed at time t_i and suppose the underlying unobserved 'fair log-price' at that time is $\theta_i = \theta(t_i)$, i.e. θ_i is the 'efficient log-price' $\breve{X}_i$ in Section 3.4.1. Now assume that $\theta(t)$ is Brownian motion with mean zero and infinitesimal variance σ^2 and that given θ_i the variables X_i are independently distributed with mean θ_i and variance η^2. Furthermore at time zero, $\theta_0 \sim N(m_0, c_0^2)$.

Referring to (3.15) we have $F_k = G_k = 1$, $Q_k = \eta^2$ and $W_k = \sigma^2(t_k - t_{k-1})$. It follows from the conclusion of Section 3.4.3 that

$$\theta_k | x_{1:k} \sim N(m_k, c_k^2),$$

where

$$m_{k+1} = m_k + \frac{c_k^2 + \sigma^2(t_{k+1} - t_k)}{c_k^2 + \sigma^2(t_{k+1} - t_k) + \eta^2}\left(x_{k+1} - m_k\right).$$

In other words m_{k+1}, the posterior expectation of $\theta_{k+1}|x_{1:k+1}$ is given by the recursion

$$m_{k+1} = m_k\lambda_k + (1 - \lambda_k)x_k, \quad \text{where} \quad \lambda_k = \frac{c_k^2 + \sigma^2\Delta_k}{c_k^2 + \sigma^2\Delta_k + \eta^2},$$

and $\Delta_k = t_{k+1} - t_k$. This can now be compared with (3.5) and we see that in general, although the recursion coefficients differ, the recursive form is the same. There is simplification when $\{t_i\}$ are equally spaced, in which case c_k converges to a limit as $k \to \infty$. In this limiting case, the Kalman filter is the same as the EWMA.

3.5 Discontinuities

Movements of the market over a protracted period of time, hours or days, are generally not of interest to high-frequency traders. Nevertheless the recovery of prices after significant news events is a very visible example of mean reversion and for this reason we conclude with a brief discussion of the impact of disruptive events on financial markets. Certain events are known to produce temporary changes to the trading environment. Foreign exchange markets are affected by the expiry of currency options typically scheduled at 3pm Tokyo time or 10am New York time. The New York expiry usually has the largest effect because it incorporates both European and North American sentiment. There are also daily currency fixings at around 9am in Tokyo and 4pm in London. These fixings determine the prices of currencies for commercial transactions. Some currency pairs may see significant upturn in trading prior to the fixing time with attendant price movements. In many cases the price will return to the previous level.

Of course, not all jumps are as sharply identifiable as that shown in Figures 3.1 and 3.2.

These figures illustrate the rapid price change that occurs at around 13:30 on the first Friday of each month when Non-Farm Payroll numbers are announced. Similar jumps occur at other known times when, for example, FX fixings are made each day. Figure 3.2 shows the total volumes in the first 5 rungs of the order book (logarithmic scale). We

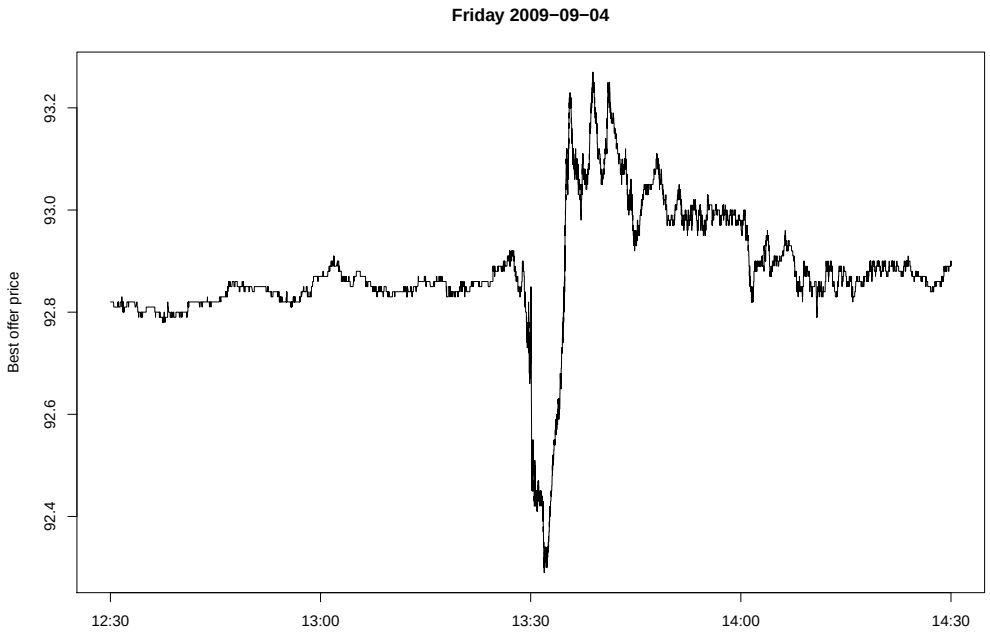


Figure 3.1: USDJPY exchange rates.

can see that immediately prior to the news event, the volume dries up, corresponding to the cautious behaviour of traders in the market, since they are wary about the direction and magnitude of potential price movement. After the event the volume returns to the previous level but it is much less stable.

Scheduled news announcements such as the release of US unemployment figures can produce major price movement in a wide range of instruments including Interest Rate Futures and foreign exchange rates. Market makers usually exercise caution immediately prior to such announcements, withdrawing or substantially modifying resting orders with the effect of thinning out the visible order book. Under these circumstance the best prices at bid and offer may differ substantially and are maintained with only small volumes available at these prices. As a result when market activity resumes and the impact of the announcement is absorbed, prices may swing wildly before settling into their usual

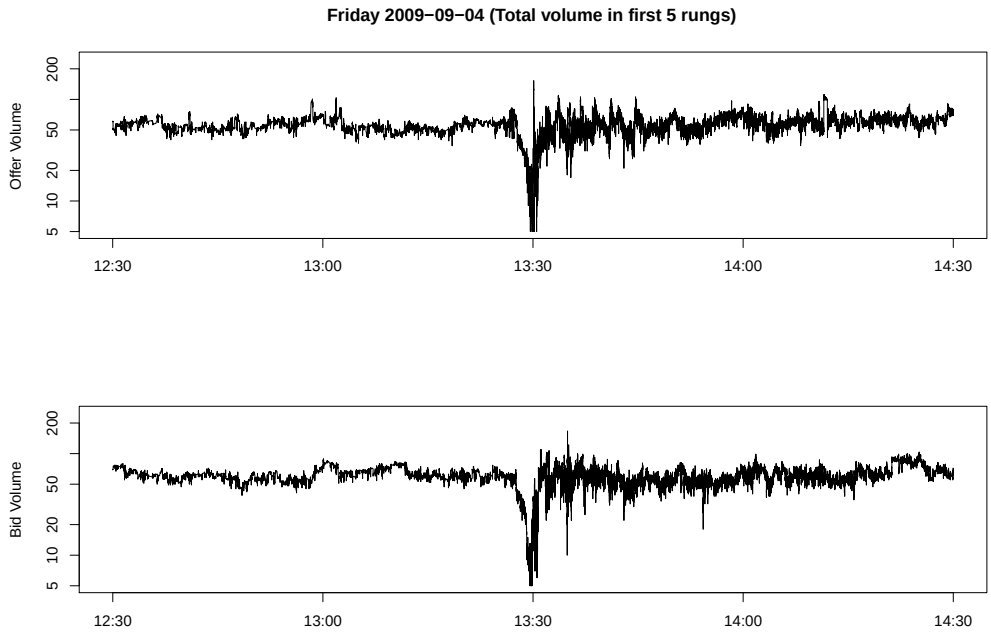


Figure 3.2: USDJPY exchange rates.

pattern.

Unscheduled news events relating to local or international incidents can also produce large price movements or *jumps*, primarily in the asset impacted by the news content but also secondarily in correlated instruments, local and internationally. There is an extensive literature supporting the importance of jumps in understanding the mechanism of price movement. See for example Bollerslev et al. (2008) and Corsi et al. (2010). Evans and Lyons (2005) show that in some situations the market may still be absorbing and adjusting to major news releases for hours and sometimes days after the announcement. Relative movements can be seen in the first and second day with trading activity still elevated in the 3rd and 4th days.

Unexpectedly large trades can have similar effects. A buy trade that takes out several levels of the order book will expose a gap between the best bid and ask prices in the book.

Resiliency provided by market makers will act to fill the gap but the resulting mid-price may drift upwards and fail to revert to the value that it had prior to the trade.

The movements seen at the time of scheduled news releases are not of primary concern to high-frequency traders who close down their positions in advance. They are however of major importance to hedge fund managers who hold large positions in the market.

The statistical methods used for these unusual events are highly specific, making use of external features coupled with ARMA modelling. There is extensive literature on the subject. With relevance to foreign exchange markets, see for example Love (2005) and for Eurex interest rate futures, see Almgren (2012) and the references therein.

Chapter 4

Volatility

In statistical terms the study of financial volatility is a scale problem. It is concerned with the fluctuation of price about a particular direction of movement rather than the direction itself. For most purposes the location component of the statistical problem can be ignored. In its most common usage, volatility refers to the standard deviation of the change in the log-price of an asset over some fixed period of time (although in the financial literature, volatility may also refer to the variance). Individual volatilities also play an important role when considering correlations between different assets.

We will see that there are two main approaches to volatility and covolatility estimation. One is built around the concept of 'stochastic volatility' where volatility is viewed as a latent unobserved stochastic process. The other is centred on 'time-change' where volatility is considered to be constant on an alternative non-physical time scale. Both approaches enable the cumulative volatility to be estimated over a given interval of physical time.

In this chapter and the next, we look back at two remarkable papers by Zhou (1995,1996). This is partly to emphasise their historical importance in the early development of the

subject but also to show they provide common ground for the two main estimation approaches described above. Additionally by developing a systematic method of proof for the theorems in Zhou (1996) we are able to show that Zhou's methods can provide an 'infill consistent' estimator of cumulative volatility as conjectured by Hansen and Lunde (2006), albeit under rather strong conditions on the infill observation times.

There is an extensive literature on volatility and covolatility in financial mathematics and econometrics. McAleer and Medeiros (2008) and Aït-Sahalia and Mykland (2009) provide substantial reviews. The recently published compendium edited by Bauwens et al. (2012) has contributions from 39 authors, with almost 1000 references.

Although the practicalities of high-frequency trading favour the time-change approach, we outline key results in modern stochastic volatility theory and indicate conditions under which powerful asymptotic results have been obtained. We start by describing the standard model framework for inference, outlining key results and defining a basic class of volatility estimators. We explain the use of the *signature plot* as a graphical diagnostic but argue that the variogram provides a more natural plot providing greater insight. Various volatility estimators that have been proposed in the literature are shown to be equivalent to simple summaries of the empirical variogram.

The time-change approach opens up the possibility of using classical Kalman filtering methods in the Dynamic Linear Model framework. These techniques are illustrated in Section 4.3. We then consider criticisms that have been levied at the standard model. A Monte Carlo goodness of fit test for the model is proposed and its significance is assessed for financial data. Finally we apply our modified volatility estimator to the problem of estimating daily patterns of volatility in the US Dollar – Japanese Yen exchange rate.

4.1 Latent price – physical time

There is a long tradition in mathematical finance of using continuous time models for describing the behaviour of asset returns. Empirical evidence shows that logarithmic returns on equities, currencies and commodities can be approximately modelled as normal mixtures with varying standard deviations. For a detailed discussion with application to the movement of foreign exchange rates, see Dacorogna et al. (2001).

Much of the theoretical work on volatility problems is carried out in the framework of stochastic differential equations (SDEs) and Itô calculus. A basic and traditional model for the latent 'efficient' log-price, $\breve{X}(t)$, of a financial asset is

$$d\breve{X}(t) = \sigma(t)dB(t), \tag{4.1}$$

where $B(t)$ is standard Brownian motion and the infinitesimal standard deviation $\sigma(t)$ is either deterministic or stochastic. The value $\sigma(t)$ is also called the *instantaneous volatility* at time t. Usually t represents ordinary physical time – sometimes termed 'wall-clock' time. Various models for the stochastic evolution of $\sigma(s)$ have been proposed; predominant among these are the ARCH class of models and related generalisations. For a wide historical review of the mathematical modelling of financial time series, see (Shephard, 2015).

Note that the stochastic differential equation above has no drift term. This is in accordance with a fundamental principle of mathematical finance, the martingale property in which the current log-price is supposed to incorporate all available information and hence represent the consensus prediction of expected log-price going forward, at least in the short term. The minimal assumption of no arbitrage in the price process leads, under some

regularity conditions, to the consequence that the price process is a semi-martingale; see Delbaen and Schachermayer (1994).

4.1.1 Realised volatility

Suppose $\breve{X}_1$ is the latent efficient log-price at time t_1 and $\breve{X}_2$ is the value at time t_2. The change in value, $\breve{X}_2 - \breve{X}_1$ is called the *return* or *log-return* over the *return period* (t_1, t_2). In the above model (4.1), the return over a unit of time $(0, 1)$ has mean zero and variance

$$v = \text{Var}[\breve{X}(1) - \breve{X}(0)] = \int_0^1 \sigma^2(s)ds, \tag{4.2}$$

conditional on σ.

The *theoretical volatility* over $(0, 1)$ is the square root of this quantity. Using the additive property of variances for processes with conditionally independent increments (conditional on σ), the theoretical volatility (4.2) can be written as

$$\sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} \sigma^2(s)ds = \sum_{i=1}^{n} \mathbb{E}\,(R_i^2),$$

where $R_i = \breve{X}_i - \breve{X}_{i-1}$ is the return over the sub-interval (t_i, t_{i-1}) and $(t_i, i = 0, 1, \ldots, n)$ is an increasing sequence of times with $t_0 = 0$ and $t_n = 1$.

Assuming that the returns $\{R_i = r_i\}$ are observable, $\sum_{i=1}^{n} r_i^2$ is an obvious estimate of v. With observations $x_i = x(t_i)$ at $t_i, i = 0, \ldots, n$, the statistic

$$\hat{v}(x) = \sum_{i=1}^{n} (x_i - x_{i-1})^2,$$

is called the *realised variance* of x over $(0, 1)$. The *realised volatility* RVol(x) is the

square root of $\hat{v}(x)$ but note again the confusion between variance and volatility in the financial terminology.

The basic philosophy of financial mathematics is there is hidden latent efficient price $\breve{X}$ which is not observable directly. Merton (1980) noticed that if the hypothetical values of $\breve{X}$ are available on an arbitrarily fine time scale, v can be recovered with arbitrary precision, under the model in (4.1). Subsequently Andersen et al. (2001), Barndorff-Nielsen and Shephard (2002) and Mykland and Zhang (2006) among others, show this rigorously by considering the realised variance

$$\hat{v}(\breve{X}; n) = \sum_{i=1}^{n} \left(\breve{X}(t_i^{(n)}) - \breve{X}(t_{i-1}^{(n)}) \right)^2$$

for a sequence of partitions $(t_0^{(n)} = 0, t_1^{(n)}, \ldots, t_n^{(n)} = 1)$ where $\max_i |t_i^{(n)} - t_{i-1}^{(n)}| \to 0$ as $n \to \infty$.

Furthermore Barndorff-Nielsen and Shephard (2002), under quite general conditions, show that $\hat{v}(\breve{X}; n)$ is asymptotically normally distributed. In their analysis, the instantaneous volatility is assumed to have locally square integrable sample paths, while being stationary and stochastically independent of the Brownian motion $B(t)$, and most importantly, it assumed that the latent efficient price is available. Taking $t_i^{(n)} = i\Delta, i = 0, \ldots, n$, $\Delta = n^{-1}$, they show

$$\Delta^{-1/2} \frac{\hat{v}(\breve{X}; n) - \int_0^1 \sigma^2(s) ds}{\sqrt{2 \int_0^1 \sigma^4(s) ds}} \to N(0, 1), \tag{4.3}$$

as $\Delta \to 0$.

This type of result is called *infill asymptotics* – corresponding to an increasingly high frequency of observations within an interval of fixed total length. Here Δ^{-1} plays the rôle

of n in the traditional area of asymptotic statistical analysis.

In a series of paper reported in Barndorff-Nielsen and Shephard (2004b) and the references therein, the authors introduce the *bipower* estimate of integrated variance, namely

$$\hat{v}^{[1,1]}(x;n) = \sum_{i=1}^{n-1} \left| x(t_i^{(n)}) - x(t_{i-1}^{(n)}) \right| \left| x(t_{i+1}^{(n)}) - x(t_i^{(n)}) \right|.$$

The estimator, $\hat{v}^{[1,1]}(\breve{X};n)$, is shown to be robust to additional jump terms in the model specification of the latent log-price process $\breve{X}$ and hence by comparison with the realised variance they are able to detect and quantify the contributions of jumps in latent log-price volatility.

Barndorff-Nielsen and Shephard (2004b, Section 5.4) report that, jointly,

$$\frac{\Delta^{-1/2}}{\sqrt{\int_0^1 \sigma^4(u)du}} \begin{pmatrix} \hat{v}(\breve{X};n) - v \\ \hat{v}^{[1,1]}(\breve{X};n) - v \end{pmatrix} \to N\left\{ 0, \begin{pmatrix} 2 & 2 \\ 2 & 2.60907 \end{pmatrix} \right\},$$

as $\Delta \to 0$, under the conditions of (4.3). These results show how much statistical information is lost in using the bipower estimator with latent log-prices.

Annualised volatility In the financial world, daily realised volatility, $\text{RVol}_{\text{Day}}(x)$, is usually extrapolated to give an estimate of yearly volatility, by assuming additivity for variances. The extrapolated figure is then multiplied by 100 and reported as a percentage. So with a nominal 252 trading days in the year, the annualised volatility is

$$\text{RVol}_{\text{Annual}} \approx 100\sqrt{252}\, \text{RVol}_{\text{Day}}(x)\%.$$

4.1.2 Signature plots and variograms

It is tempting to conclude that the log-price process should be sampled at the highest frequency to produce the best estimate of v. As a check it is helpful to plot $\hat{v}_n$ for various values of the sampling interval, Δ. The resulting graph is called a *signature plot* (Andersen et al., 2000). Typically the curve is not flat, as would be predicted by the model in (4.1). Very often the curve increases as the Δ becomes small and the sampling frequency increases, see Figure 4.1.

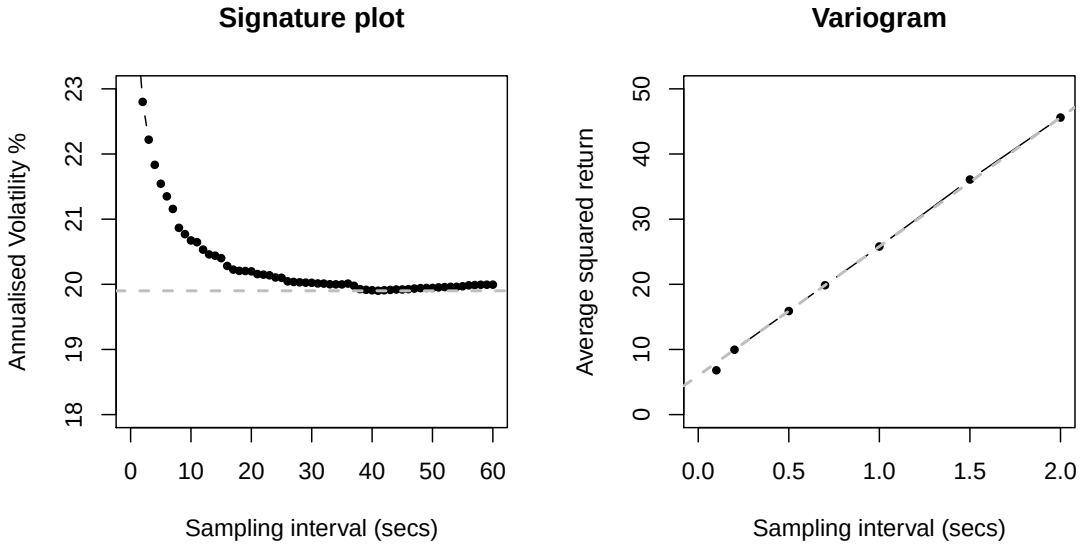


Figure 4.1: Signature and variogram plot: USD/EUR trade data 2008-10-09 (EBS)

The variogram plot is an alternative way of showing the relationship between volatility and sampling frequency. It is used extensively in geostatistics (Cressie, 1993). The theoretical variogram of a one-dimensional process X with stationary increments is defined to be

$$\Gamma(\tau) = \mathbb{E}\left[X(t+\tau) - X(t)\right]^2, \quad \tau > 0.$$

It can be estimated by

$$\hat{\Gamma}(\tau) = \frac{\sum_{(i,k)\in M(\tau)}[x(t_i) - x(t_k)]^2}{|M(\tau)|}, \quad \tau > 0,$$

where $M(\tau)$ is the set of pairs (i, k) such that $t_k = t_i + \tau$ and $|M(\tau)|$ is the cardinality of $M(\tau)$. In practice, it may not be possible to find points (t_i, t_k) that are exactly separated by τ, in which case a certain tolerance is allowed in defining $M(\tau)$. Figure 4.1 compares the estimated variogram with the signature plot. Note that in this case the variogram is approximately linear with a positive intercept. For discussion of the use of variograms in volatility estimation with an illustration, see Section 4.2.6.

4.2 Observed price – physical time

In an early paper, Roll (1984) suggested that observations of the process in (4.1) should be modified by the introduction of a random error term. Roll was motivated by problems in dealing with trade data when there is no information about whether the trade is from the bid or the ask side of the order book. In its simplest form the revised model for observations at times $\{t_i\}$ is

$$X_i = \breve{X}_i + \epsilon_i, \quad i \in \mathbb{Z}, \tag{4.4}$$

where $\breve{X}_i = \breve{X}(t_i)$ is the latent efficient log-price, $X_i = X(t_i)$ is the observed value, and $\{\epsilon_i\}$ are iid with constant variance η^2. The model has become widely accepted in mathematical finance as a suitable framework for volatility analysis. In this model the efficient price $\breve{X}(t_i)$ cannot be observed and the observed price $X(t_i)$is the efficient price plus perturbations. The perturbations $\{\epsilon_i\}$ are often termed *microstructure noise*.

With X given by (4.4), the realised variance is no longer an unbiased estimator of the integrated variance v_T. This noise component produces an inflated effect that makes the volatility more than what it should be. Assuming equally spaced sampling with $\Delta = 1/n$, it follows that

$$\mathbb{E}\,\hat{v}(X) = \frac{2\eta^2}{\Delta} + \int_0^1 \sigma^2(s)ds = \frac{2\eta^2}{\Delta} + v. \tag{4.5}$$

There is a discontinuity as $\Delta \to 0$, in qualitative agreement with observed signature plots (Figure 4.1). Similarly, the variogram estimator has approximate expectation

$$\mathbb{E}\,\hat{\Gamma}(\Delta) \approx 2\eta^2 + \Delta \int_0^1 \sigma^2(s)ds, \tag{4.6}$$

for small Δ, i.e. linear with slope equal to the integrated variance $\int_0^1 \sigma^2(s)ds$ and intercept twice the microstructure variance – again in agreement with observation (Figure 4.1)

There is an extensive literature on methods of improving the estimator $\hat{v}(X)$ in the context of model 4.4. As noted in Section 4.1.1, when the underlying efficient price is observable then its volatility can be obtained with arbitrary precision using data observed on an arbitrary high-frequency timescale. However historically, as prices started to be observed with increasingly high frequency, it was noticed that microstructure noise became dominant. So instead of providing increased accuracy the realised volatility became more and more biased. Compromises were made to observe the price over time intervals such as 1 minutes or 1 second. A basic understanding of these compromises is contained in an early paper by Zhou (1996), submitted to the Journal of Business & Economic Statistics in 1993, over 20 years ago. These ideas are discussed in the next section.

4.2.1 Zhou's contribution

Zhou (1996) describes his experiences in working with tick-by-tick data in foreign exchange trading where a succession of proposed exchange rates are seen at irregular times on the Reuters news feed. He observed that successive returns of these prices over short periods of time are highly negatively correlated. For trade data this might be explained by the bid-ask bounce phenomenon illustrated in Figure 2.4. However, Zhou found similarly large negative correlation when using the bid price alone. He then proposed model (4.4) as an explanation. He showed how to estimate the underlying volatility unbiasedly, stated the variance of his estimator and indicated how this result could be obtained.

Zhou's estimator for the integrated variance, $v = \int_0^1 \sigma^2(s)ds$, is

$$G = \sum_{i=1}^{n} R_{i,1} R_{i+1,3}, \tag{4.7}$$

where $R_{i,k} = X_i - X_{i-k}$, the return over the period (t_{i-k}, t_i). Zhou's key observation is that the error terms in

$$\begin{aligned} R_{i,1} &= \breve{X}_i - \breve{X}_{i-1} + \epsilon_i - \epsilon_{i-1} \\ R_{i+1,3} &= \breve{X}_{i+1} - \breve{X}_{i-2} + \epsilon_{i+1} - \epsilon_{i-2}, \end{aligned} \tag{4.8}$$

are independent, and since $R_{i+1,3}$ can be written as a sum of independent increments

$$R_{i+1,3} = \breve{X}_{i+1} - \breve{X}_i + \breve{X}_i - \breve{X}_{i-1} + \breve{X}_{i-1} - \breve{X}_{i-2} + \epsilon_{i+1} - \epsilon_{i-2} \tag{4.9}$$

then

$$\mathbb{E}\, R_{i,1} R_{i+1,3} = \mathbb{E}\,(\breve{X}_{i-1} - \breve{X}_i)^2 = \int_{t_{i-1}}^{t_i} \sigma^2(s)ds \stackrel{\text{defn}}{=} v_i^2. \tag{4.10}$$

Taking $t_0 = 0$ and $t_n = 1$, it follows that G is an unbiased estimator of $v = \int_0^1 \sigma^2(s)ds$. Zhou is the first person to suggest the estimator G. .

As pointed out by Zhou, the variance of G is complicated for arbitrary non-constant $\{v_i\}$, and Zhou does not derive an expression for the variance in the general case. We derive the general variance in Theorem 4.3.

However by assuming that $v_i = v/n, i = 1, 2, \ldots, n$, Zhou obtains the following.

Theorem 4.1 (Zhou, 1996)**.** *Let* $\breve{X}_i - \breve{X}_{i-1} \sim N(0, v/n)$ *and* $\epsilon_i \sim N(0, \eta^2)$*, independently* $i \in \mathbb{Z}$*, then*

$$\operatorname{Var} G = \frac{v^2}{n^2}(6n - 2) + 8v\eta^2 + (8n - 6)\eta^4. \tag{4.11}$$

Remark. There is a typographical error in Zhou's statement of the theorem, corrected above – the coefficient in last term should be $(8n - 6)$ not $(8n - 4)$. Zhou indicates a method of proof without full details. To obtain the result it is necessary to assume that the random variables are normally distributed or at least have zero kurtosis. The method of proof we develop below will be helpful in later parts of this chapter.

Although the assumption of constant $\{v_i\}$ seems restrictive, it will be applicable when working on volatility stabilising activity-timescale, for example, when observing prices at transaction times. See the discussion in Section 4.3.

We start by proving the theorem under Zhou's conditions and derive a more general version later in Theorem 4.3.

Proof. Let $Z = (Z_1, \ldots, Z_{2n+6})^{\mathrm{T}}$ be a vector of iid $N(0, 1)$ variables, then from (4.8)

and (4.9), we see that G has the same distribution as the quadratic form

$$Z^{\mathrm{T}}[\alpha I_R : \eta J_R]^{\mathrm{T}}[\alpha I_S : \eta J_S]Z, \tag{4.12}$$

where $\alpha = \sqrt{(v/n)}$ and I_R, J_R, I_S and J_S are $n \times (n+3)$ dimensional Toeplitz matrices with first rows given respectively by

$$\begin{aligned}
&\text{Row 1: } I_R = (0, 0, 1, 0, \ldots, 0)\\
&\text{Row 1: } J_R = (0, -1, 1, 0, \ldots, 0)\\
&\text{Row 1: } I_S = (0, 1, 1, 1, 0, \ldots, 0)\\
&\text{Row 1: } J_S = (-1, 0, 0, 1, 0, \ldots, 0)
\end{aligned} \tag{4.13}$$

The remaining elements in their first columns are zero.

From standard multivariate normal theory (Rencher and Christensen, 2012), we have

$$\mathbb{E}\, Z^{\mathrm{T}} H Z = \text{trace} H \quad \text{and} \quad \text{Var}\, Z^{\mathrm{T}} H Z = 2 \text{ trace} \left\{ \left(\frac{H + H^{\mathrm{T}}}{2} \right)^2 \right\}.$$

In our case

$$H = \begin{pmatrix} \alpha^2 I_R^{\mathrm{T}} I_S & \alpha\eta I_R^{\mathrm{T}} J_S \\ \alpha\eta J_R^{\mathrm{T}} I_S & \eta^2 J_R^{\mathrm{T}} J_S \end{pmatrix},$$

so that

$$\begin{aligned}
\mathbb{E}\, G = \text{trace} H &= \alpha^2 \text{ trace } I_R^{\mathrm{T}} I_S + \eta^2 \text{ trace } J_R^{\mathrm{T}} J_S \\
&= \alpha^2 n + 0 = \frac{v}{n} n = v.
\end{aligned}$$

For an arbitrary symmetric square matrix

$$\text{trace } A^2 = \sum_i \sum_j A_{ij}^2 = ||A||_2^2.$$

The variance of G is then

$$\text{Var}\, G = \frac{1}{2}\alpha^4 ||I_R^{\mathrm{T}} I_S + I_S^{\mathrm{T}} I_R||_2^2 + \alpha^2\eta^2 ||I_R^{\mathrm{T}} J_S + I_S^{\mathrm{T}} J_R||_2^2 + \frac{1}{2}\eta^4 ||J_R^{\mathrm{T}} J_S + J_S^{\mathrm{T}} J_R||_2^2. \quad (4.14)$$

Inspecting the pattern of these matrix products we have, by induction,

$$\begin{aligned} ||I_R^{\mathrm{T}} I_S + I_S^{\mathrm{T}} I_R||_2^2 &= 12n - 4 \\ ||I_R^{\mathrm{T}} J_S + I_S^{\mathrm{T}} J_R||_2^2 &= 8n \\ ||J_R^{\mathrm{T}} J_S + J_S^{\mathrm{T}} J_R||_2^2 &= 16n - 12. \end{aligned}$$

Substituting these values in (4.14) with $\alpha = \sqrt{(v/n)}$ gives the required variance. □

From 4.11, it is easy to see that increasing the observation frequency infinitely does not improve the accuracy of the volatility estimation. The frequency of observation is determined by the number of observations, n. If n increases and everything else is kept constant, the term which produces the problem is $8n\eta^4$. Increasing the observation frequency does not improve the accuracy. Nevertheless, even with this equation, there will be an optimal value for n. It is near $\sigma^2\sqrt{3}/(2\eta^2)$ when η^2/σ^2 is small. For example, when there is significant noise in the data, fewer data sometimes are better than more data. This allows us to quantify the penalty we pay by increasing the frequency. Again Zhou was the first person to discover this optimal observation frequency (Zhou, 1996). Before him, it was just common knowledge that data should not be observed too frequently when there

is observation noise in the data.

4.2.2 Longer return periods

In the same paper, Zhou (1996) proposed an improved estimator based on return periods of length k. His idea is to use observations every kth tick but look at all the different starting points. For example, start at 1, go to $k+1$; start at 2 go to $k+2$; start at 3 go to $k+3$ and so on, then overlap the returns, i.e. take the average. The technique was eventually became known as *subsampling*. For some reason or other the term *subsampling* has stuck, although it can be better described *oversampling*.

Zhou's improved estimator can be written as

$$G_k = \frac{1}{k}\sum_{i=1}^{n} R_{i,k}R_{i+k,3k}. \tag{4.15}$$

This idea was picked up again ten years later by the Per Mykland school, among others. Zhou showed that the variance of G_k is bounded as follows

$$\operatorname{Var} G_k \leq v^2\left(6\frac{k}{n} + 8\frac{\eta^2}{kv} + 8\frac{n\eta^4}{k^2v^2}\right).$$

Zhou shows that G_k is biased for the integrated variance, with expectation

$$\mathbb{E}\, G_k = \int_0^1 \sigma^2(s)ds + \sum_{i=1}^{n}\frac{i}{k}(v_{i-k} - v_{n-i+1}). \tag{4.16}$$

Under the same special conditions of constant $\{v_i = v/n\}$ he obtains an upper bound for the variance of the estimator. For completeness, under the same conditions, we will give

an explicit formula for the variance of a more general estimator namely

$$G_{k,j} = \frac{1}{k}\sum_{i=1}^{n} R_{i,k}R_{i+j,k+2j}. \tag{4.17}$$

Theorem 4.2. *Let* $\breve{X}_i - \breve{X}_{i-1} \sim N(0, v/n)$ *and* $\epsilon_i \sim N(0, \eta^2)$*, independently* $i \in \mathbb{Z}$*, then*

$$\begin{aligned}\operatorname{Var} G_{k,j} = \frac{v^2}{3n^2k}\left[(12jk + 4k^2 + 2)n - (k+j)(k^2 + 3jk - 1)\right] \\ + \frac{8v\eta^2}{k} + \frac{(8n - 4j - 2k)\eta^4}{k^2}.\end{aligned} \tag{4.18}$$

Proof. The method of proof is similar to that in Theorem 4.1. This time I_R, J_R, I_S and J_S are $n \times (n + 2j + k + 1)$ dimensional Toeplitz matrices with first rows

$$\begin{aligned}&\text{Row 1: } I_R = (0^{j+1}, 1^k, 0^{n+j})\\ &\text{Row 1: } J_R = (0^j, -1, 0^{k-1}, 1, 0^{n+j})\\ &\text{Row 1: } I_S = (0, 1^{k+2j}, 0^n)\\ &\text{Row 1: } J_S = (-1, 0^{k+2j-1}, 1, 0^n),\end{aligned}$$

and again the remaining elements in their first columns are zero. The notation x^n represents a sequence with the value x repeated n times. We find

$$\begin{aligned}||I_R^{\mathrm{T}} I_S + I_S^{\mathrm{T}} I_R||_2^2 &= \frac{2k}{3}[(12jk + 4k^2 + 2)n - (k+j)(k^2 + 3jk - 1)]\\ ||I_R^{\mathrm{T}} J_S + I_S^{\mathrm{T}} J_R||_2^2 &= 8nk\\ ||J_R^{\mathrm{T}} J_S + J_S^{\mathrm{T}} J_R||_2^2 &= 16n - 8j - 4k.\end{aligned}$$

The proof is by induction – details are omitted for brevity. The variance has the same

form as (4.14). By substitution we have the required variance. □

By taking $j = k$ in (4.18) we have the variance of Zhou's second estimator

$$\text{Var}\, G_k = \frac{v^2}{3n^2k}\left[(16k^2+2)n - 8k^3 + 2k\right] + \frac{8v\eta^2}{k} + \frac{(8n-6k)\eta^4}{k^2},$$

which is indeed smaller than the upper bound given by Zhou, namely

$$\text{Var}\, G_k \leqslant \frac{6v^2k}{n} + \frac{8v\eta^2}{k} + \frac{8n\eta^4}{k^2}.$$

The two expressions differ in the first term. For large $k < n$ the dominant coefficient of v^2k/n is $5\frac{1}{3}$ in the exact variance and 6 in the bound. Note that the variance diverges as $n \to \infty$ for fixed k, so the estimator is not consistent. Nevertheless, the variance is finite and for given η^2/v and n , k can be chosen to make the variance as small as possible. Zhou gives guidance in this respect.

To summarise, Zhou (1996) introduced two new ideas. One is to deal with modelling microstructure noise and the other is to introduce the ideas of subsampling. He made two major contributions in this early paper, but it took a long time to find its way into the financial academic literature.

4.2.3 Inhomogeneous variance components

We return to Theorem 4.1 and now consider the case when $\{v_i\}$, defined in (4.10), are not constant.

Theorem 4.3. *Let* $\breve{X}_i - \breve{X}_{i-1} \sim N(0, v_i)$ *and* $\epsilon_i \sim N(0, \eta^2)$*, independently* $i \in \mathbb{Z}$*, then*

G in (4.7) *is an unbiased estimator of* $v = \int_0^1 \sigma^2(u)du$*, with*

$$\text{Var}\, G = v_1 v_0 + v_n v_{n+1} + 4\sum_{i=1}^{n-1} v_i v_{i+1} + 2\sum_{i=1}^{n} v_i^2 + \left(2(v_0 + v_{n+1}) + 6(v_1 + v_n) + 8\sum_{i=1}^{n-1} v_i\right)\eta^2 + (8n-6)\eta^4 \quad (4.19)$$

Proof. We start by following the proof of Theorem 4.1. Let Φ be a diagonal matrix with elements $(\sqrt{v_i} : i = -1, 0, \ldots, n+1)$ then since $(\Phi Z)_i \sim N(0, v_i)$ we see that G has the same distribution as the quadratic form

$$Z^{\mathrm{T}}[I_R\Phi : \eta J_R]^{\mathrm{T}}[I_S\Phi : \eta J_S]Z, \quad (4.20)$$

using our previous notation with I_R, J_R, I_S and J_S as given in (4.13).

Continuing to follow the proof in Theorem 4.1, the matrix H becomes in this case

$$H = \begin{pmatrix} \Phi\, I_R^{\mathrm{T}} I_S\, \Phi & \eta\Phi\, I_R^{\mathrm{T}} J_S \\ \eta J_R^{\mathrm{T}} I_S\, \Phi & \eta^2 J_R^{\mathrm{T}} J_S \end{pmatrix},$$

so that

$$\begin{aligned} \mathbb{E}\, G = \text{trace}H &= \text{ trace } (\Phi I_R^{\mathrm{T}} I_S \Phi) + \eta^2 \text{ trace } (J_R^{\mathrm{T}} J_S) \\ &= \text{ trace } (\Phi I_R^{\mathrm{T}} I_S \Phi) + 0. \end{aligned}$$

Since $I_R^{\mathrm{T}} I_S$ is diagonal with elements $(0, 0, 1, \ldots, 1, 0)$, the product $\Phi I_R^{\mathrm{T}} I_S \Phi$ is diagonal

with elements $(0, 0, v_1, \ldots, v_n, 0$. It follows that

$$\mathbb{E}\, G = \text{ trace } (\Phi I_R^{\mathrm{T}} I_S \Phi) = \sum_{i=1}^{n} v_i = v,$$

so that G is unbiased.

For the variance as in (4.14)

$$\begin{aligned} \text{Var}\, G = \frac{1}{2}||\Phi(I_R^{\mathrm{T}} I_S + I_S^{\mathrm{T}} I_R)\Phi||_2^2 \\ + \eta^2||\Phi(I_R^{\mathrm{T}} J_S + I_S^{\mathrm{T}} J_R)||_2^2 + \frac{1}{2}\eta^4||J_R^{\mathrm{T}} J_S + J_S^{\mathrm{T}} J_R||_2^2 \end{aligned} \tag{4.21}$$

To go further we need to introduce some notation, defining $\delta[q]$ to be an $n \times (n+3)$ matrix with unit elements at $(i, q+i-1), i = 1, \ldots, n$ and zero elements elsewhere. In other words, $\delta[p]$ is a Toeplitz matrix which has a first row with 1 in column p and zero elsewhere.

With this notation

$$\begin{aligned} I_R &= \delta[3] \\ J_R &= \delta[3] - \delta[2] \\ I_S &= \delta[2] + \delta[3] + \delta[4] \\ J_S &= \delta[4] - \delta[1]. \end{aligned} \tag{4.22}$$

Similarly, we define $\delta[p, q]$ to be an $(n+3) \times (n+3)$ matrix with unit elements at $(p+i-1, q+i-1), i = 1, \ldots, n$ and zero elements elsewhere.

Since both $\delta[q]$ and $\delta[p, q]$ are essentially $(n \times n)$ identity matrices embedded in a matrix

of zeroes, it is straightforward to show

$$\delta[p]^{\mathrm{T}}\delta[q] = \delta[p,q].$$

The matrix products in 4.21 can then be written as

$$I_R^{\mathrm{T}}J_S + I_S^{\mathrm{T}}J_R = \delta[2,3] + \delta[3,3] + \delta[3,4] + \delta[4,3] - \delta[2,2] - \delta[3,2] - \delta[3,1] - \delta[4,2]$$
$$I_R^{\mathrm{T}}I_S + I_S^{\mathrm{T}}I_R = \delta[2,3] + \delta[3,2] + \delta[3,4] + \delta[4,3] + 2\delta[3,3]$$

The combination of these terms has the effect adding and subtracting $(n \times n)$ identity matrices at various locations in an $(n+3) \times (n+3)$ matrix of zeroes. For illustration $(n = 7)$,

$$I_R^{\mathrm{T}}I_S + I_S^{\mathrm{T}}I_R = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 2 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{4.23}$$

and

$$I_R^{\mathrm{T}} J_S + I_S^{\mathrm{T}} J_R = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \end{pmatrix} \tag{4.24}$$

Returning to (4.21), multiplying by the diagonal matrix Φ and summing the squares we have

$$||\Phi(I_R^{\mathrm{T}} I_S + I_S^{\mathrm{T}} I_R)\Phi||_2^2 = 2(v_1 v_0 + v_n v_{n+1}) + 8\sum_{i=1}^{n-1} v_i v_{i+1} + 4\sum_{i=1}^{n} v_i^2$$

$$||\Phi(I_R^{\mathrm{T}} J_S + I_S^{\mathrm{T}} J_R)||_2^2 = 2(v_0 + v_{n+1}) + 6(v_1 + v_n) + 8\sum_{i=1}^{n-1} v_i$$

and since we already know from Theorem 4.1 that

$$||J_R^{\mathrm{T}} J_S + J_S^{\mathrm{T}} J_R||_2^2 = 16n - 12$$

the result follows. □

4.2.4 Symbolic algebra

The proof of Theorem 4.3 is straightforward since the estimator has a relatively simple form. However, there is a requirement to use more complicated estimators, to bring multi-tick time periods into variance analysis. Multi-tick estimators then overlap in complicated ways. Take for instance the general estimator:

$$G := \sum_{i=1}^{n} R_{i,k} R_{i+j,k+2j} \tag{4.25}$$

We have a k-tick lag and a j-length-shift for correlation. In the time dependent inhomogenous problem for this estimator, the formulae will become prohibitively difficult to write down. What we can do is show how the general estimators decompose into polynomials over 'fundamental terms' and therefore how the variance of general estimators decomposes into linear combinations of expectations of fundamental terms ('fundamental expectations'). At this level of recursion, we show that the general formulae are amenable to treatment in symbolic algebra language. To that end, we have written a general computer package.

We have written the software in the object-oriented language C++. The programme starts by defining a short suite of C++ classes which encapsulate monomial and polynomial algebra. A user can define a polynomial over multiple variables, multiply, add and subtract them, etc. Essentially the user can do algebra. Now for a fixed n we can see any G as a polynomial over 'symbols' for $B_i = X_i - X_{i-1}$ and ϵ_i. Our routine treats G (and potentially other estimators) in this way - as stochastic polynomials.

We then add functionality for taking 'expectation' of such stochastic polynomials. The expectation of a stochastic polynomial is a polynomial in a symbolic algebra consisting

of symbols for the time dependent volatilities v_i. By writing the polynomial G in terms of B_i's and ϵ_i's, taking expectation of such a stochastic polynomial *commutes* to taking products of expectations of the powers in the monomials, and this then reduces to Gaussian moments. So we have a symbolic algebra of stochastic polynomials closed under arithmetic and taking of expectations. This means that

$$\mathrm{Var}[G] = \mathrm{E}[G^2] - \mathrm{E}[G]^2 \tag{4.26}$$

is just a polynomial to be displayed. The asymptotic growth at n is now the difference polynomial $\mathrm{Var}[G]_{n+1} - \mathrm{Var}[G]_n$, another polynomial in our algebra.

We then check our computer implementation versus Zhou's original problem and the time-dependent inhomogenous Zhou problem. We have calculated inhomogenous Zhou in two different ways - using our Toeplitz technology and also above, in the spirit of Zhou, by direct computation. We can confirm that our symbolic stochastic algebra programme reproduces all the asymptotic formulae we have computed by hand.

Having done all this work, is there any value in studying inhomogenous Zhou? This is not obvious. What should the model represent? Is the model trying to capture a long time span such as a day, suggesting that there are periods of the day where due to volume we are in one volatility regime, and at other times in another regime? We do not think this is an appropriate model specification to capture that feature of trading. We would much prefer to identify the different periods of the day and calibrate different homogenous Zhou models for each. Or are we suggesting that at the appropriate atomic period for application of the model, say even a few minutes or seconds, a deterministic cycle of volatility is inherent? The market lunges and then retreats, and this pattern is fundamental - there is a short trading cycle as opposed to a short trading period? This is a very interesting

contention and inhomogenous Zhou could test it.

4.2.5 An infill consistent estimator

In the intervening years since the appearance of Zhou's paper in 1996 there has been a substantial effort to find so-called *infill consistent* volatility estimators under model (4.4), i.e. estimators that converge in probability to the underlying volatility as n increases or equivalently as the interval between observations diminishes. The papers of Zhang et al. (2005), Zhang (2006), Barndorff-Nielsen et al. (2008) and Jacod et al. (2009) are important landmarks. See also Bauwens et al. (2012) for further background.

Here we will show that with a small modification, Zhou's second estimate can be made infill consistent. This possibility was conjectured by Hansen and Lunde (2006), however no consideration was given to the bias in (4.16) and its impact on the mean squared error of the estimator. Our proposed modification removes the bias and has the following form:

$$G_k^* = \frac{1}{k} \sum_{i=1}^{n+k-1} R_{i+1,k+2} R_{i,k}^*, \tag{4.27}$$

where $R_{i,k} = X_i - X_{i-k}$ and

$$R_{i,k}^* = \begin{cases} X_i - X_0 & \text{if } i = 1, \ldots, k \\ X_i - X_{i-k} & \text{if } i = k+1, \ldots, n \\ X_n - X_{i-k} & \text{if } i = n+1, \ldots, n+k-1. \end{cases}$$

To verify that G_k^* is unbiased we need to show

$$\mathbb{E}\, R_{i+1,k+2}R_{i,k}^* = \begin{cases} \sum_{j=1}^{i} v_j & \text{if } i = 1, \ldots, k \\ \sum_{j=i-k+1}^{i} v_j & \text{if } i = k+1, \ldots, n \\ \sum_{j=i-k+1}^{n} v_j & \text{if } i = n+1, \ldots, n+k-1. \end{cases}$$

For example, when $1 \leqslant i \leqslant k$, we have

$$\begin{aligned} R_{i,k}^* &= \breve{X}_i - \breve{X}_0 + \epsilon_i - \epsilon_0 \\ R_{i+1,k+2} &= \breve{X}_{i+1} - \breve{X}_i + \breve{X}_i - \breve{X}_0 + \breve{X}_0 - \breve{X}_{i-k-1} + \epsilon_{i+1} - \epsilon_{i-k-1}, \end{aligned}$$

and since the error terms and the increments in $\breve{X}$ are independent

$$\mathbb{E}\, R_{i+1,k+2}R_{i,k}^* = \mathbb{E}\,(\breve{X}_i - \breve{X}_0)^2 = \sum_{j=1}^{i} v_j.$$

The other cases follow similarly so that

$$\mathbb{E} \sum_{i=1}^{n+k-1} R_{i+1,k+2}R_{i,k}^* = k \sum_{i=1}^{n} v_i = k \int_0^1 \sigma^2(s)ds, \tag{4.28}$$

as required.

We now prove two lemmas.

Lemma 4.1. *Let $\breve{X}_i - \breve{X}_{i-1} \sim N(0, \alpha^2)$ and $\epsilon_i \sim N(0, \eta^2)$, independently $i \in \mathbb{Z}$, then*

$$\begin{aligned} \text{Var}\,(G_k^* | \alpha^2, \eta) = \frac{\alpha^4}{6k}\left[n(8k^2 + 24k + 4) - (k+2)(k+1)^2\right] \\ + \frac{4\alpha^2\eta^2}{3k}\left[6n + (k+4)(k-1)\right] + \frac{(8n - 8 + 2k)\eta^4}{k^2}. \end{aligned} \tag{4.29}$$

Proof. As in Theorem 4.1 the estimator G_k^* has the representation

$$Z^{\mathrm{T}}[\alpha I_R : \eta J_R]^{\mathrm{T}}[\alpha I_S : \eta J_S]Z,$$

where now I_R, J_R, I_S and J_S are $(n+k-1) \times (n+2k+1)$ dimensional matrices. I_S and J_S are Toeplitz matrices with first rows

$$\text{Row 1: } I_S = (0, 1^{k+2}, 0^{n+k-2})$$
$$\text{Row 1: } J_S = (-1, 0^{k+1}, 1, 0^{n+k-2}),$$

and all other elements in their first columns equal to zero. The matrices I_R and J_R have partitions

$$I_R = (\mathbf{0}^{(n+k-1)\times(k+1)} : B : \mathbf{0}^{(n+k-1)\times k})$$
$$J_r = (\mathbf{0}^{(n+k-1)\times(k+1)} : C : D : \mathbf{0}^{(n+k-1)\times k}),$$

where $\mathbf{0}^{r\times c}$ is an $r \times c$ dimensional matrix of zeros, C is a column vector with $C_i = -1, i = 1, \ldots, k$ and $C_i = 0$ otherwise and B, D are Toeplitz matrices with first columns

$$\text{Column 1: } B = (1^k, 0^{n-1})^{\mathrm{T}}$$
$$\text{Column 1: } D = (1, 0^{k-1}, -1, 0^{n-2})^{\mathrm{T}},$$

and all other first row elements zero. We find by induction

$$||I_R^{\mathrm{T}} I_S + I_S^{\mathrm{T}} I_R||_2^2 = \frac{k}{3}\left[n(8k^2+24k+4) - (k+2)(k+1)^2\right]$$
$$||I_R^{\mathrm{T}} J_S + I_S^{\mathrm{T}} J_R||_2^2 = \frac{4k}{3}[6n + (k+4)(k-1)]$$
$$||J_R^{\mathrm{T}} J_S + J_S^{\mathrm{T}} J_R||_2^2 = (16n - 16 + 4k),$$

and substituting in (4.14) as before we have the variance of G_k^* as claimed. □

Lemma 4.2. *Let* $\breve{X}_i - \breve{X}_{i-1} \sim N(0, v_i)$ *and* $\epsilon_i \sim N(0, \eta^2)$*, independently* $i \in \mathbb{Z}$ *and let* $\mathrm{Var}\,(G_k^*|\{v_i\}, \eta)$ *be the variance of* G_k^* *under these conditions. Suppose* $v_i \leqslant \alpha^2, i \in \mathbb{Z}$*, then*

$$\mathrm{Var}\,(G_k^*|\{v_i\}, \eta) \leqslant \mathrm{Var}\,(G_k^*|\alpha^2, \eta), \tag{4.30}$$

where $\mathrm{Var}\,(G_k^*|\alpha^2, \eta)$ *is given in* (4.29)*.*

Proof. As in Theorem 4.3 the variance of G_k^* with non-constant $\{v_i\}$ is given by an expression of the form (4.21) with each of the three component terms representing the sum of squares of elements of the associated matrix. Pre- or post multiplication by the diagonal matrix $\Phi = \mathrm{diag}\{\sqrt{v_i}\}$, introduces a multiplicative factor for each element. In the constant variance case, the diagonal matrix is simply $\Phi_0 = \mathrm{diag}\{\alpha\}$. Since $v_i \leqslant \alpha^2$, it follows that the square of each element when multiplying by Φ is no larger than when multiplying by Φ_0. □

We are now in a position to construct a consistent estimator of the underlying volatility when the density of observation points tends to infinity.

Theorem 4.4. *Let* S *be the interval* $(-1, 2)$ *and for each* n*, let* $\mathcal{T}^{(n)} = (t_i^{(n)} \in S)$ *be an increasing sequence of observation times with* $t_0^{(n)} = 0$ *and* $t_n^{(n)} = 1$*. Suppose* $\sigma(s), s \in S$

is bounded by $\sigma_{\max}$ *and* $|t_i^{(n)} - t_{i-1}^{(n)}| < A/n, t_i^{(n)} \in S$*, then under Model* (4.4) *the estimator* $G^*_{k(n)}(n)$ *is infill consistent for* $v = \int_0^1 \sigma^2(s)ds$*, when* $k(n) \sim n^{2/3}$ *as* $n \to \infty$*.*

Proof. Since $\sigma(s) \leqslant \sigma_{\max}$, and $|t_i^{(n)} - t_{i-1}^{(n)}| < A/n$,

$$v_{i,n} = \int_{t_{i-1}^{(n)}}^{t_i^{(n)}} \sigma^2(u)du \leqslant \frac{A\sigma^2_{\max}}{n}.$$

From Lemma 4.2, it follows that the variance of $G^*_k(n)$ is no larger than (4.29) with $A\sigma^2_{\max}/n$ substituted for α^2. After this substitution, we have

$$\begin{aligned}\text{Var}\,(G^*_k|\{v_i\},\eta) \leqslant \frac{A^2\sigma^4_{\max}}{6n^2k}\left[n(8k^2+24k+4) - (k+2)(k+1)^2\right] \\ + \frac{4A\sigma^2_{\max}\eta^2}{3kn}\left[6n + (k+4)(k-1)\right] + \frac{(8n-8+2k)\eta^4}{k^2} \quad (4.31)\end{aligned}$$

For large values of $k < n$, the dominant terms are

$$\frac{8kA^2\sigma^4_{\max}}{6n} \quad \text{and} \quad \frac{8n\eta^4}{k^2},$$

from the first and last componenents, respectively. Taking $k(n) = n^{2/3}$, both terms converge to zero as $n \to \infty$. It follows that the variance converges to zero and is of order $n^{-1/3}$ as $n \to \infty$. Since the estimator is unbiased the estimator is infill consistent, by Chebyshev's inequality. □

We have demonstrated infill consistency with conditions on σ and the observation time sequence $\{t_i^{(n)}\}$. If $\sigma(s)$ is continuous then it is necessarily bounded on $(-1, 2)$. The time sequence constraint involves the constant A which can be fixed at an arbitrarily high level. It is also possible to require only $|t_i^{(n)} - t_{i-1}^{(n)}| < A/n^\gamma$, where $3/4 < \gamma < 1$. Consequently

the constraint is not overly restrictive.

4.2.6 Modern developments

Some of the more recent research on volatility estimation can be understood most readily with reference to the estimated variogram. For illustration, Figure 4.2 shows a variogram simulated from model (4.4). The observations are assumed to be made at regularly spaced

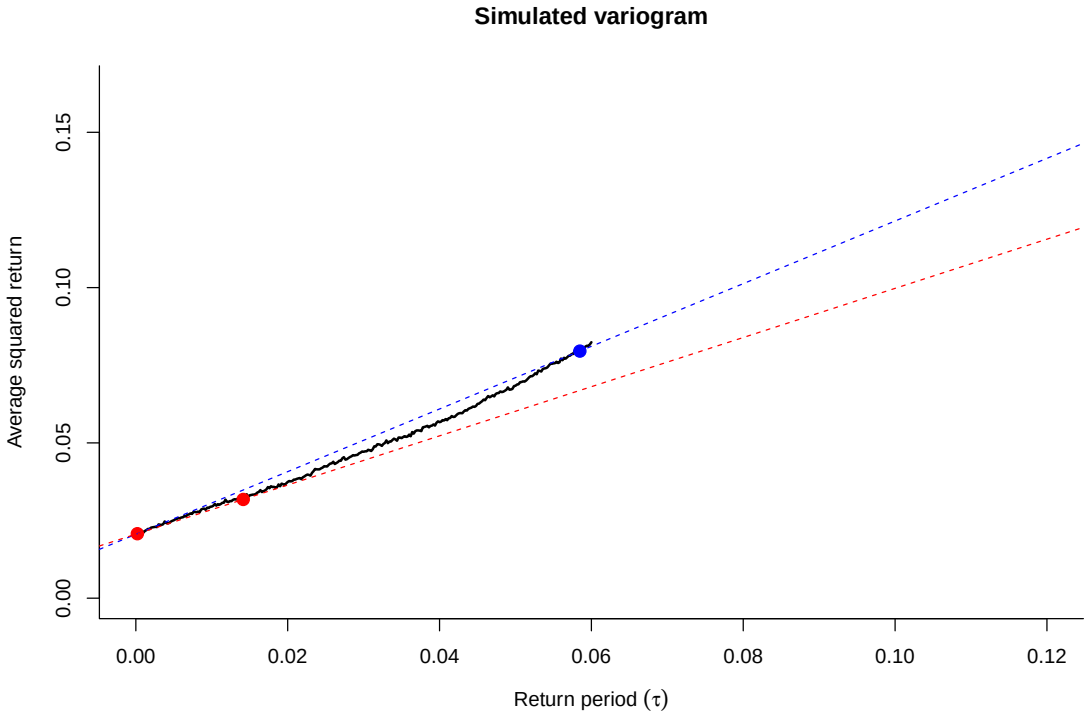


Figure 4.2: Simulated variogram from model (4.4) with $\eta = 0.1$, $\sigma = 1$, $n = 5000$ and $\tau = k/n, k = 1, \ldots, 300$. Points are marked in red at $k = 1, n^{1/2}$ and in blue at $n^{2/3}$. Extrapolated values using the estimators of Zhang et al. (2005) and Zhang (2006) are indicated by the dashed lines in blue and red, respectively.

times $t_i = i/n, i = 0, \ldots n$. The objective is to estimate $v = \int_0^1 \sigma^2(s)ds = \mathbb{E}\left[\breve{X}_n - \breve{X}_0\right]^2$.

The variogram estimator is

$$\hat{\Gamma}(k/n) = \frac{\sum_{i=k}^{n}[X(t_i) - X(t_{i-k})]^2}{n-k+1}, \quad k = 1, \ldots, n.$$

From (4.6) under the model (4.4) we expect to see an approximately linear plot, with slope v. To estimate the slope, two points can be chosen, say $1/n$ and k/n and the slope estimated by

$$\hat{V} = \frac{\hat{\Gamma}(k/n) - \hat{\Gamma}(1/n)}{k/n - 1/n}.$$

One of the ways explaining, other methods of estimating volatility is trying to reproduce this line which increases as return period increases to estimate the slope of the line. One simple way to estimate the slope is to choose two points on the line on the variogram plot, to estimate the slope from this two points, (two red dots). This is essentially the two scale estimator considered by Zhang et al. (2005). The estimator is biased but the authors show that it is infill consistent when $k = k(n)$ is chosen to grow as $n^{2/3}$, in which case the MSE (mean square error) of the estimator is of order $n^{-1/3}$. This is the same order of convergence as the estimator in Section 4.2.5.

Zhang (2006) then goes on to consider an estimator that combines all of the values of $\hat{\Gamma}(k/n)$ up to $k \sim n^{1/2}$, effectively fitting a straight line through these values. She shows that the MSE of her 'multiscale estimator' is or order $n^{-1/2}$. In an other word, Zhang chooses several points along this variogram and combines them in the way to fit a straight line through the variogram plot and uses the slope of the fitted line as the volatility estimate.

The estimator proposed by Jacod et al. (2009) can also be thought of as a variogram estimator, where the process X is replaced by a moving average of X over k successive time points and the variogram is computed with these average values. They show that by

choosing k of order $n^{1/2}$ the MSE of their estimator is of order $n^{-1/2}$. This rate can be shown to be optimal when the volatility is constant.

4.2.7 Kernel methods

In a series of papers Barndorff-Nielsen et al. (2004, 2008, 2009) have developed a range of kernel-based volatility estimators. These estimators are weighted linear combinations of the empirical autocovariances of the returns. They are directly related to the estimators of Zhang et al. (2005), Zhang (2006) and Zhou (1996). For example, Zhou's first estimator (4.7) can be written as the linear combination $G = \hat{\gamma}_0 + \hat{\gamma}_1 + \hat{\gamma}_{-1}$ where

$$\hat{\gamma}_h = \sum_{i=1}^{n} R_{i,1} R_{i+h,1}, \quad h \in \mathbb{Z}.$$

Kernel based methods were originally developed in the classical problem of estimating the variance of X from the observed values of a stationary sequence $X_1, \ldots, X_n$ (Newey and West, 1987). Barndorff-Nielsen et al. show that the asymptotic behaviour of simple kernel-based volatility estimator is strongly influenced by "edge effects". They show how to select kernel weightings to achieve the optimal order of infill asymptotics. The kernel weightings are applied to the autocovariances to form linear combinations of the autocovariances. There is a connection to spectral density estimation and Fourier methods via the representation of the spectrum in terms of autocovariances.

4.3 Observed price – activity time

The modelling of time-varying volatility is important when the data are limited to prices at an arbitrary set of discrete times. For exchange traded assets, other measures of activity are known to be tied strongly to price fluctuations, as described previously in Section 3.1.3. There is now substantial evidence that volatility can be stabilised when measured on an 'activity' time scale (Clark, 1973; Ané and Geman, 2000; Griffin and Oomen, 2008). Transaction time and 'tick' time are common choices. Alternative measures of 'activity' time are closely guarded intellectual property in the world of high-frequency algorithmic trading.

The principal advantage of working on a volatility stabilised time scale is that simple recursive methods with constant coefficients based on the Kalman filter can be developed. Coefficients are determined by maximising empirical profitability in simulation, real-time application, or with historical data.

Measuring volatility on a suitably modified time-scale provides a simple way of dealing with time-varying volatility. Switching clocks is a fundamental idea in finance. The paradigm in the market is event-based time. As modifications of chronological time, a clock should run as the same speed as trading activity that might be taking account of volume traded or the actual activity on the order book. The standard inhomogeneous log-price model in physical time,

$$d\breve{X}(t) = \sigma(t)dB(t),$$

then has $\sigma^2(t) = \psi(t)\beta^2$, where $\psi(t)$ is the observed activity intensity at time t and β^2 is a constant factor be estimated.

If we let $s = \Psi(t) = \int_0^t \psi(u)du$ be the new time-scale and $\breve{Y}(s) = \breve{X}(t)$, then

$$d\breve{Y}(s) = \beta dB(s),$$

so the log-price process is transformed into Brownian motion with constant infinitesimal variance. The integrated variance over a period of time $(0, T)$ is then

$$\int_0^1 \sigma^2(s)ds = \beta^2 \int_0^1 \psi(s)ds = \beta^2\Psi(1),$$

and as a result the assessment of integrated variance is reduced to the estimation of the constant parameter β^2.

There is particular simplification when the activity clock is simply a count of activity events. Observation times on this clock are effectively successive integers. If the clock successfully stabilises the volatility, the returns of $\breve{X}_i$ have constant variance α^2 and an estimate $\hat{\alpha}^2$ can be obtained by Zhou's subsampling method (4.15). The estimated integrated variance over the physical time interval $(0, 1)$ is then $\hat{\alpha}^2$ multiplied by the number of activity events in $(0, 1)$.

4.3.1 Maximum likelihood and Kalman filter

With a volatility stabilising clock, observations form a homogeneous ARIMA sequence, and the sequence of returns is MA[1]. In this framework, the problem simplifies to finding the MLEs of the parameters of a simple moving average sequence for which the Kalman filter is well adapted. In this formulation the asymptotic variance of the MLE of α^2 is known to be of order $n^{-1/2}$, see Aït-Sahalia et al. (2005). The convergence rate is very slow – and it does not go to 0 rapidly as we would think. It is the noise that causes the

problem. Nevertheless, Kalman filter allows us to write down the likelihood, to explore the likelihood space to maximize it.

Other authors have suggested *pre-whitening* the sequence of returns by representing them in the form

$$R_i = X_i - X_{i-1} = \breve{X}_i - \breve{X}_{i-1} + \epsilon_i - \epsilon_{i-1} = W_i - \theta W_{i-1},$$

where (W_i) are iid $N(0, v_W)$ and the quantities θ and v_W^2 are given by the equations

$$\alpha^2 + 2\eta^2 = (1 - \theta^2)v_W^2$$
$$\eta^2 = \theta v_W^2.$$

The *pre-whitened* (W_i) are then recovered by applying an EWMA to the sequence (R_i). This approach is investigated by Corsi et al. (2001) who provide numerical comparisons of various volatility estimates based on these ideas.

4.4 Model criticism

In this section we develop a Monte Carlo goodness-of-fit test for the standard model (4.4). The test is aimed specifically at the assumptions underlying the volatility estimators in Section 4.2.2. We apply the test in an empirical study of EBS data on the Euro/USdollar exchange rate between August 2008 and October 2009.

The standard model (4.4) has been criticised by various authors. Hansen and Lunde (2006) provide a very thorough analysis of the model assumptions, illustrating with data from 30 stocks in the Dow Jones Industrial Average. Robustness to model specification

is emphasised by Jacod et al. (2009) in their construction of pre-averaging estimators.

Our test is based on the observation that under (4.4) with fixed $k \geqslant 1$ and $j \geqslant 1$, the collection of disjoint returns

$$W_i = R_{ik+(i-1)j,k}, \quad i = 1, \ldots, N, \tag{4.32}$$

are independently $N(0, \nu_i^2)$ with

$$\nu_i^2 = \int_{\tau_{i-1}}^{\tau_i} \sigma^2(s)ds + 2\eta^2$$
$$\tau_i = t_{ik+(i-1)j}, \quad i = 1, \ldots, N,$$

and $N = \lfloor (n+j)/(k+j) \rfloor$. It follows that

$$\sum_{i=1}^{p} W_i \overset{\text{distn}}{=} B\left(\sum_{i=1}^{p} \nu_i^2\right), \quad p = 1, \ldots, N,$$

where $B(t), t > 0$ is a standard Brownian process.

In general, if a process has independent normally distributed increments then by changing the time scale, it ends up as a Brownian motion observed at a sequence of times. Volatility is not constant, it changes, and increments have different variances. That is the assumption in this model. Therefore, if we choose the correct time scale, we can convert this process into Brownian motion and we can test whether that was correct by checking if it is consistent with Brownian motion properties. One of the property is Brownian bridge.

Suppose that

$$S(t) \overset{\text{distn}}{=} B(\Psi(t)), \quad 0 < t < T,$$

for an increasing function Ψ then, by a time transformation,

$$\tilde{B}(u) = \frac{S(\Psi^{-1}[u\Psi(T)]) - uS(T)}{\sqrt{\Psi(T)}}, \quad 0 < u < 1$$

is a Brownian bridge. To test whether Ψ is the correct time transformation we can consider the maximum value of $|\tilde{B}(u)|$, as in the Kolmogorow-Smirnov test. Note that

$$\max_{0<u<1} \left| \frac{S\{\Psi^{-1}[u\Psi(T)]\} - uS(T)}{\sqrt{\Psi(T)}} \right| = \max_{0<t<T} \left| \frac{S(t) - S(T)\Psi(t)/\Psi(T)}{\sqrt{\Psi(T)}} \right|,$$

so the test rejects when the maximum value of $S(t) - S(T)\Psi(t)/\Psi(T), 0 < t < T$ exceeds some critical value.

For our test, we mimic this approach and use $T^* = \max_p |T_p|$ as a test statistic where

$$T_p = \sum_{i=1}^{p} W_i - \frac{\sum_{i=1}^{p} W_i^2}{\sum_{i=1}^{N} W_i^2} \sum_{i=1}^{N} W_i, \quad p = 1, \ldots, N \tag{4.33}$$

and the terms $\sum_{i=1}^{p} W_i^2$ are used as a proxy for $\sum_{i=1}^{p} \nu_i^2$.

We want to see how often this trajectory deviates from the expected trajectory and whether or not it is too far from where it is supposed to be. We assess the significance of T^* with a Monte Carlo p-value, independently randomising the sign of each $W_p, p = 1, \ldots, N$. The observed value of T^* is then compared with (say) 999 randomised values $T^*(m), m = 1, \ldots, 999$ to obtain the Monte Carlo p-value, in the usual manner – see Barnard (1963).

Application to Euro/USdollar exchange rate

Data for Euro/USdollar trades on EBS for the period August 2008 and October 2009 were made available by William Hooper of Xdirect. The Monte Carlo test was applied

to each full trading day with four (k, j) pairs and 999 Monte Carlo samples in each case. Histograms of the p-values are shown in Figure 4.3.

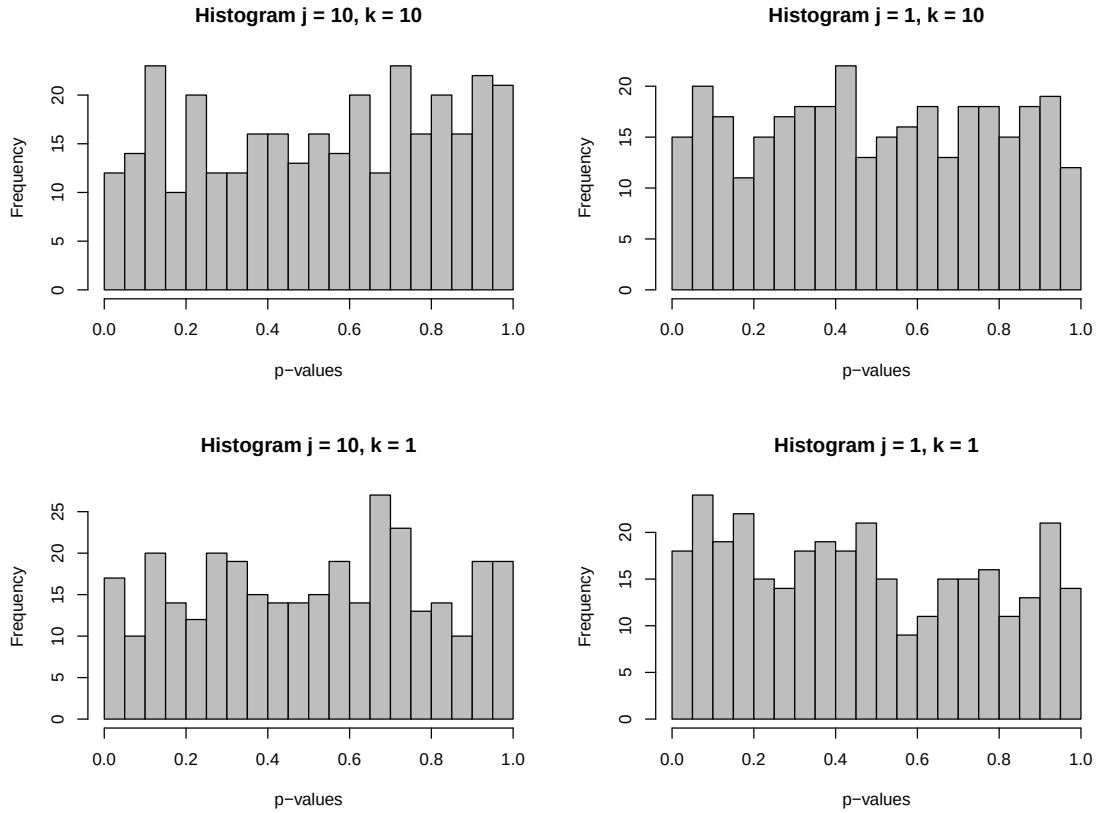


Figure 4.3: Monte Carlo p-values based on the test statistic (4.33) where $\{W_i\}$ are defined in (4.32). The data are tick-by-tick trade prices of Euro/USdollar transactions on the EBS exchange for 328 trading days between August 2008 and October 2009

As can be seen, the p-value distributions for all these (k, j) pairs are relatively flat showing no reason to reject model (4.4). See Hansen and Lunde (2006) and Chapter 6 for further discussion.

4.5 Robustness and volume effects

The observed log-price in high-frequency data is commonly assumed to be a combination of the unobservable theoretically efficient log-price and some noise component ϵ due to the imperfections of the trading and recording process, the so-called *microstructure* of the market. Model (4.4) attempts to capture this structure. Less restrictively, we can think of the ϵ process as a combination of discretisation (i.e. rounding error) and independent random noise. When the bid/ask spread is small, the discretisation component will be dominant. More generally, the noise term will capture effects due to bid/ask bounce, targeted price reaction and trade and quote volume. Jacod et al. (2009) argue convincingly that their pre-averaging approach using smoothed values of the log-price data will provide estimators that are robust to model misspecification. Here we will see that stability can also be achieved by using prices derived from deeper levels of the order book.

4.5.1 Volume effects

In this section, we will investigate the daily pattern of volatility in depth-dependent pricing of the US Dollar – Japanese Yen exchange rate, making use of the estimator developed in Section 4.2.5.

The data sets are the quote ladders in the EBS order book for USD–JPY foreign exchange. Observations are tick-by-tick with quotes every 100 milliseconds.

As a first step, the *regular prices* at each time, were determined for various depth of volume: 2, 5 and 10 million US dollars. For example the regular ask price at 2 million dollars is the price in Yen per Dollar that is required to buy that amount of US currency. The regular mid-price is then calculated as the geometric mean of the regular bid and ask

prices. By taking logarithms, these become $\{X_i\}$, the data to be processed.

The volatility was then estimated for each hour of each day in each month, using all tick-by-tick data. The results were then averaged to produce the plots in Figure 4.4.

From the three figures we see immediately that volume-dependent volatility is highest in the concluding hours of each day, when trading is light. In such periods market makers are less active which has the effect of increasing the spread and as a result regular prices have to be found deeper into the order book. Furthermore, there is a visible bump in each curve at around 1:30 PM each day, corresponding to the time at which important economic news items are usually announced. This underlines the importance of identifying and taking account of points in time at which there are substantial price movements, predictable or otherwise.

We observe that there is a consistent reduction in the volatility estimates when prices are calculated using an increasing proportion of the order book. Interestingly we also find that the size of η, a measure of the microstructure effect, is substantially smaller with these volume-dependent estimates of volatility compared with values calculated using the simple mid-price.

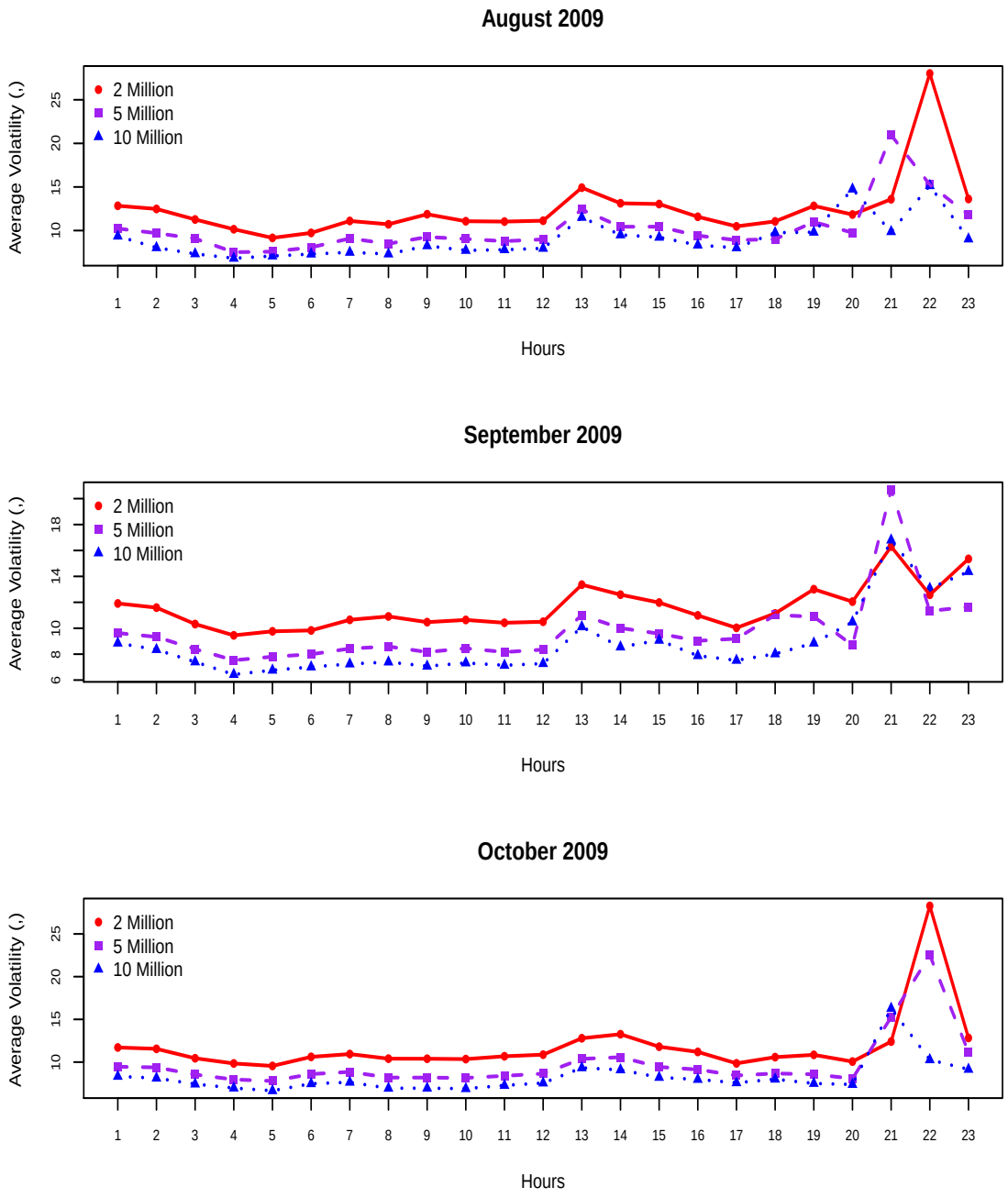


Figure 4.4: USDJPY Hourly Average Volatility in August, September and October 2009 by using regular prices constructed from quote-by-quote prices of different volumes.

Chapter 5

Covolatility

Up to this stage, we have focussed on the volatility of a single series. We now turn to the task of estimating the correlation between two financial time series. Volatility estimation plays an essential role since the volatilities of the component series appear in the denominator of the correlation calculation. However, estimation of the numerator presents a new problem because log-prices for the components are typically observed asynchronously, for example when using log-prices at the times of trades.

Not only did Zhou (1996) suggest how to estimate volatility, but remarkably he also suggested how to estimate covolatility with asynchronous data (Zhou, 1995). That is important to note because much later, ten years later, Hayashi and Yoshida (HY) also came up with a way of estimating covolatility. It turns out that Zhou had discovered the HY estimator before HY did. One of our contributions in this chapter is to give recognition to Zhou for his discovery before HY. We improve the HY estimate and come up with our own correlation estimate. We show that this estimator has superior performance to the Zhou-Hayashi-Yoshida estimate.

The models in Chapter 4 can be extended to multivariate asset processes in a straightforward manner. However there are complications in calculating cross-correlations between assets when the assets are traded asynchronously. Nevertheless under certain conditions it is possible to estimate cross-correlations efficiently. We review methods that have been proposed over the last 20 years and identify key contributions. We pay particular attention to a neglected paper by Zhou (1995) and show that it contains essential ideas that generalise and pre-date more recent work. Our search for recursively defined estimators of covolatility that can be used in real-time high-frequency trading lead us to further generalisations of Zhou's estimator. Properties of the estimator are discussed and its variance is derived when working in tick time with constant parameters. In a simulation study the estimator is shown to outperform the popular covolatility estimator of Hayashi and Yoshida (2005). The chapter concludes with a discussion of the effectiveness of quasi-MLE methods with optimisation based on the Kalman filter and the use of an EM algorithm for asynchronous episodes.

5.1 Background

The analysis of multivariate data is a central concern of financial management, not least in portfolio design where the objective is to exploit correlations between price movements of portfolio components to minimise risk. In high-frequency trading, short term correlations between assets can be exploited in the design of fast aggressive trading strategies and in the rapid defence of market making positions.

In the theoretical financial literature, joint movements of log-returns of multiple assets are often modelled by correlated Brownian motion. Under the assumption that all of the assets are priced synchronously (in practice an unrealistic requirement) correlations

and covariances can then be recovered from their realised versions. The procedures are described in Barndorff-Nielsen and Shephard (2004a) along with the derivation of various limit theorems for the associated estimators.

There are various ways of dealing with asynchronous data. The simplest is to impose a common grid of 'observation' times for all the assets involved and then to interpolate prices for each of the assets at these grid points. In an early study, Epps (1979) showed that correlations obtained by this method depend on the size of the spacing, h, between the grid points. With the data that he had available, he found that the correlation reduced as $h \to 0$. This has become known as the *Epps effect*. A possible explanation in terms of persistent lag-correlation has been put forward by Tóth and Kertész (2009a,b) but as Zhou (1995) and others have noted, the effect can be explained simply by supposing that observations of the efficient log-price are contaminated by independent random noise. As the sampling frequency increases, i.e. as $h \to 0$, returns are dominate by uncorrelated noise.

For data derived from the 'fighting screens' described by Zhou where promotional foreign exchange quotes are posted erratically on the Reuters screens, this may be an adequate explanations. For high-frequency exchange-traded assets in a modern trading environment more elaborate models are required, since observations of prices and volumes for quotes and trades are available more or less continuously with microsecond time resolution. Nevertheless considerable theoretical research has been devoted to the 'efficient log-price plus noise' model and the model is usually accepted as a starting point for analysis. In this context, major contributions have been made by Zhang (2011) using subsampling methods, extending similar work on volatility estimation. Barndorff-Nielsen et al. (2011) show that kernel-based methods are closely related. In their work they advocate synchronising the processes at 'renewal points', i.e. the succession of epochs at

which a fresh price has been recorded for each of the assets. The most recent price for each asset is then used in correlation calculations. They argue that the neglected 'stale' prices contribute relatively little to the precision of covariance estimation. Their method has the advantage of providing non-negative definite estimates of the covariance matrix. Their theoretical results and simulation studies are based on a flexible class of SDE based stochastic volatility models. Another recent and intuitively attractive approach, achieving comparable efficiency, is that of *pre-averaging* where prices are smoothed over a rolling window before calculating cross-correlations (Jacod et al., 2009; Podolskij et al., 2009).

Hayashi and Yoshida (2005) and earlier Zhou (1995) take a different approach; the work of de Jong and Nijman (1997) is related. They use asynchronous data directly to obtain unbiased covariance estimators by summing cross-products of overlapping returns. Zhou shows that the estimator remains unbiased when there is observation noise and introduces a subsampling technique to deal with the consequential loss of precision. Zhou argues there is an optimal sampling frequency. He is then able to provide guidance on the optimal choice of sampling frequency. In particular he is able to show infill consistency when observation error is negligible and the sampling frequency increases. The cross-product estimators of Zhou, Hayashi and Yoshida are the main focus of this chapter primarily because their linear structure can be exploited to provide a recursive estimator of covolatility that can be rapidly updated for high-frequency trading. The necessary modifications are described in Section 5.2.4.

5.2 Zhou-Hayashi-Yoshida estimators

Suppose we have two assets A and B. The log-price of A is observed at discrete times $t \in \mathcal{T}$ and the log-price of B at discrete times $u \in \mathcal{U}$. For simplicity, we assume that

a single 'price' is provided at the given time points, for example the logarithm of the geometric mean of the bid and ask prices, or the regular mid-price. The clock can run on physical time or some variant of tick time (TTS), for example the cumulative combined count of ticks for assets A and B. Both $\mathcal{T}$ and $\mathcal{U}$ will usually be irregularly spaced and the intersection $\mathcal{T} \cap \mathcal{U}$ may be empty.

Let $\breve{X}(t)$ and $\breve{Y}(u)$ be the latent efficient log-prices of assets A and B, respectively, at given time points. In their initial analyses Zhou (1995) and Hayashi and Yoshida (2005) assume that the pair of log-prices $(\breve{X}, \breve{Y})$ is bivariate driftless Brownian motion with infinitesimal variances $\sigma_1^2(t)$ and $\sigma_2^2(t)$. In other words

$$\begin{aligned} d\breve{X}(t) &= \sigma_1(t) dB_1(t) \\ d\breve{Y}(t) &= \sigma_2(t) dB_2(t), \end{aligned} \tag{5.1}$$

where B_1, B_2 are correlated Brownian motions with $\langle dB_1, dB_2 \rangle = \rho(t)dt$.

At the first stage Zhou (1995) and Hayashi and Yoshida (2005) assume the log-prices can be observed without noise in an observation window $[0, 1]$ with increasing sequences of observation times $\mathcal{T} = (t_0, \ldots, t_m)$ and $\mathcal{U} = (u_0, \ldots, u_n)$, where $t_0 = u_0 = 0$ and $t_m = u_n = 1$. Observation times are independent of the price process.

Their objective is to estimate

$$c = \int_0^1 \sigma_{12}(t)dt, \quad \text{where} \quad \sigma_{12}(t) = \sigma_1(t)\sigma_2(t)\rho(t).$$

Their estimators are defined as follows.

5.2.1 Definitions

Let

$$\begin{aligned} R_i &= \breve{X}(t_i) - \breve{X}(t_{i-1}),\ i = 1, \ldots, m \\ S_j &= \breve{Y}(u_j) - \breve{Y}(u_{j-1}),\ j = 1, \ldots, n, \end{aligned}$$

be the returns of the latent log-prices of assets A and B, respectively.

Let J_i^X be the ith interval of time $(t_{i-1}, t_i]$ for asset A and let τ_i^X be the length of this return period. Similarly, let $J_j^Y = (u_{j-1}, u_j]$ and τ_j^Y be the length of J_j^Y.

Now define τ_{ij} to be the length of the intersection between J_i^X and J_j^Y, i.e. τ_{ij} is the length of time that the ith return period of asset A and the jth return period of asset B have in common. This length could of course be zero.

Furthermore, let

$$\begin{aligned} t_j^+ &= \min\{t_i, t_i \geqslant u_j\}, \\ t_j^- &= \max\{t_i, t_i \leqslant u_j\}, \end{aligned}$$

and

$$\begin{aligned} u_i^+ &= \min\{u_j, u_j \geqslant t_i\}, \\ u_i^- &= \max\{u_j, u_j \leqslant t_i\}, \end{aligned}$$

so that t_j^+ is the smallest time in $\mathcal{T}$ that is larger than or equal to u_j and t_j^- is the largest time in $\mathcal{T}$ that is smaller than or equal to u_j and similarly for u_i^+ and u_i^+.

Zhou's estimator is

$$\mathrm{Z} = \sum_{j=1}^{n} (\breve{X}(t_j^+) - \breve{X}(t_{j-1}^-))(\breve{Y}(u_j) - \breve{Y}(u_{j-1})). \tag{5.2}$$

Each term in (5.2) is a product of returns. See Figure 5.1.

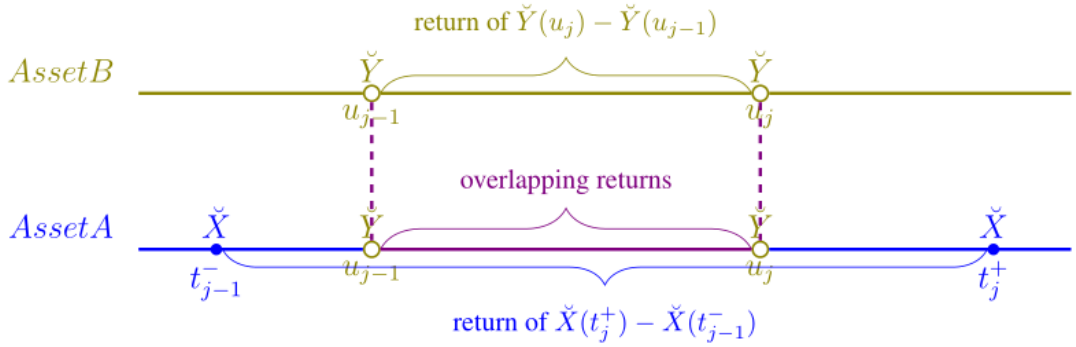


Figure 5.1: Illustration of the contribution of terms in (5.2). The interval (t_{j-1}^-, t_j^+) contains (u_{j-1}, u_j).

Zhou's covariance estimate only involves return periods that *overlap* since $X(t_j^+) - X(t_{j-1}^-)$ can be written as contributions from three *disjoint* intervals I, II, III as shown in Figure 5.2.

$$\breve{X}(t_j^+) - \breve{X}(t_{j-1}^-) = (\breve{X}(t_j^+) - \breve{X}(u_j)) + (\breve{X}(u_j) - \breve{X}(u_{j-1})) + (\breve{X}(u_{j-1}) - \breve{X}(t_{j-1}^-))$$

Zhou notes that his estimator can be expressed alternatively as

$$\sum_{i=1}^{m} (\breve{X}(t_i) - \breve{X}(t_{i-1}))(\breve{Y}(u_i^+) - \breve{Y}(u_i^-)).$$

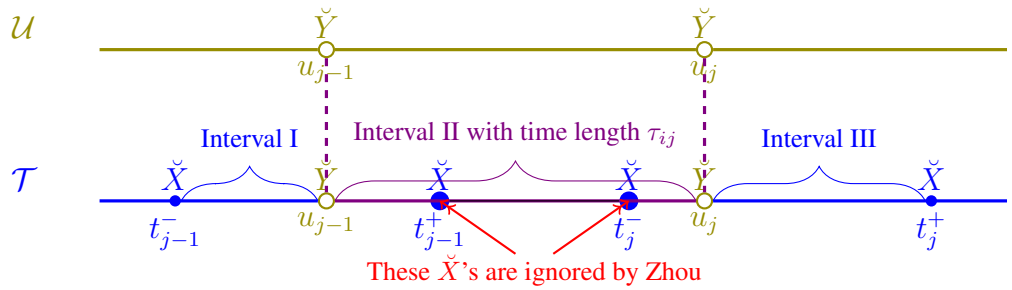


Figure 5.2: Zhou ignores some returns if there were events $\breve{X}$ between the $\breve{Y}$ observations.

The estimator of Hayashi and Yoshida (2005) is

$$\text{HY} = \sum_{i=1}^{m}\sum_{j=1}^{n} R_i S_j \mathbb{I}(\tau_{ij} > 0). \tag{5.3}$$

Hayashi and Yoshida (2005) take *every* return there is in the A process and all the returns in the B process: R_i's and S_j, but only multiply them together if there is an intersection between the return period, i.e. taking the product for every return period of $\breve{X}$ process and the return period of $\breve{Y}$ process, with the indicator 0 if the intersection of the return periods is empty and 1, otherwise. See Figure 5.3

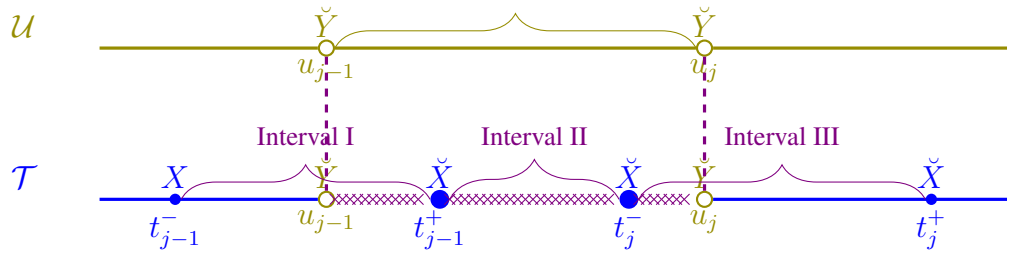


Figure 5.3: Hayashi and Yoshida (2005) covariance estimator approach.

5.2.2 Equivalence of Zhou and Hayashi-Yoshida estimators

We claim the estimators are the same. This is straightforward since (5.2) can be rewritten as

$$\sum_{j=1}^{n}(\breve{X}(t_j^+) - \breve{X}(t_{j-1}^-))S_j \tag{5.4}$$

and

$$\breve{X}(t_j^+) - \breve{X}(t_{j-1}^-) = \sum_{i=1}^{n} R_i \mathbb{I}(\tau_{ij} > 0).$$

There is a similar reduction for the second form of Zhou's estimator. Since the estimators are the same, from now on we will refer to this as the ZHY estimator.

Unbiasedness: Under the above assumptions the ZHY estimator is unbiased for c. This follows from (5.4). Writing $R_j^+ = \breve{X}(t_j^+) - \breve{X}(t_{j-1}^-)$ for notational convenient, we note that S_j is the return of $\breve{Y}$ over the interval $(u_{j-1}, u_j]$ and R_j^+ is the return of $\breve{X}$ over an interval that contains $(u_{j-1}, u_j]$, so that

$$R_j^+ = \left[\breve{X}(t_j^+) - \breve{X}(u_j)\right] + \left[\breve{X}(u_j) - \breve{X}(u_{j-1})\right] + \left[\breve{X}(u_{j-1}) - \breve{X}(t_{j-1}^-)\right].$$

From the basic property of bivariate Brownian motion, the first and last bracketed terms are independent of S_j and

$$\mathbb{E}\left(\breve{X}(u_j) - \breve{X}(u_{j-1})\right)S_j = \mathbb{E}\left(\breve{X}(u_j) - \breve{X}(u_{j-1})\right)(\breve{Y}(u_j) - \breve{Y}(u_{j-1})) = \int_{u_{j-1}}^{u_j} \sigma_{12}(s)ds.$$

The unbiasedness then follows by summing over j. Both Zhou (1995) and Hayashi and Yoshida (2005) show that the ZHY estimator *based on the latent efficient log-price* is infill consistent as the density of the observation points tends to infinity.

5.2.3 Observation noise

The ZHY estimator remains unbiased when subject to independent observation noise (Zhou, 1995), i.e. when

$$\begin{aligned} X(t_i) &= \breve{X}(t_i) + \epsilon_{i,1}, \quad i = 0, \ldots, m \\ Y(u_j) &= \breve{Y}(u_j) + \epsilon_{j,2}, \quad j = 0, \ldots, n \end{aligned} \tag{5.5}$$

where $\breve{X}, \breve{Y}$ are latent bivariate Brownian log-prices and $\{\epsilon_{i,1}\}$,$\{\epsilon_{j,2}\}$ are independent mean zero sequences with variances η_1^2 and η_2^2, respectively.

Although the ZHY estimator is unbiased, its variance is inflated – as shown by Zhou (1995) who provides an approximation to the variance. By extending the techniques used in volatility estimation, Zhang (2011) shows how to modify the ZHY estimator to provide infill consistency.

5.2.4 Activity time – constant parameters

For the remainder of this chapter we assume that the prices of both assets evolve on a common activity clock, such that the parameters σ_1, σ_2 and σ_{12} are constant. A simple example could be the clock that counts the combined cumulative number of ticks of both assets. In practice even with the best chosen clock, there will still be slow changes in the parameters and for this reason volatilities and covolatility are estimated in a moving window. As previously explained, there are speed advantages in expressing this as a recursive calculation.

Since the ZHY estimator is linear in the time contributions, it lends itself to this type of

modification. For example, with the formulation (5.2), ZHY is replaced by

$$\text{ZHY}_m = \sum_{i=1}^{m} \phi^{m-i}(X(t_i) - X(t_{i-1}))(Y(u_i^+) - Y(u_i^-)), \quad 0 < \phi < 1,$$

so that ZHY_m can be updated with the recursion

$$\text{ZHY}_{m+1} = \phi\text{ZHY}_m + (X(t_{m+1}) - X(t_m))(Y(u_{m+1}^+) - Y(u_{m+1}^-)).$$

In practice, traders will experiment with various values of ϕ to maximise the profitability of algorithms that depend on their covolatility estimator.

However, as we shall see the ZHY estimator is not the only linear estimator of σ_{12} when this parameter is assumed to be constant. In the next section we introduce a general class of linear estimators and investigate their properties.

5.3 Extended ZHY estimator

We now propose an extension of the ZHY estimator in the form

$$\sum_{i=1}^{m}\sum_{j=1}^{n} R_i S_j w_{ij},$$

where we introduce weighting factors $\{w_{ij}\}$ that are zero for non-intersecting return periods J_i^X and J_j^Y.

Note that the sum of products can be progressively updated as new data arrive, either in a moving window or recursively as described in Section 5.2.4 – an important consideration in online processing.

Since

$$\mathbb{E}\left(\sum_{i=1}^{m}\sum_{j=1}^{n} R_i S_j w_{ij}\right) = \sigma_{12}\sum_{i=1}^{m}\sum_{j=1}^{n}\tau_{ij}w_{ij}$$

it follows that the extended ZHY estimator

$$\hat{\sigma}_{12} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} R_i S_j w_{ij}\mathbb{I}(\tau_{ij}>0)}{\sum_{i=1}^{m}\sum_{j=1}^{n}\tau_{ij}w_{ij}} \tag{5.6}$$

is unbiased for σ_{12} (assumed constant).

The ZHY estimator is a special case where $w_{ij} = \mathbb{I}(\tau_{ij} > 0)$.

5.3.1 Variance of the estimator

Theorem 5.1. *For given observation structure $\tau = \{\tau_{ij}\}$, under the assumptions of Section 5.2 with constant σ_1,σ_2 and σ_{12}, the extended* ZHY *estimator has variance*

$$\text{Var}\,(\hat{\sigma}_{12}|\tau) = \sigma_1^2\sigma_2^2 A_1 + \sigma_{12}^2 A_2,$$

where

$$\begin{aligned} A_1 &= \left[\textstyle\sum_{ij} w_{ij}^2\tau_i^X\tau_j^Y\right] \Big/ \left(\textstyle\sum_{ij}\tau_{ij}w_{ij}\right)^2 \\ A_2 &= \left[\textstyle\sum_i(\sum_j w_{ij}\tau_{ij})^2 + \sum_j(\sum_i w_{ij}\tau_{ij})^2 - \sum_{ij} w_{ij}^2\tau_{ij}^2\right] \Big/ \left(\textstyle\sum_{ij}\tau_{ij}w_{ij}\right)^2. \end{aligned} \tag{5.7}$$

Proof. The variance of $\sum_{i=1}^{m}\sum_{j=1}^{n} R_i S_j w_{ij}$ can be written as

$$\sum_{ij} w_{ij}^2 \text{Var}\,(R_i S_j) + \sum_{ij\neq kl} w_{ij}w_{kl}\text{Cov}\,(R_i S_j, R_k S_l), \tag{5.8}$$

and since

$$\begin{pmatrix} R_i \\ S_j \\ R_k \\ S_l \end{pmatrix} \sim N \left(0, \begin{pmatrix} \sigma_1^2\tau_i^X & \sigma_{12}\tau_{ij} & 0 & \sigma_{12}\tau_{il} \\ \sigma_{12}\tau_{ij} & \sigma_2^2\tau_j^Y & \sigma_{12}\tau_{kj} & 0 \\ 0 & \sigma_{12}\tau_{kj} & \sigma_1^2\tau_k^X & \sigma_{12}\tau_{kl} \\ \sigma_{12}\tau_{il} & 0 & \sigma_{12}\tau_{kl} & \sigma_2^2\tau_l^Y \end{pmatrix} \right),$$

we have

$$\begin{aligned} \text{Var}\left(R_iS_j\right) &= \sigma_1^2\sigma_2^2\tau_i^X\tau_j^Y + \sigma_{12}^2\tau_{ij}^2 \\ \text{Cov}\left(R_iS_j, R_kS_l\right) &= \tau_{il}\tau_{kj}\sigma_{12}^2, \quad ij \neq kl, \end{aligned}$$

by standard distribution theory.

The first term in (5.8) reduces to

$$\sigma_1^2\sigma_2^2 \sum_{ij} v_{ij}^2\tau_i^X\tau_j^Y + \sigma_{12}^2 \sum_{ij} v_{ij}^2\tau_{ij}^2.$$

The summation in second term of (5.8) can be split into 3 cases

$$\sum_{ij\neq kl} = \sum_{i=k,j\neq l} + \sum_{j=l,i\neq k} + \sum_{i\neq k,j\neq l}.$$

For the first case

$$\begin{aligned} \sum_{i,j\neq l} w_{ij}w_{il}\text{Cov}\left(R_iS_j, R_kS_l\right) &= \sigma_{12}^2 \sum_{i,j\neq l} w_{ij}w_{il}\tau_{ij}\tau_{il} \\ &= \sigma_{12}^2 \left[\sum_{i,j,l} w_{ij}w_{il}\tau_{ij}\tau_{il} - \sum_{ij} v_{ij}^2\tau_{ij}^2 \right] \\ &= \sigma_{12}^2 \left[\sum_i (\sum_j w_{ij}\tau_{ij})^2 - \sum_{ij} v_{ij}^2\tau_{ij}^2 \right], \end{aligned}$$

and similarly for the second case.

For the third case, where $i \neq k$ and $j \neq l$, the summation becomes

$$\sigma_{12}^2 \sum_{i \neq k, j \neq l} w_{ij} w_{kl} \tau_{il} \tau_{kj}.$$

But if $\tau_{il}, \tau_{kj}, w_{ij}$ and w_{kl} are all non-zero then both J_i^X and J_k^X must intersect both J_j^Y and J_l^Y which is impossible since J_i^X and J_k^X are disjoint, as are J_j^Y and J_l^Y. It follows that the summation in the third case is zero.

Collecting terms, the result follows. □

Corollary 5.1. *Under the assumptions of Section 5.2 the basic Zhou-Hayashi-Yoshida estimator has conditional variance*

$$\text{Var}\,(\text{ZHY}|\tau) = \sigma_1^2 \sigma_2^2 B_1 + \sigma_{12}^2 B_2,$$

where

$$\begin{aligned} B_1 &= \textstyle\sum_{ij} \tau_i^X \tau_j^Y \mathbb{I}(\tau_{ij} > 0) \\ B_2 &= \textstyle\sum_i (\tau_i^X)^2 + \sum_j (\tau_j^Y)^2 - \sum_{ij} \tau_{ij}^2. \end{aligned} \tag{5.9}$$

Proof. Noting that $\sum_{ij} \tau_{ij} = 1$, the total length of the interval $(0, 1)$, and $\sum_i \tau_{ij} = \tau_j^X$ and $\sum_j \tau_{ij} = \tau_i^Y$, the result follows from (5.7), by substituting $w_{ij} = \mathbb{I}(\tau_{ij} > 0)$. □

5.4 A special case

A special case of the extended estimator is got by replacing the returns (R_i, S_j) in the basic ZHY estimator by interpolated versions reflecting the amount of overlap. For asset

A the interpolated value is $\tau_{ij}R_i/\tau_i^X$ and $\tau_{ij}S_j/\tau_j^Y$ for asset B. If the interpolated values are treated as actual log-returns then a natural estimator would involve the sum of their products divided by τ_{ij}. The estimator then has the form

$$\sum_{i=1}^{m}\sum_{j=1}^{n} R_i S_j \tau_{ij}/(\tau_i^X \tau_j^Y),$$

so that

$$\tilde{\sigma}_{12} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} \tau_{ij} R_i S_j/(\tau_i^X \tau_j^Y)}{\sum_{i=1}^{m}\sum_{j=1}^{n} \tau_{ij}^2/(\tau_i^X \tau_j^Y)},$$

is an unbiased estimator of σ_{12}, with variance $\sigma_1^2\sigma_2^2 C_1 + \sigma_{12}^2 C_2$, where

$$C_1 = 1/\sum_{ij}\theta_{ij}$$
$$C_2 = \left[\sum_i(\sum_j \theta_{ij})^2 + \sum_j(\sum_i \theta_{ij})^2 - \sum_{ij}\theta_{ij}^2\right] \Big/ \left(\sum_{ij}\theta_{ij}\right)^2,$$

and $\theta_{ij} = \tau_{ij}^2/(\tau_i^X \tau_j^Y)$.

5.4.1 Comparison with ZHY

The variances of ZHY and the new estimator $\tilde{\sigma}_{12}$ differ in the coefficients of $\sigma_1^2\sigma_2^2$ and σ_{12}^2. We can show that C_1 is never larger than B_1. To see this note that

$$\sum_{ij}\theta_{ij} = \sum_{ij}\tau_{ij}\mathbb{I}(\tau_{ij} > 0)\phi_{ij},$$

where $\sum_{ij}\tau_{ij}\mathbb{I}(\tau_{ij} > 0) = 1$ and $\phi_{ij} = \tau_{ij}/(\tau_i^X \tau_j^Y$.

In other words, $\sum_{ij}\theta_{ij}$ can be thought of as $\mathbb{E}(\phi)$, the mean of ϕ_{ij} with the discrete

probability mass function $\{p_{ij}\}$ where $p_{ij} = \tau_{ij}\mathbb{I}(\tau_{ij} > 0)$.

By Jensen's inequality $\mathbb{E}\,(1/\phi) \geqslant 1/\mathbb{E}\,(\phi)$, so that

$$\sum_{ij} \frac{\tau_i^X \tau_j^Y}{\tau_{ij}} \tau_{ij}\mathbb{I}(\tau_{ij} > 0) \geqslant \left[\sum_{ij} \frac{\tau_{ij}}{\tau_i^X \tau_j^Y} \tau_{ij}\mathbb{I}(\tau_{ij} > 0)\right]^{-1} = 1/\sum_{ij} \theta_{ij},$$

and $C_1 \leqslant B_1$ as claimed.

The weights $w_{ij} = \tau_{ij}/(\tau_i^X \tau_j^Y)$ are in fact the optimal values as regards minimising A_1 in (5.7) for the general extended estimator. This can be seen by considering the problem of minimising the numerator $\sum_{ij} v_{ij}^2 \tau_i^X \tau_j^Y \mathbb{I}(\tau_{ij} > 0)$ with respect to $\{w_{ij}\}$ under the denominator constraint $\sum_{ij} \tau_{ij} w_{ij}$. Introducing the Lagrange multiplier λ and differentiating with respect to w_{ij} we have

$$2w_{ij}\tau_i^X \tau_j^Y + \lambda \tau_{ij} = 0,$$

so that

$$w_{ij} \propto \tau_{ij}/(\tau_i^X \tau_j^Y).$$

Numerically we find C_2 is not always smaller than and B_2, but we can compare their average values and their empirical distributions when observation times are generated by Poisson processes, i.e. when $(t_1, \ldots, t_m)$ and $(u_1, \ldots, u_n)$ are associated with Poisson processes with rates λ_1 and λ_2, respectively. This is the approach adopted by Hayashi and Yoshida (2008) in their asymptotic analysis of the ZHY estimator as λ_1 and λ_2 tend to infinity. It is also the framework in which Griffin and Oomen (2011) compare the variance of the basic ZHY estimator with various estimators based on artificially synchronised prices. They derive expressions for the expected values of B_1 and B_2 valid for large values

of λ_1 and λ_2. These results can be also be derived directly from (5.9) by taking $M = m$ and $N = n$ to have Poisson distributions conditioned to be positive and $(t_1, \ldots, t_m)$ and $(u_1, \ldots, u_n)$ to be Poisson process event times conditional on $t_m = u_n = 1$.

For $\mathbb{E}(B_2)$ we then have conditionally on $M = m$ and $N = n$

$$\mathbb{E}(B_2|m, n) = m\mathbb{E}[(\tau_1^X)^2] + n\mathbb{E}[(\tau_1^X)^2] - (m + n - 1)\mathbb{E}(\tau^2),$$

where τ_1^X, τ_1^Y are typical values of τ_i^X and τ_j^Y and τ is a typical gap in the ordered union of $(t_1, \ldots, t_m)$ and $(u_1, \ldots, u_n)$.

Since

$$\mathbb{P}(\tau_1^X > x|m) = (1 - x)^{m-1}, \quad 0 < x < 1,$$

it follows that $\mathbb{E}[(\tau_1^X)^2|m] = 2/[m(m + 1)]$ so that unconditionally

$$\mathbb{E}[M(\tau_1^X)^2] = \sum_{m=1}^{\infty} \frac{2\,\mathrm{e}^{-\lambda_1}\lambda_1^m}{(m + 1)m!(1 - \mathrm{e}^{-\lambda_1})} = \frac{2}{\lambda_1} - \frac{2}{\mathrm{e}^{\lambda_1} - 1} \sim \frac{2}{\lambda_1}.$$

Similarly, $\mathbb{E}[(\tau_1^Y)^2] \sim 2/\lambda_2$ and $\mathbb{E}[\tau^2] \sim 2/(\lambda_1 + \lambda_2)$, so that

$$\mathbb{E}(B_2) \sim \frac{2}{\lambda_1} + \frac{2}{\lambda_2} - \frac{2}{\lambda_1 + \lambda_2}, \tag{5.10}$$

as λ_1 and λ_2 tend to infinity.

In the same way it can be shown that

$$\mathbb{E}(B_1) \sim \frac{2}{\lambda_1} + \frac{2}{\lambda_2}. \tag{5.11}$$

in agreement with Griffin and Oomen (2011). We have confirmed the result both alge-

braically and numerically.

Note however that Griffin and Oomen give the value

$$\frac{2}{\lambda_1+\lambda_2}\left(\frac{\lambda_1}{\lambda_2}+\frac{\lambda_2}{\lambda_1}\right)$$

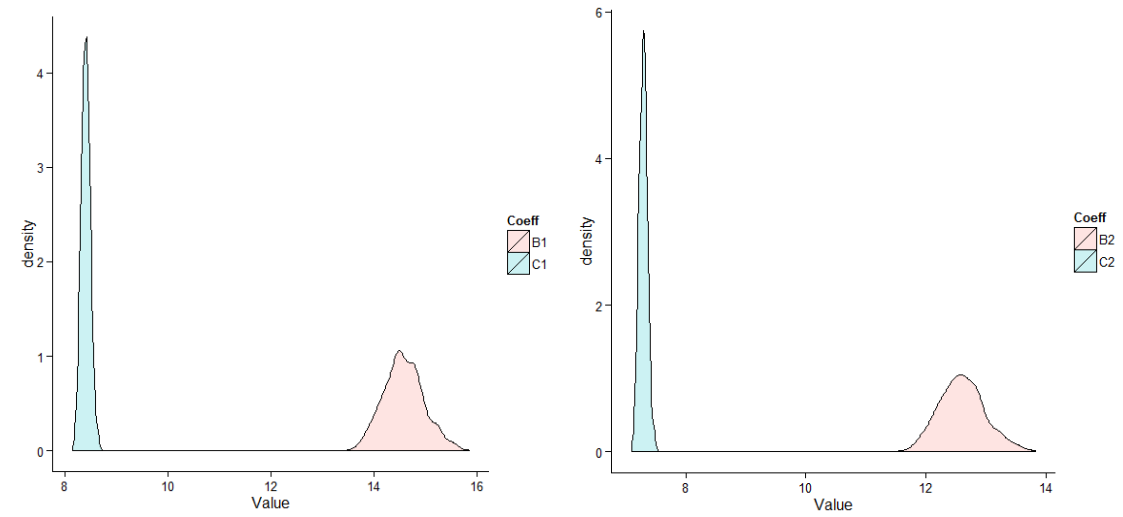
for $\mathbb{E}(B_2)$. This appears to be in error. We have confirmed this error both algebraically and numerically by simulation.

5.4.2 Numerical comparisons

From the general form of the variances we see it is only necessary to compare the coefficients of $\sigma_1^2\sigma_2^2$ and σ_{12}^2 and that these only depend on the timestamp structure. We have shown that C_1 is never larger than B_1 however when $m+n$ is small we have found cases where C_2 is larger than B_2. These seem to be rare cases. To illustrate the typical relative performance of $\tilde{\sigma}_{12}$ and ZHY, we present numerical results working with timestamps from independent Poisson processes. Along the way, we were able to confirm the accuracy of (5.11) and (5.10). The results are presented in Figure 5.4.

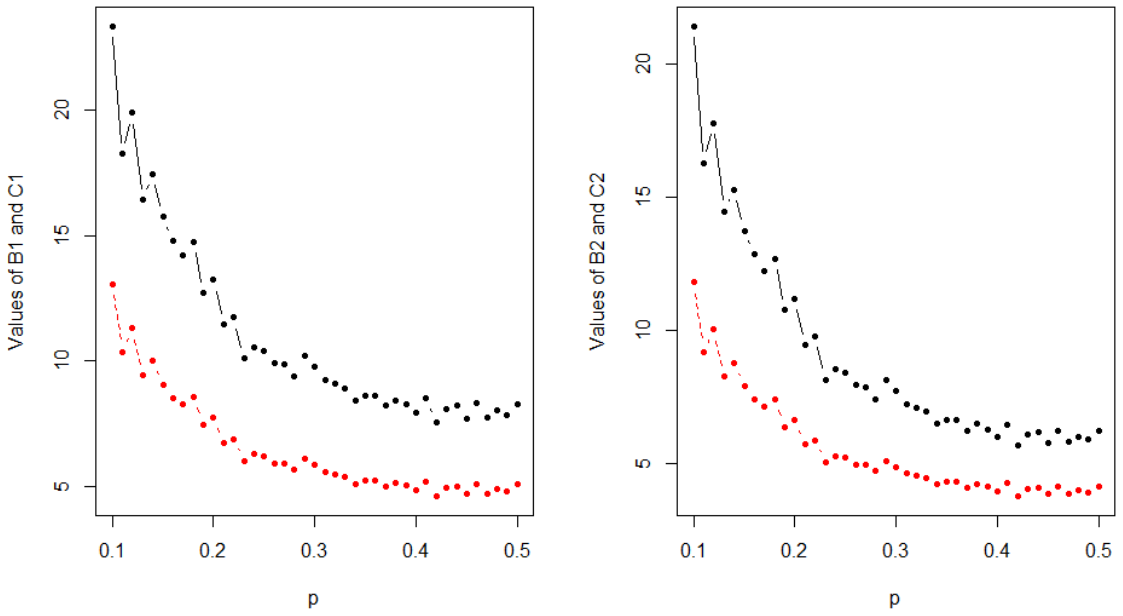
We see that in both cases the coefficients in the extended ZHY has are substantially smaller those for the standard ZHY estimator. We also compare the values of the coefficients for various value of the ratio $p=\lambda_1/(\lambda_1+\lambda_2)$ in Figure 5.4. Again the extended ZHY estimator is consistently better.

For completeness we will derive the exact variance for the general class of estimator proposed in Section 5.3 as a function of the time-stamp structure. Some additional notation is required.



(a)

(b)



(c)

Figure 5.4: Panel (a) compares simulated values of B_1 and C_1. Panel (b) compares B_2 and C_2. In both cases there are 1000 simulations with $\lambda_1 = 1000$ and $\lambda_2 = 5000$ and the coefficients are scaled by $\lambda_1 + \lambda_2$ for convenient axis values. Panel (c) shows the effect of varying the ratio $p = \lambda_1/(\lambda_1 + \lambda_2)$.

5.4.3 Additional notation

The time-stamp structure is determined by $(\mathcal{T}, \mathcal{U})$, the two sequences of observation times between 0 and 1. When these sequences are interleaved, the gaps in the combined sequence have lengths $\{\tau_{ij}\}$, where τ_{ij} is the length of the overlap between τ_i^X and τ_j^Y. Since $\mathcal{T} = (t_0, \ldots, t_m)$ and $\mathcal{U} = (u_0, \ldots, u_n)$, the number of these gaps is $L = m+n-1$. It is convenient to number the gaps as $p = 1, \ldots, L$, from left to right, and then define $w_{[p]}$ to be the value of w_{ij} associated with the pth gap, $p = 1, \ldots, L$, i.e. the gap where J_i^X and J_j^Y overlap. For notational convenience we also put $w_{[0]} = w_{[N+1]} = 0$.

Since $w_{[p]}$ is associated with a gap in the combined sequence, it will either lie between timestamps that are both from $\mathcal{T}$, both from $\mathcal{U}$ or otherwise. We denote the latter case where there is alternation, by Q

Theorem 5.2. *Under the assumptions* (5.5)*, with normally distributed observation error, the extended ZHY estimator has variance*

$$\text{Var}\,(\hat{\sigma}_{12}) = \sigma_1^2\sigma_2^2 A_1 + \sigma_{12}^2 A_2 + \left[\nu_1^2\sigma_2^2 D_1 + \nu_2^2\sigma_1^2 D_2 + \nu_1^2\nu_2^2 D_3\right]/T^2(v),$$

where

$$
\begin{aligned}
A_1 &= \big[\sum_{ij} w_{ij}^2 \tau_i^X \tau_j^Y\big]/T^2(w) \\
A_2 &= \big[\sum_i (\sum_j w_{ij}\tau_{ij})^2 + \sum_j (\sum_i w_{ij}\tau_{ij})^2 - \sum_{ij} w_{ij}^2\tau_{ij}^2\big]/T^2(w) \\
D_1 &= 2\sum_i \tau_i^X [\sum_j w_{ij}^2 - \sum_j w_{ij}w_{i,j+1}] \\
D_2 &= 2\sum_j \tau_j^Y [\sum_i w_{ij}^2 - \sum_i w_{ij}w_{i+1,j}] \\
D_3 &= 4\sum_{p=1}^{L-1} (w_{[p]} - w_{[p+1]})^2 + 2\sum_{p\in Q} w_{[p-1]}w_{[p+1]} \\
T(w) &= \sum_{ij} \tau_{ij} w_{ij}
\end{aligned}
\tag{5.12}
$$

Proof. From (5.6) the variance of $\hat{\sigma}_{12}$ is Var $\big(\sum_{ij} R_i S_j w_{ij}\big)/T^2(w)$, where

$$
\text{Var}\big(\sum_{ij} R_i S_j w_{ij}\big) = \sum_{ij} w_{ij}^2 \text{Var}\left(R_i S_j\right) + \sum_{ij\neq kl} w_{ij}w_{kl}\text{Cov}\left(R_i S_j, R_k S_l\right). \tag{5.13}
$$

By the assumptions of the theorem, the vector $(R_i, S_j, R_k, S_l)^{\mathrm{T}}$ is multivariate normal with

$$
\begin{aligned}
\text{Var}\left(R_i\right) &= \sigma_1^2\tau_i^X + 2\nu_1^2 \quad \text{Var}\left(R_k\right) = \sigma_1^2\tau_k^X + 2\nu_1^2 \\
\text{Var}\left(S_j\right) &= \sigma_2^2\tau_j^Y + 2\nu_2^2 \quad \text{Var}\left(S_l\right) = \sigma_2^2\tau_l^Y + 2\nu_2^2
\end{aligned}
$$

and

$$
\begin{aligned}
\text{Cov}\left(R_i, S_j\right) &= \sigma_{12}\tau_{ij} & \text{Cov}\left(R_k, S_j\right) &= \sigma_{12}\tau_{kj} \\
\text{Cov}\left(R_i, S_l\right) &= \sigma_{12}\tau_{il} & \text{Cov}\left(R_k, S_l\right) &= \sigma_{12}\tau_{kl} \\
\text{Cov}\left(R_i, R_k\right) &= -\nu_1^2\mathbb{I}_1(i\sim k) & \text{Cov}\left(S_j, S_l\right) &= -\nu_2^2\mathbb{I}_2(j\sim l),
\end{aligned}
$$

where the indicator $\mathbb{I}_1(i\sim k)$ is 1 when J_i^X is adjacent to J_k^X and similarly $\mathbb{I}_2(j\sim l) = 1$ when J_j^Y is adjacent to J_l^Y.

It follows again by standard normal distribution theory that

$$\begin{aligned}
\text{Var}\,(R_iS_j) &= (\sigma_1^2\tau_i^X + 2\nu_1^2)(\sigma_2^2\tau_j^Y + 2\nu_2^2) \\
\text{Cov}\,(R_iS_j, R_iS_l) &= \sigma_{12}^2\tau_{ij}\tau_{il} - \nu_2^2\mathbb{I}_2(j\sim l)(\sigma_1^2\tau_i^X + 2\nu_1^2), \quad j\neq l \\
\text{Cov}\,(R_iS_j, R_kS_j) &= \sigma_{12}^2\tau_{ij}\tau_{kj} - \nu_1^2\mathbb{I}I_1(i\sim k)(\sigma_2^2\tau_j^Y + 2\nu_2^2), \quad i\neq k \\
\text{Cov}\,(R_iS_j, R_kS_l) &= \sigma_{12}^2\tau_{il}\tau_{kl} + \nu_1^2\nu_2^2\mathbb{I}_1(i\sim k)\mathbb{I}_2(j\sim l), \quad i\neq k, j\neq l
\end{aligned}$$

The first term on the right-hand side of (5.13) reduces to

$$\begin{aligned}
\sum_{ij}\text{Var}\,(R_iS_j)w_{ij}^2 = \sigma_1^2\sigma_2^2\sum_{ij}\tau_i^X\tau_j^Y v_{ij}^2 + 2\nu_2^2\sigma_1^2\sum_i\tau_i^X\sum_j v_{ij}^2 \\
+ 2\nu_1^2\sigma_2^2\sum_j\tau_j^Y\sum_i v_{ij}^2 + 4\nu_1^2\nu_2^2\sum_{ij}v_{ij}^2. \qquad (5.14)
\end{aligned}$$

The second summation on the right-hand side of (5.13) can be split into 3 cases

$$\sum_{ij\neq kl} = \sum_{i=k,j\neq l} + \sum_{j=l,i\neq k} + \sum_{i\neq k,j\neq l}.$$

For the first case,

$$\begin{aligned}
&\sum_{i=k,j\neq l} w_{ij}w_{il}\text{Cov}\,(R_iS_j, R_iS_l) \\
&= \sigma_{12}^2\sum_{i,j\neq l}w_{ij}w_{il}\tau_{ij}\tau_{il} - \nu_2^2\sum_{i,j\neq l}(\sigma_1^2\tau_i^X + 2\nu_1^2)w_{ij}w_{il}\mathbb{I}_2(j\sim l) \\
&= \sigma_{12}^2\big[\sum_i(\sum_j w_{ij}\tau_{ij})^2 - \sum_{ij}v_{ij}^2\tau_{ij}^2\big] - \nu_2^2\sum_i(\sigma_1^2\tau_i^X + 2\nu_1^2)P_i^Y \qquad (5.15)
\end{aligned}$$

where P_i^Y is the sum of products $w_{ij}w_{il}$ for all pairs (j,l) such that J_j^Y and J_l^Y are adjacent, i.e. $P_i^Y = 2\sum_j w_{ij}w_{i,j+1}$.

Similarly for the second case

$$\sum_{j=l,i\neq k} w_{ij}w_{kj}\text{Cov}\,(R_iS_j, R_kS_j)$$
$$= \sigma_{12}^2\big[\sum_j(\sum_i w_{ij}\tau_{ij})^2 - \sum_{ij} v_{ij}^2\tau_{ij}^2\big] - \nu_1^2\sum_j(\sigma_2^2\tau_j^Y + 2\nu_2^2)P_j^X, \tag{5.16}$$

where P_j^X is the sum of products $w_{ij}w_{kj}$ for all pairs (i,k) such that J_i^X and J_k^X are adjacent, i.e. $P_j^X = 2\sum_i w_{ij}w_{i+1,j}$.

For the third case

$$\sum_{i\neq k,j\neq l} w_{ij}w_{kl}\text{Cov}\,(R_iS_j, R_kS_l)$$
$$= \sum_{i\neq k,j\neq l}\sigma_{12}^2 w_{ij}w_{kl}\tau_{il}\tau_{kj} + \nu_1^2\nu_2^2\sum_{i\neq k,j\neq l} w_{ij}w_{kl}\mathbb{I}_1(i\sim k)\mathbb{I}_2(j\sim l)$$
$$= \quad 0 + \nu_1^2\nu_2^2\sum_{i\neq k,j\neq l} w_{ij}w_{kl}\mathbb{I}_1(i\sim k)\mathbb{I}_2(j\sim l), \tag{5.17}$$

since if $\tau_{il}, \tau_{kj}, w_{ij}$ and w_{kl} are all non-zero then both J_i^X and J_k^X must intersect both J_j^Y and J_l^Y which is impossible since J_i^X and J_k^X are disjoint, as are J_j^Y and J_l^Y.

From (5.17), referring to Figure 5.5, we see that

$$\sum_{i\neq k,j\neq l} w_{ij}w_{kl}\mathbb{I}_1(i\sim k)\mathbb{I}_2(j\sim l) = 2\sum_{p\in Q} w_{[p-1]}w_{[p+1]}, \tag{5.18}$$

since when (J_i^X, J_k^X) and (J_j^Y, J_l^Y) are both neighbouring pairs, the products $w_{ij}w_{kl}$ are non-zero unless there is an alternation at one of the gaps τ_{ij}, τ_{il},τ_{kj} or τ_{kl}. Furthermore there are only two such arrangements since the ordering (i,k) has to agree with the ordering (j,l).

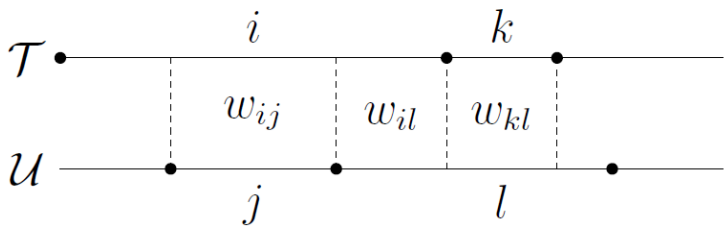


Figure 5.5: An alternation between $\mathcal{T}$ and $\mathcal{U}$ timestamps at the gap τ_{il}. The ordering (i,k) has to agree with the ordering (j,l) to have both w_{ij} and w_{kl} non-zero.

Collecting coefficients of $\sigma_1^2\sigma_2^2$ and σ_{12}^2, the expressions for A_1 and A_2 follow as in Theorem 5.1.

The coefficient of $\nu_1^2\sigma_2^2$ comes from (5.15) and (5.14). Similarly, the coefficient of $\nu_1^2\sigma_2^2$ is from (5.16) and (5.14).

Collecting terms in $\nu_1^2\nu_2^2$ from (5.14), (5.15) and (5.16) we have

$$4\sum_{ij} w_{ij}^2 - 4\sum_{j} w_{ij}w_{i,j+1} - 4\sum_{i} w_{ij}w_{i+1,j} = 4\sum_{p}(w_{[p]} - w_{[p+1]})^2,$$

and together with (5.18) we have D_3. □

There is simplification in the basic ZHY case where $w_{ij} = \mathbb{I}_{ij}$.

Corollary 5.2. *Under the assumptions of Theorem 5.2, the basic Zhou-Hayashi-Yoshida estimator has conditional variance*

$$\text{Var}\,(\text{ZHY}|\tau) = \sigma_1^2\sigma_2^2 B_1 + \sigma_{12}^2 B_2 + 2\nu_1^2\sigma_2^2 + 2\nu_2^2\sigma_1^2 + 4\nu_1^2\nu_2^2|Q|,$$

where

$$\begin{aligned} B_1 &= \textstyle\sum_{ij} \tau_i^X \tau_j^Y \mathbb{I}(\tau_{ij} > 0) \\ B_2 &= \textstyle\sum_i (\tau_i^X)^2 + \sum_j (\tau_j^Y)^2 - \sum_{ij} \tau_{ij}^2, \end{aligned} \tag{5.19}$$

and $|Q|$ is the number of alternating gaps.

Proof. For B_1 and B_2, the result follows as in Corollary 5.1; it also follows that $T(w) = 1$. The coefficient of $\nu_1^2\sigma_2^2$ can be obtained from D_1 in Theorem 5.2 with $w_{ij} = \mathbb{I}(\tau_{ij} > 0)$. Note that $\sum_j \mathbb{I}(\tau_{ij} > 0)$ is the number of J_j^Y returns that overlap J_i^X and $\sum_j \mathbb{I}(\tau_{ij} > 0)\mathbb{I}(\tau_{i,j+1} > 0)$ is the number of timestamps in $\mathcal{U}$ that lie inside J_i^X, i.e. one fewer than the number of J_j^Y returns that overlap J_i^X. It follows that

$$\sum_j \mathbb{I}^2(\tau_{ij} > 0) - \sum_j \mathbb{I}(\tau_{ij} > 0)\mathbb{I}(\tau_{i,j+1} > 0) = 1,$$

and similarly for D_2. Finally, $w_{ij} = \mathbb{I}(\tau_{ij} > 0)$ implies $w_{[p]} = 1, p = 1, \ldots, N$ so the first term in D_3 is zero and the second term is the number of alternations $|Q|$. □

As in the noiseless case we can consider the average value of the variance when the timestamps are from two independent Poisson processes with rates λ_1 and λ_2, respectively. Beyond the noise free case discussed in Section 5.4.1, this only requires the evaluation of $\mathbb{E}(|Q|)$. But the probability of an alternation is $2\alpha(1-\alpha)$ where $\alpha = \lambda_1/(\lambda_1 + \lambda_2)$, and since the expected number of gaps is of order $\lambda_1 + \lambda_2$, it follows that $\mathbb{E}(|Q|) \sim \lambda_1\lambda_2/(\lambda_1 + \lambda_2)$, in agreement with the result obtained by Griffin and Oomen (2011).

5.4.4 Further extensions

The estimators we have considered so far can be extended by *subsampling*. This was proposed originally by Zhou (1995) and follows naturally from the subsampling methods that he proposed for volatility estimation in Zhou (1996). The basic idea is use a sample of every kth observation to estimate the parameter and then to average the resulting values. Zhou shows that this will reduce the standard error and he makes recommendations about

the optimal value of k. See also Griffin and Oomen (2011) and the references therein for recent applications of this technique.

5.5 Maximum likelihood estimation

We now look briefly at the problem of maximum likelihood estimation under strong model assumptions, specifically the case where the efficient price can be observed without noise and the infinitesimal parameters are constant. We note however, as before, that it may be possible to remove microstructure noise using the methods of Robert and Rosenbaum (2012) and that we can also deal with time-varying parameters by suitable rescaling of time.

When $(\breve{X}, \breve{Y})$ is bivariate Brownian motion, as in Section 5.1, the likelihood of the observed value $\bar{r}$ of the combined vector of returns $\bar{R} = (R, S)$ is

$$L(\theta) = \frac{\exp\left(-\frac{1}{2}\bar{r}^{\mathsf{T}} V^{-1} \bar{r}\right)}{(2\pi)^{(m+n)/2}\sqrt{\det(V)}},$$

where

$$V = \begin{pmatrix} \sigma_1^2 D_1 & \sigma_{12} B \\ \sigma_{12} B^{\mathsf{T}} & \sigma_2^2 D_2 \end{pmatrix},$$

and

$$D_1 = \text{diag}(\tau_i^X; i = 1, \ldots, m), \quad D_2 = \text{diag}(\tau_j^Y; j = 1, \ldots, n)$$

and B is a matrix with elements $B_{ij} = \tau_{ij}; i = 1, \ldots, m; j = 1, \ldots, n$.

In principle, the parameters $\theta = (\sigma_1, \sigma_2, \sigma_{12})$ can then be estimated by maximum likeli-

hood and the correlation recovered from

$$\hat{\rho} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2},$$

but the vector $\bar{r}$ is very long and it is computationally very expensive to obtain the maximum likelihood estimate. The EM algorithm offers a possible computational approach.

5.5.1 EM algorithm

We start by supposing that we have full synchronous data for both assets. In this case it is easy to find the MLE. Recall that to apply the EM algorithm we need to calculate the expected value of the full log-likelihood given our partially observed data (Dempster et al., 1977).

Suppose that we have synchronous price information for assets A and asset B at every time $t \in \mathcal{T} \cup \mathcal{U}$. Let $\tilde{\tau}_1, \tilde{\tau}_2, \ldots, \tilde{\tau}_L$ be the return times for the full data and denote the associated log-returns of the assets by $(\tilde{R}_k, \tilde{S}_k); k = 1, 2, \ldots, L$. In this case, we simply have a sequence of independent pairs. The kth pair has a bivariate normal distribution with mean zero and variance-covariance matrix $\tilde{\tau}_k \Sigma$.

The full likelihood is then

$$\tilde{L}(\theta) = \frac{\exp\left[-\frac{1}{2}(1-\rho^2)^{-1}\left(\sigma_1^{-2} h_{1,1} - 2\rho(\sigma_1\sigma_2)^{-1} h_{1,2} + \sigma_2^{-2} h_{2,2}\right)\right]}{(2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2})^N (\prod \tilde{\tau}_k)^{1/2}},$$

where

$$H_{1,1} = \sum_{k=1}^{N} \tilde{R}_k^2/\tilde{\tau}_k, \quad H_{1,2} = \sum_{k=1}^{N} \tilde{R}_k \tilde{S}_k/\tilde{\tau}_k, \quad \text{and} \quad H_{2,2} = \sum_{k=1}^{N} \tilde{S}_k^2/\tilde{\tau}_k.$$

It is well-known, and straightforward to show, that the MLEs are

$$\tilde{\sigma}_1^2 = N^{-1}H_{1,1}, \quad \tilde{\sigma}_{12} = N^{-1}H_{1,2}, \quad \tilde{\sigma}_2^2 = N^{-1}H_{2,2}$$

To apply the EM algorithm we need the expectation of $\log(\tilde{L}(\theta)$, in particular the expectations of $H_{1,1}, H_{1,2}$ and $H_{2,2}$ given $\bar{R} = \bar{r}$, as follows.

First note that it is only necessary to consider the joint distribution of

$$Z = (\tilde{R}_k, \tilde{S}_k, \bar{R})^\mathsf{T}$$

for each $k = 1, 2, \ldots, N$, since the conditional expectations we want are sums of terms involving $(\tilde{R}_k, \tilde{S}_k)$.

The distribution is multivariate normal with variance-covariance matrix

$$\begin{pmatrix} \tilde{\tau}_k\Sigma & \tilde{\tau}_k\alpha^\mathsf{T} & \tilde{\tau}_k\beta^\mathsf{T} \\ \tilde{\tau}_k\alpha & \sigma_1^2 D_1 & \sigma_{12}C \\ \tilde{\tau}_k\beta & \sigma_{12}C^\mathsf{T} & \sigma_2^2 D_2 \end{pmatrix},$$

where α is an $m \times 2$ matrix with ith row $(\sigma_1^2, \sigma_{12})^\mathsf{T}$ and zero elements elsewhere and β is an $n \times 2$ matrix with jth row $(\sigma_{12}, \sigma_2^2)^\mathsf{T}$ and zero elements elsewhere, where (i, j) are the indices of the time periods J_i^X and J_j^Y that intersect to give $\tilde{\tau}_k$.

The conditional distribution of $(\tilde{R}_k, \tilde{S}_k)$ given $\bar{R} = \bar{r}$ is then multivariate normal with mean

$$\boldsymbol{\mu} = \tilde{\tau}_k \, (\alpha^\mathsf{T}, \beta^\mathsf{T}) \begin{pmatrix} \sigma_1^2 D_1 & \sigma_{12}C \\ \sigma_{12}C^\mathsf{T} & \sigma_2^2 D_2 \end{pmatrix}^{-1} \bar{r} = \tilde{\tau}_k (\alpha^\mathsf{T}, \beta^\mathsf{T}) V^{-1} \bar{r},$$

and variance-covariance matrix

$$\Lambda = \tilde{\tau}_k \Sigma - \tilde{\tau}_k^2 (\alpha^\mathsf{T}, \beta^\mathsf{T}) V^{-1} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

The conditional expectation of $\tilde{R}_k \tilde{S}_k$ given $\bar{R} = \bar{r}$ is then $\Lambda_{12} + \mu_1 \mu_2$, with similar expressions for the expectations of $\tilde{R}_k^2$ and $\tilde{S}_k^2$. This enables the EM algorithm to be set up.

The form of the matrix V simplifies its inversion. Using the Boltz–Banachiewicz matrix identity (Puntanen and Styan, 2005) we have

$$\begin{pmatrix} \sigma_1^2 D_1 & \sigma_{12} C \\ \sigma_{12} C^\mathsf{T} & \sigma_2^2 D_2 \end{pmatrix}^{-1} = \begin{pmatrix} \sigma_1^{-2}(D_1^{-1} + \rho^2 F E^{-1} F^\mathsf{T}) & -\rho/(\sigma_1\sigma_2) F E^{-1} \\ -\rho/(\sigma_1\sigma_2) E^{-1} F^\mathsf{T} & \sigma_2^{-2} E^{-1} \end{pmatrix}, \quad (5.20)$$

where $E = D_2 - \rho^2 C^\mathsf{T} D_1^{-1} C$ and $F = D_1^{-1} C$. Note that D_1 is diagonal so its inverse can be found readily. The algorithm can be easily modified to incorporate this simplification.

For large values of $\min(m, n)$, matrix inversion (or equivalently the solution of the linear system) is the dominant computational task. However as we shall see in the next section there are other ways in which the cost of calculation can be reduced.

5.5.2 Asynchronous episodes

One reason for implementing the EM algorithm was to provide a robust check on estimators obtained by other methods. We now look more closely at the problem of finding the MLE.

First note that if, at any time, both assets are traded simultaneously then the process after this time is independent of the process before. We define an *asynchronous episode* to be a period during which every trade is asynchronous and in which each return period of asset A intersects with at least one return period of asset B. After removing initial and final periods where only one asset is traded, the succession of synchronisation points divides the process into a succession of independent asynchronous episodes, as illustrated in Figure 5.6

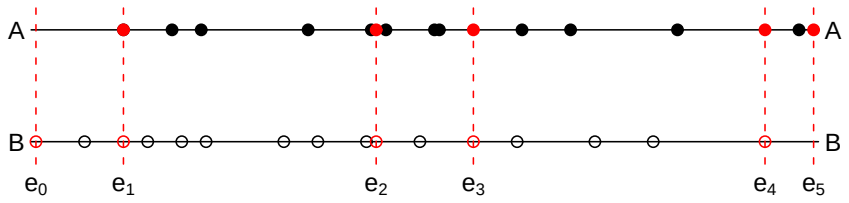


Figure 5.6: Times of trades of assets A and B. The points $(e_1, \ldots, e_4)$ are synchronisation points. These points separate the process into independent asynchronous episodes. In this illustration, prior to e_1 only asset B provides information and after e_4 only asset A

The likelihood can then be written as

$$L(\theta) = L_0^*(\theta) \prod_{k=1}^{K} L_k(\theta) L_1^*(\theta),$$

where $L_0^*(\theta)$ and $L_1^*(\theta)$ depend on the data from a single asset (at the beginning and end of the sequence) and $L_k(\theta)$ depends on data from the kth asynchronous episode. Note that if synchronisation is frequent this produces substantial savings in the cost of calculating the likelihood.

When $L_0^*(\theta)$ depends only on asset A, for example, we have

$$L_0^*(\theta) = \frac{\exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n_0} r_i^2/\tau_i^X\right]}{\prod_{i=1}^{n_0}\sqrt{2\pi\sigma_1^2/\tau_i^X}},$$

where n_0 is the number of return periods prior to the first asynchronous episode and r_i is the ith return for that asset.

For a typical asynchronous episode we have

$$L_k(\theta) = \frac{\exp\left(-\frac{1}{2}\bar{r}^{\mathsf{T}}V^{-1}\bar{r}\right)}{(2\pi)^{(m+n)/2}\sqrt{\det(V)}},$$

where $\bar{r}$, m, n and V now refer to the information in that episode. For notational simplicity we will not annotate these terms with k.

Using (5.20) we have

$$L_k(\theta) = \frac{\exp\left(-\frac{1}{2}(M_1 - 2M_{12} + M_2)\right)}{(2\pi)^{(m+n)/2}\sqrt{\det(V)}},$$

where

$$\begin{aligned} M_1 &= r^{\mathsf{T}}(D_1^{-1} - \rho^2 F E^{-1} F^{\mathsf{T}})r, \\ M_{12} &= \rho r^{\mathsf{T}} F E^{-1} s/(\sigma_1\sigma_2), \\ M_2 &= s^{\mathsf{T}} E^{-1} s/\sigma_2^2, \end{aligned}$$

and the $\det(V)$ can be calculated as $\sigma_1^{2n}\sigma_2^{2m}\det(D_1)\det(D_2 - \rho^2 B^{\mathsf{T}} D_1^{-1} B)$.

The combined likelihood then becomes

$$L(\theta) = \frac{\exp\left(-\frac{1}{2}(Q_1(\rho)/\sigma_1^2 - 2Q_{12}(\rho)/(\sigma_1\sigma_2) + Q_2(\rho)/\sigma_2^2)\right)}{(2\pi)^{(m+n)/2}\sigma_1^m\sigma_2^n\sqrt{v(\rho)}}, \tag{5.21}$$

where $v(\rho)$ arises as a product of determinants and $Q_1(\rho)$, $Q_2(\rho)$ and $Q_{12}(\rho)$ are sums of quadratic terms from the components of the partitioned likelihood, depending only on ρ.

5.5.3 Profile likelihood for correlation

Maximising (5.21) with respect to σ_1^2 and σ_2^2 for fixed ρ is straightforward. The maximising values can be given explicitly into terms of $Q_1(\rho)$, $Q_2(\rho)$ and $Q_{12}(\rho)$. Hence the profile likelihood for ρ can be calculated and confidence intervals for ρ evaluated by considering the values of ρ at which the profile log-likelihood falls significantly below its maximum. This algorithm has been implemented and applied to data relating the USD-JPY exchange rate and the price of 10 year US treasury bond futures. The results are not included here.

5.5.4 Kalman filter

As with volatility estimation in Chapter 4 the likelihood can be evaluated by the Kalman filter and then maximised routinely. For large scale problems involving data at multiple time points this is a substantial computational task. For smaller problems this becomes feasible. Timing comparisons were made between the partitioning methods described in Section 5.5.2 and the Kalman filter using a subset of the USD-JPY exchange rate and 10 year US Treasury bond futures data. The results indicate that the partitioning method is competitive with the Kalman filter when implemented in R using the 'fast' Kalman filter

package. We have not given numerical comparisons here because they are highly sensitive to implementation details.

5.5.5 Recursive covolatility estimation

It should be noted that maximum likelihood estimation will never be a serious competitor for covolatility estimation in high-speed trading. It is a batch processing algorithm that may be of use in historical analysis. The advantage of the ZHY estimator and its extensions is that they can be calculated in linear time, linear in the number of data points. Furthermore, we can produce exponentially weighted versions of the estimators that can be updated using simple recursion as described in a more general context in Chapter 3 and specifically for volatility estimation in Chapter 4. This enables changing patterns of covolatility to be monitored in real time.

Chapter 6

Lead-lag relationships

Arbitrage traders take advantage of correlations between financial instruments. They try to detect and exploit pricing anomalies. Such opportunities can arise when the price movements of one asset lags behind those of another. Cross-correlations and lead-lag indicators are important in trying to find out which asset leads.

When the assets are known to be correlated traders can make predictions of the eventual price of the lagging instrument and place orders to buy or sell accordingly. Market makers who post prices in these markets have closely coupled interest since lead-lag relationships will determine the price they are prepared to quote to other traders. For high-frequency trading it is important to determine such effects on very short time scales, since some assets follow the path of others with a small time lag.

Lead-lag effects are not constant throughout the day and may vary from day to day. There is intraday seasonality associated with specific times such as the announcement of macroeconomic figures and the US market opening. Strong asymmetric cross-correlation functions are empirically observed, especially in the Futures and Equities markets. Nor-

mally the most liquid assets with short inter-trade duration, narrow bid-ask spread and small volatility lead smaller stocks. These lead-lag relationships become more and more pronounced as we zoom in on significant events.

In this chapter we start by looking at established methods for measuring lead-lag effects and illustrate their use, noting that the issue of asynchronicity arises immediately for price determination, as in Chapter 5. We then focus on lead-lag analysis in low latency trading networks where trigger and response events are discrete and recorded to nanosecond precision. We demonstrate how a detailed picture of lagged response can be built up for communication across competing high-speed networks linking multiple exchanges. We then develop and implement an algorithm to deal with confounding effects arising from closely spaced trigger events.

6.1 Choice of observables

There are many ways of defining 'price' and 'observation time' when working with financial time series derived from exchange sources. Lead-lag relationships are highly dependent on the observational framework chosen. As previously explained in Section 2.4.2, price can be a simple mid-price or a weighted mid-price, weighting inversely by the volume of asset available at the quoted bid and offer prices. More elaborate combinations of prices and volumes at various depths in the order book are also utilised.

As we will see below, the advantage of working with quoted prices and volumes is that they are available continuously in time. The disadvantage in practice is that the associated datasets are large. Changes in quoted prices and volumes may be fleeting, existing for fractions of a second before the quoted price and volume is modified or cancelled. Such

behaviour is reminiscent of the promotional FX prices on Reuters screens as witnessed by Zhou (1996) in the days of telephone trading.

Alternatively the prices of completed trades can be used. These will move between the bid and offer prices depending on whether the trader is buying or selling – giving rise to the well known bid/ask bounce effect (Roll, 1984). To deal with these fluctuations the quoted mid-price at the time of the trade can be used. Another possibility is to use prices generated by the 'uncertainty zone' methods of Robert and Rosenbaum (2012).

There are also many options in defining timestamps that will serve as the observational clock. The times when trades occur is an obvious candidate. The times when the mid-price moves is another. High-frequency traders will also look at the BBOT (best bid/offer and trade) timestamps, the times when the quoted volumes at the best bid and offer prices change, either because volume has been added or cancelled or a trade has occurred.

6.2 Measuring lead-lag for prices

As in Chapter 5, we suppose that there are two assets A and B with prices X and Y that are recorded at discrete times $\mathcal{T} = (t_0, \ldots, t_m)$ and $\mathcal{U} = (u_0, \ldots, u_n)$. For lead-lag analysis, we would like to know how the association between the returns of X and Y changes when the timestamps of Y are shifted relative to the timestamps of X. In other words how the association between the returns of $X(t)$ and $Y(t+h)$ changes with lag h.

Covolatility can be used as a measure of association. In Chapter 5 we discussed how to estimate covolatility in the case $h = 0$. In principle these methods can be adapted to estimate the covolatility between the returns of $X(t)$ and $Y(t+h)$ for any value of h.

6.2.1 Synchronous data

The simplest case is when $\mathcal{T} = \mathcal{U}$ and the observation times are equally spaced with spacing Δ. This arrangement is possible when using quote data and only observing the quoted prices at a sequence of times $i\Delta, i = 0, 1, \ldots$. In this case the lagged cross-covariance between the return sequences with lags $h = k\Delta$ provides a basic measure of association so that the effect of changing h can be investigated

6.2.2 Asynchronous data

More generally covolatility can be estimated using any of the methods in Chapter 5. In particular, shifting the observation times of Y by h, the lagged ZHY estimator becomes:

$$\text{ZHY}(h) = \sum_{ij} R_i S_j(h) \mathbb{I}(\tau_{ij}(h) > 0), \quad h \in \mathbb{R},$$

where

$$S_j(h) = Y(u_j + h) - Y(u_{j-1} + h)$$

and

$$\tau_{ij}(h) = |J_i^X \cap (u_{j-1} + h, u_j + h)|,$$

is the length of the intersection between J_i^X and $(u_{j-1} + h, u_j + h)$, i.e. the jth return interval of Y shifted by the amount h. See Section 5.2.1 for definitions of J_i^X, etc. As in previous Chapters, a physical or TTS clock can be used.

In summary, the ZHY estimator is computed using lagged observations where the times of Y are shifted by h. By plotting $\text{ZHY}(h)$ against h, the values of h that maximise $\text{ZHY}(h)$ can be determined. This approach was suggested by Hoffmann et al. (2013) and

further studied by Huth and Abergel (2014) and Alsayed and McGroarty (2014). Huth and Abergel suggest assessing the strength of lead-lag by comparing the areas under ZHY(h) on either side of zero. They show that this measure is related to the notion of Granger causality (Granger, 1969).

Other measures of covolatility can be adapted in a similar manner, see for example de Jong and Nijman (1997) and references therein. Alsayed and McGroarty (2014) compare the performance of the ZHY(h) lead-lag estimator with an estimator based on the previous tick (PT) covariance estimator. They show that the PT covariance produces biased estimators of h when one of the assets is observed more frequently than the other. By contrast the ZHY based estimator they advocate is shown to be unbiased.

In their theoretical analysis, Huth and Abergel (2014), assume that log-prices move as correlated Brownian motion and that observation times are drawn from independent Poisson processes. In the same paper they construct the ZHY(h) plot for French equity and Index Futures data and show that the next midquote movement of a stock can be predicted from the past evolution of the Index Futures with high probability.

6.2.3 Comparison with extended ZHY

We follow Huth and Abergel (2014) and simulate log-prices $(X(t), Y(t+h))$ as correlated Brownian motions with zero means and constant parameters $\sigma_1 = \sigma_2 = 1$ and $\sigma_{12} = 0.9$. The processes are assumed to run on physical time measured in seconds and for the simulation study, h is chosen to be 0.2 seconds. Again following Huth and Abergel (2014) observation times for X and Y are taken from independent Poisson process with rate parameters λ_1 and λ_2, respectively. For a similar setup, see Section 5.4.1.

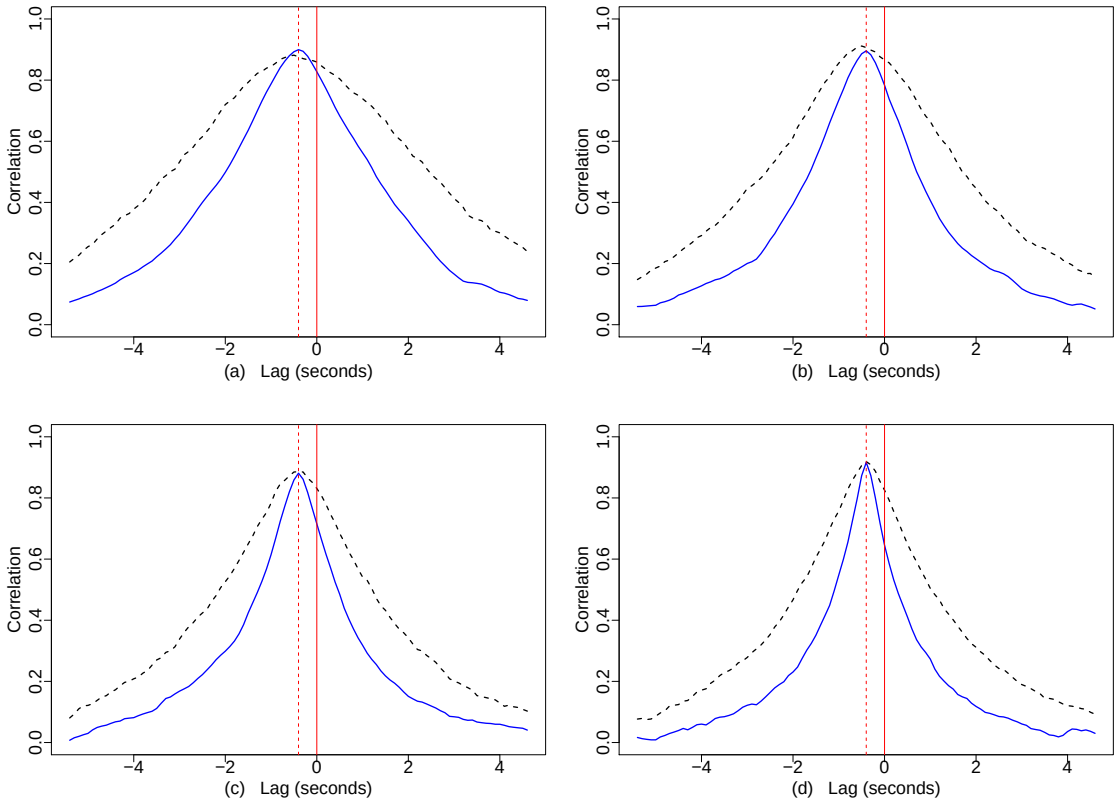


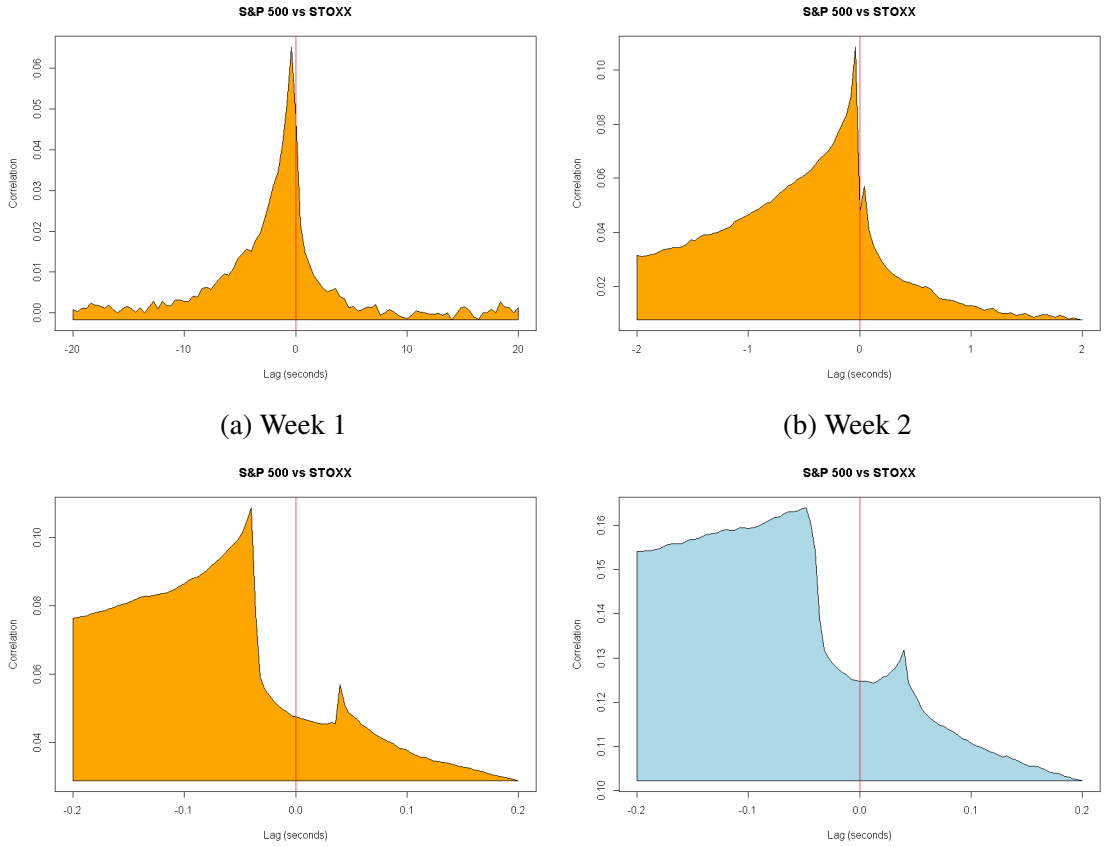
Figure 6.1: Comparison of simulated lead-lag profiles obtained with ZHY(h) and the extended ZHY estimator of Section 5.4. The dashed curves are ZHY(h). The ratios λ_1/λ_2 are 1, 2, 5 and 10 in (a), (b), (c) and (d), respectively. The dashed vertical line corresponds to the known lag -0.2 secs.

Figure 6.1, compares the performance of the standard ZHY based estimator used by Huth and Abergel (2014) with the extended ZHY developed in Section 5.4.

It is clear that our extended ZHY provides a sharper estimate of the known lag under these model assumptions.

6.2.4 Application – Index Futures

Figure 6.2 shows the extended ZHY(h) plot for recent Futures data where asset A is the S&P 500 E-mini Futures contract traded on CME in Chicago and B is the Euro Stoxx Index Futures contract traded on the Eurex exchange in Frankfurt. The data are mid-prices at the time of trades. If the peak is at a negative value, it indicates that on average Frankfurt price movement is delayed relative to Chicago activity.



(a) Week 1

(b) Week 2

(c) Week 1

(d) Week 2

Figure 6.2: Lead-lag plots with the extended ZHY covariance of Section 5.4 for S&P 500 Futures in Chicago versus Euro Stoxx Futures in Frankfurt – two weeks in 2014. Prices are mid-prices at the times of trades. Plots are shown for two time resolutions.

The broad picture is clear – namely that the Chicago Futures contract leads the Frankfurt Futures contract. The major peak is around $h = -50$ milliseconds, indicating that Frankfurt responses to Chicago activity are delayed by this amount on average. There is also a minor peak at $h = 50$ milliseconds, indicating that less commonly what happens in Frankfurt will have some impact on what happens in Chicago.

It is also clear that Figure 6.2 does not resemble the plots in Figure 6.1, made under Poisson tick time, Brownian price assumptions. A possible explanation lies in the discrete

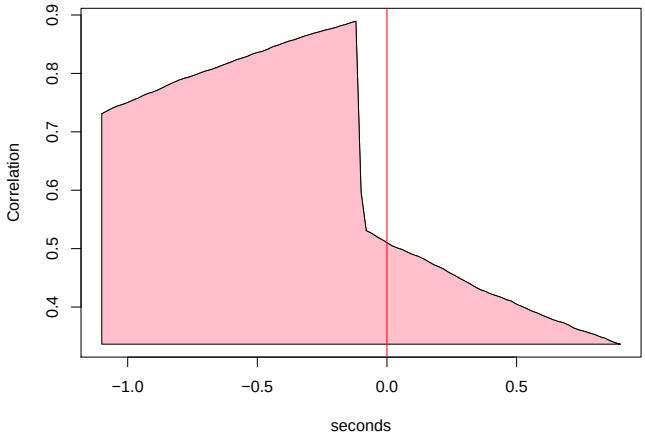


Figure 6.3: Simulated asymmetric lead-lag plot where lag is introduced in both tick and calendar time. See text for details of the simulation.

nature of time at this atomic level of analysis. Brief bursts of activity are followed by relatively long periods with no trades. Lead-lag effects may be more sensibly described in tick time as illustrated in Figure 6.3 and described below.

Suppose that $\{t_i\}$ is a sequence of tick times and that $\{Z_i\}$ and $\{Y_i\}$ are associated random price sequences with $(Z_i - Z_{i-1}, Y_i - Y_{i-1})$ iid pairs of correlated bivariate normal variables. Now let $X_i = Z_{i+1}$, so that X leads Y by one tick in tick time (TTS). The simulation in Figure 6.3 shows what happens when the ZHY(h) plot is calculated. Asymmetric

behaviour similar to Figure 6.2 is reproduced.

Other explanations in terms of auto-correlation effects are possible. See for the example the references in Griffin and Oomen (2011) relating to controlling lead-lag bias in covariance estimation. Why does it have asymmetric behaviour? To be honest we don't really know. This will be a very interesting area for further research.

6.3 High-frequency lead-lag

Competitive performance is a major concern for high-frequency traders who invest substantial amounts of money in leasing or building their own fast connections between exchanges. The introduction of fibre-optic and microwave connections between exchanges marked major steps in improving connection latency. Nowadays, high-speed connections can transmit data between Chicago and Frankfurt in under 40 milliseconds. The bulk of the transmission time is taken up by the transatlantic portion of the route. Around a dozen transatlantic cables are available, running between New York and the cable landing point near Bude in Cornwall, UK. Faster microwave networks link Chicago with New York and link Bude to London, down to Dover, across the Channel and then on to Frankfurt (Trader, 2014a,b).

The broad picture provided by our ZHY(h) plot is not able to expose the detailed microstructure of high-speed lagged responses. High-frequency traders want to know how much slower or faster their connections are than others, i.e. what is the latency profile of competitors between specific exchanges. In the next section we demonstrate how to reveal this profile, discuss possible weaknesses and propose and implement an improved method.

6.3.1 Event-based latency profiles

The objective is to investigate the different network and infrastructure speeds used by high-frequency trading firms when communicating between the major financial exchanges. A very simple way to do this is to wait for a large trigger trade in Chicago for example, then see how soon people trade in the same direction in other exchanges such as New York, London and Frankfurt after this event.

If a significant event occurs in Chicago, e.g. some large trade at a particular time, traders in Frankfurt will receive the signal at different times, depending on how fast their network connection and trading infrastructure is. We would like to build a latency profile for their responses.

6.3.2 Exploratory analysis

Our first proposal exploits the bursty nature of trade data. S&P 500 Futures trades in Chicago can be viewed as trigger events. The times between trades of Futures contracts on CME have a heavy tailed distribution. We can find numerous S&P 500 Futures trades that are separated from trades of the same contract by significant gaps of time, both before and after. Treating these special trades as triggers we can look at potential BBOT (best bid, offer and trade) responses involving the Euro Stoxx contract in Frankfurt for example, i.e. trades, addition, cancellation or modification of volume at the best prices. For each such event we look back to the most recent Chicago trigger event and prepare a histogram of the elapsed time between the trigger and the response.

Figure 6.4 shows the result of such a calculation. Basic features of the response profile can be identified: the first peak, a peak at around 39 milliseconds corresponding to

high-frequency traders with comprehensive microwave connectivity in Europe and the US and the second peak, a lesser peak at around 45 milliseconds corresponding to slower fibre-optic connection on the mainland routes. There is an intermediate peak for traders who have sub-optimal microwave connections. Responses prior to 35 milliseconds are background noise.

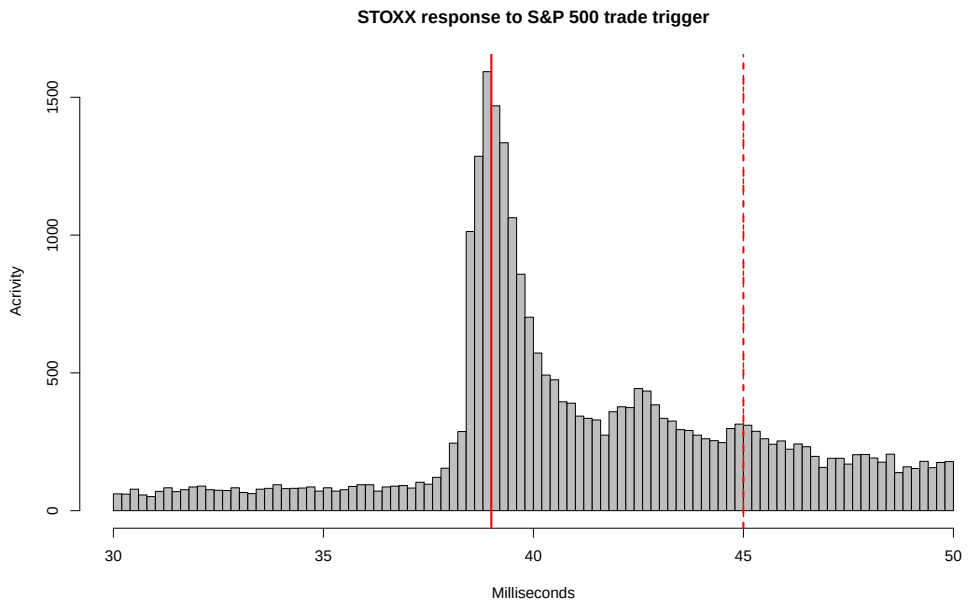


Figure 6.4: Response delay for BBOT activity of Euro Stoxx Futures with sparse S&P Futures triggers, for a typical day early in 2014.

Figure 6.4 is not a very complicated profile. What are much more interesting are the European profiles. For example, from Stockholm to Milan, there are several routes one can take: go through London or go by Zurich or go by Frankfurt or by Paris. From North Europe to South Europe, one can get a very complicated latency profile. Investigation of such profiles is commercially sensitive and requires access to high-precision transaction timestamps at multiple exchanges.

6.3.3 A theoretical framework

Although there are many Chicago trades that fit our separation criterion, there may be concern that their impact on the response profile is not representative of all trades. The problem is that unseparated trades may produce responses that overlap with responses from other trades that are close in time.

The simple histogram plot is not able to separate noise from signal. From Figure 6.4 we can see, there is a lot of noise before 35 millisecond and also a lot of noise in the tail too. Before 35 millisecond, this is nearly certainly background noise. The problem is how to eliminate the noise.

To formalise the problem we denote the sequence of trigger times at exchange A by $\mathcal{T} = (t_1, \ldots, t_m)$, i.e. the trigger times at exchange A are t_1 to t_m. Each trigger independently produces a cluster of delayed responses at exchange B. i.e. each trigger independently produces a cluster of delayed responses from trading company x, trading company y, trading company z, etc. Different trading companies will respond at different times, depending on what kind of network speed they have. The number of responses is not fixed. It will be variable depending for example on the size of the trigger trade, or the state of the trading company's inventory or position. Of course a trading company may decide not to respond at all, so the number of responses that each trigger produces will change depending on circumstances.

To allow for this variability, we model the number of responses for each trigger as a Poisson variable with mean μ. We also assume that individual response delays for each trigger are drawn independently from a distribution F, i.e. the individual response delays vary between different trading companies; some respond fast and some respond slow. As a consequence, there is pattern of events generated in exchange B, every time there is an

event in exchange A.

In addition there will be responses at exchange B that are not the direct result of activity on exchange A. These arise from people trading on exchange B using information that does not come from exchange A. We call this background. For example, activity in Frankfurt will come from internal behaviour of the market in Frankfurt, or information from New York, London, Paris, Milan or wherever it is. So there will be events which are not triggered by Chicago. To accommodate this, we suppose that independently there will be untriggered background events at exchange B and that these form a Poisson process with constant intensity λ_0. This noise is not what we are interested in. We are interested in the response to the Chicago trigger. Our objective is then to estimate μ, F, λ_0.

Although we would like to do everything in continuous time, for now we are going to switch to discrete time. To make progress with our formulation we will assume that $\{t_i\}$ and the response times take discrete values on a regular grid with grid points numbered in $\mathbb{N}$. We are just going to round all the times to the nearest q microseconds, so our clock will just tick every q microseconds. Therefore, we get a distribution on that discrete time scale with a probability mass function. For high-frequency applications we take the grid width to be around $q = 100$ microseconds. The delay distribution F is assumed to have finite support with probability mass function $\{p_k, k = 1, \ldots, d\}$.

Figure 6.5 illustrates what might be observed with events occurring in Chicago and some events occurring in Frankfurt.

The idea is to figure out which of the events in B are generated by which of the events in A. If we just look at it, we will have no idea at all how B relates to A. Our task is to untangle what we see in B and split it into some noise plus some signal.

The input data set is then $\{t_i\}$, the trade times at exchange A, and for each j, the number

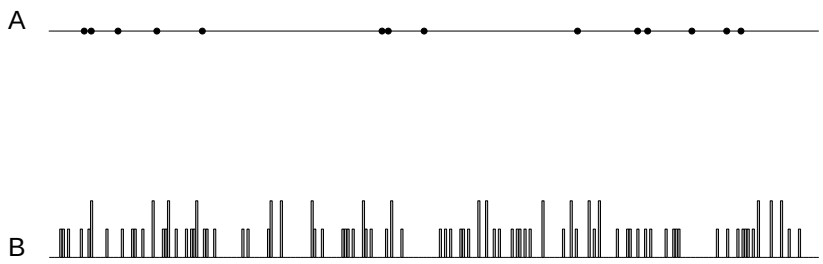


Figure 6.5: The events at exchange B are either delayed responses to triggers on A or background noise. BBOT responses on B will outnumber trade triggers on A.

of events W_j at the jth grid point on exchange B as illustrated in Figure 6.5.

Futures exchanges have substantial activity for around 14 hours a day so N the number of 100 microsecond grid points is around half a billion. Our main interest is in estimating the response profile, $\{p_k, k = 1, \ldots, d\}$ where d is a few hundred.

We know the structure of W_j. It is the sum of background noise (with parameter λ_0) plus contributions from events that occurred in exchange A, taking into account the latencies of different participants. These contributions have a Poisson distribution depending on parameters p and μ. The likelihood of the parameter set (p, μ, λ_0) becomes a simple product of Poisson likelihoods.

Since W_j is the sum of independent Poisson variables with contributions from each of the triggers and the background, it has a Poisson distribution with mean μ_j where

$$\mu_j = \lambda_0 + \mu \sum_{j-d \leqslant r \leqslant j} p(j - r + 1)\mathbb{I}_r, \tag{6.1}$$

where $\mathbb{I}_r = 1$ when $r \in \mathcal{T}$ and zero otherwise. Since the contributions to each of the

elements of $\{W_j\}$ are independent, the likelihood of p, μ, λ_0 given $\{W_j = w_j\}$ is

$$L(p, \mu, \lambda_0) = \prod_{j=1}^{N} \frac{\mathrm{e}^{-\mu_j} \mu_j^{w_j}}{w_j!}. \tag{6.2}$$

Direct optimisation is unrealistic but with the aid of a short cut we will show that the EM algorithm provides a practical solution.

6.3.4 Optimisation

For convenience let $\lambda_k = \mu p_k, k = 1, \ldots, d$. Now note if we have available $Z_k^{(s)}$, the number of responses triggered at time s that arrive at time $s + k - 1$ for $s \in \mathcal{T}$, then we can estimate λ_k directly as

$$\hat{\lambda}_k = \frac{1}{m} \sum_{s \in \mathcal{T}} Z_k^{(s)}, \tag{6.3}$$

since $\{Z_k^{(s)}, s \in \mathcal{T}\}$ are independent Poisson variables each with mean λ_k. Similarly, the background rate λ_0 can be estimated by

$$\hat{\lambda}_0 = \frac{\sum_{j=1}^{N} \left(W_j - \sum_{s \in \mathcal{T}} Z_{j-s+1}^{(s)} \right)}{N},$$

since when the triggered responses have been removed only the background events remain and these are independently Poisson with mean λ_0.

To apply the EM algorithm to maximise 6.2 we proceed as usual to find the expected value of the log likelihood which in this case is the sum of log likelihoods for the independent Poisson components. Since the distributions are Poisson this reduces to finding terms like

$\mathbb{E}\left(Z_k^{(s)}|W\right)$, with

$$W_j = \sum_{s\in\mathcal{T}} Z_{j-s+1}^{(s)} + B_j, \quad j = 1, \ldots, N, \tag{6.4}$$

where B_j is the unobserved background count. We see that $Z_k^{(s)}$ appears in W_j only when $j = s + k - 1$.

From (6.4) since W_j is the sum if independent Poisson components $\{Z_{j-s+1}^{(s)}, s \in \mathcal{T}\}$ and B_j, with means $\{\lambda_{j-s+1}, s \in \mathcal{T}\}$ and λ_0, respectively. It follows that

$$\mathbb{E}\left(Z_{j-s+1}^{(s)}|W\right) = \mathbb{E}\left(Z_{j-s+1}^{(s)}|W_j\right) = \frac{\lambda_{j-s+1}}{\mathbb{E}\left(W_j\right)} W_j, \quad s \in \mathcal{T},$$

so that

$$\mathbb{E}\left(Z_k^{(s)}|W\right) = \frac{\lambda_k}{\mathbb{E}\left(W_{k+s-1}\right)} W_{k+s-1}, \quad s \in \mathcal{T}. \tag{6.5}$$

and similarly

$$\mathbb{E}\left(B_j|W\right) = \frac{\lambda_0}{\mathbb{E}\left(W_j\right)} W_j, \tag{6.6}$$

where

$$\mathbb{E}\left(W_j\right) = \sum_{r\in\mathcal{T}} \lambda_{j-r+1} + \lambda_0, \tag{6.7}$$

from 6.4.

6.3.5 EM steps

The EM iteration starts with initial estimates of $\lambda_k, k = 0, \ldots, d$, denoted by $\{\hat{\lambda}_k^{(0)}\}$. Starting at step $h = 0$, the iteration is as follows:

(1) Calculate $\mathbb{E}\left(Z_k^{(s)}|W\right)$ using (6.5), (6.6) and (6.7) using $\{\hat{\lambda}_k^{(h)}\}$.

(2) Replace $Z_k^{(s)}$ in (6.3) by the calculated values $\mathbb{E}\left(Z_k^{(s)}|W\right)$ to obtain a revised esti-

mate of λ_k, namely

$$\hat{\lambda}_k^{(h+1)} = \frac{1}{m} \sum_{s \in \mathcal{T}} \left(\frac{\hat{\lambda}_k^{(h)} W_{k+s-1}}{\sum_{r \in \mathcal{T}} \hat{\lambda}_{k+s-r}^{(h)} + \hat{\lambda}_0^{(h)}} \right), \quad k = 1, \ldots, d,$$

and

$$\hat{\lambda}_0^{(h+1)} = \frac{1}{N} \sum_{j=1}^{N} \left(\frac{\hat{\lambda}_0^{(h)} W_j}{\sum_{s \in \mathcal{T}} \hat{\lambda}_{j-s+1}^{(h)} + \hat{\lambda}_0^{(h)}} \right).$$

(3) Iterate to convergence.

The calculation of $\mathbb{E}(W_j)$ at each step is a key component of the algorithm. The first term in (6.7) can be written as

$$\phi(j) = \sum_{r \in \mathcal{T}} \lambda_{j-r+1} = \sum_{i=1}^{N} \lambda_{j-i+1} \mathbb{I}_i = \sum_{k=1}^{d \wedge j} \lambda_k \mathbb{I}_{j-k+1},$$

where $\mathbb{I}$ is the indicator defined in (6.1). This means that ϕ is a moving average of $\mathbb{I}$, that can be calculated rapidly by the filter function in R.

Although the calculation time is linear in N given the size of N this still represents a substantial computing effort.

However the bursty nature of high-frequency data allows time-savings to be made. This is because any response period longer than d without a trigger event can only contain background noise, assuming d is sufficiently large to cover the range of possible delay times. Periods longer than d can be truncated without affecting the calculation and the periods that are removed provide a separate estimate of the background rate. The pre-processing step is as follows.

(a) First fix the grid points. For example set the spacing at 100 microseconds.

(b) Fix the range of response times of interest, say 30000 to 50000 microseconds, so

that $d = 200$.

(c) Remove time periods where there is no trigger for L grid points, say $L = 1000$, i.e. 0.1 seconds.

For a typical daily trading period, 07:00 to 21:00, with the figures as itemized above, the initial size of N is approximately 500 million. On a reasonably active day there may be $100,000$ trigger events, i.e. $100,000$ trades of the S&P Futures contract in Chicago. Choosing a day at random we found that $35,000$ trades were followed by a gap of at least 0.1 seconds. The total length of these gaps is 13.8 hours, i.e. most of the day's trading is compressed into 12 minutes! Since there are 7.2 million grid points in 12 minutes, we have reduced the size of N by a factor of 70. Further reduction can be achieved by reducing L. As a result the modified EM algorithm becomes feasible and runs in a few seconds in R.

To illustrate we apply the EM algorithm to the data of Figure 6.4. The grid width is taken to be 100 microseconds with $d = 200$. Convergence is extremely rapid, with the basic shape of the profile visible after two iterations.

There is a general view in the statistical community that the EM algorithm is slow to converge. This is not the case here. It seems that our model specification is particularly well adapted to the application of the EM algorithm.

Figures 6.6 and 6.7 show the result of the EM algorithm after 2 and then 20 iterations. Note that the algorithm progressively eliminates unwanted noise as was the original intention.

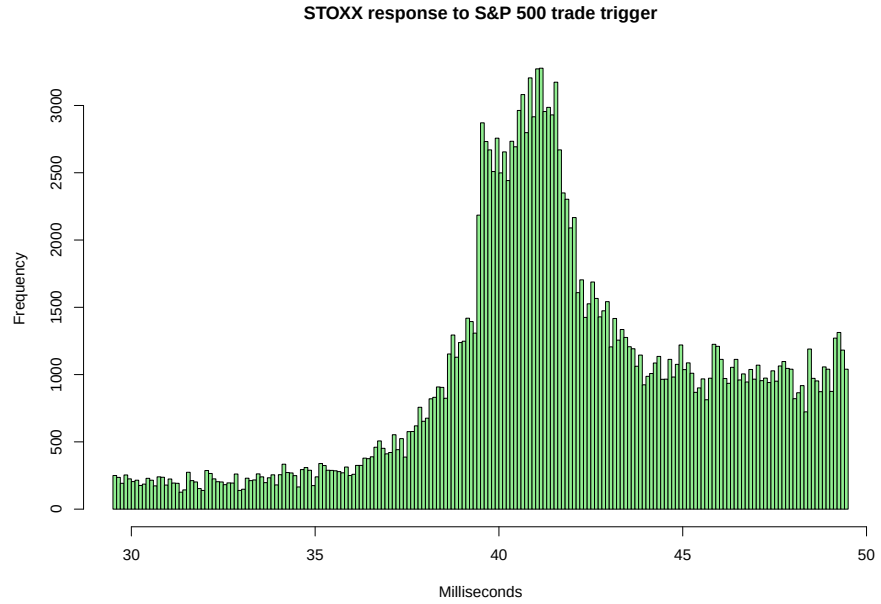


Figure 6.6: Maximum likelihood reconstruction of the profile of Euro Stoxx Futures responses to trades of S&P Futures, after 2 iterations of the EM algorithm.

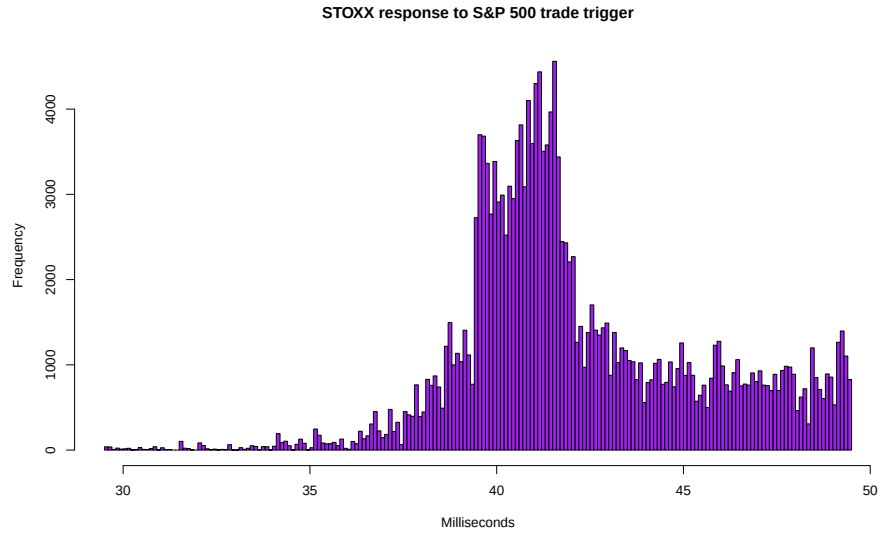


Figure 6.7: Maximum likelihood reconstruction of the profile of Euro Stoxx Futures responses to trades of S&P Futures, after 20 iterations of the EM algorithm.

Chapter 7

Conclusion

The focus in this thesis has been on high-frequency trading and the compromises that have to made in implementing statistical tools under time constraints. This leads us to consider statistics that can be computed in linear time and updated recursively. High-frequency trading forms a small part of the financial trading landscape. The fastest high-frequency traders are typically private firms trading with their own money. Hedge funds have a different perspective, with large capital bases under management. For them major disruptions are particular important and the strategies they develop can be quite different from the high-frequency trader

Now that I have a general picture of the methods that available for the estimation of volatility and covolatility, I have started to focus more closely on the practical problem of cross-Gamma hedging in the presence of jumps in asset prices. Although they do not discuss the effect of jumps, the results on Delta hedging reported in Mykland and Zhang (2008) are intriguing. It will be interesting to see whether similar revelations are possible for Gamma and cross-Gamma hedging problems when dealing with rapid price

movements.

A formal statistical methodology for modelling the impact of such jumps on the future volatility process is needed, for example to determine the best hedging strategy.

Issues with hedging

A hybrid product has a value, V, at any particular point in time. An underlying asset involved in the contract often has a spot price S, the price that is quoted for immediate settlement, i.e. payment and delivery. *Delta* is the first derivative of V with respect to S. A *Delta hedge* is a scheme that is designed to make the value of a contract insensitive to small changes in the price of the underlying asset. *Gamma* is a term reserved for the second derivative of V with respect to S. Gamma is a measure of the rate of change of Delta with respect to S. *Cross-Gamma* is the mixed second partial derivative of a hybrid product with respect to the spot prices S_1 and S_2 of two different underlying assets. It can be thought of as the sensitivity of the Delta for the first underlying to changes in the spot level of the second, or vice versa. Ideally one would like to design investment strategies that simultaneously hedge Delta, Gamma and Cross-Gamma risk.

Jumps are important because they represent a significant source of non-diversifiable risk as discussed at length in (Bollerslev et al., 2008) and the references therein. For risk management, a risk averse investor might be expected to shun investments with sharp unforeseeable movements. Jumps like that are of great importance for standard arbitrage-based arguments and derivatives pricing in particular, as the effect cannot readily be hedged by a portfolio of the underlying asset, cash, and other derivatives. The presence of jumps makes incomplete markets. The degree of market incompleteness depends on the size and intensity of jumps, which determines the magnitude of derivative hedging error (see

Naik and Lee, 1990; and Bertsimas et al., 2001).

The practical questions are: should we hedge before or after the jump, or both? If so how should we hedge it? And having detected a jump what do we do with it? Can we use the direction, size and intensity of jumps to forecast subsequent instantaneous volatility? If so, how fast can we estimate the volatility after the jump?

Bibliography

Aït-Sahalia, Y., Fan, J., and Xiu, D. (2010). High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association*, 105(492):1504–1517.

Aït-Sahalia, Y., Mykland, P., and Zhang, L. (2011). Ultra high frequency volatility estimation with dependent microstructure noise. *Journal of Econometrics*, 160(1):160–175.

Aït-Sahalia, Y. and Mykland, P. A. (2009). Estimating volatility in the presence of market microstructure noise: A review of the theory and practical considerations. In Anderson, T., Davis, R., J.-P., K., and Mikosh, T., editors, *Handbook of financial time series*, pages 577–598. Springer.

Aït-Sahalia, Y., Mykland, P. A., and Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial studies*, 18(2):351–416.

Almgren, R. (2012). High-frequency event analysis in eurex interest rate futures. Technical report, Working Paper, Quantitative Brokers.

Alsayed, H. and McGroarty, F. (2014). Ultra-high-frequency algorithmic arbitrage across international index futures. *Journal of Forecasting*, 33(6):391–408.

Andersen, T., Bollerslev, T., Diebold, F., and Labys, P. (2000). Great realisations. *Risk Magazine*, 13(3):105–109.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, 96(453).

Ané, T. and Geman, H. (2000). Order flow, transaction clock, and normality of asset returns. *Journal of Finance*, pages 2259–2284.

Bannouh, K., Van Dijk, D., and Martens, M. (2009). Range-based covariance estimation using high-frequency data: The realized co-range. *Journal of Financial Econometrics*, 7(4):341–372.

Barnard, G. (1963). Discussion: The spectral analysis of point processes (by M.S. Bartlett). *Journal of the Royal Statistical Society. Series B*, 25:294.

Barndorff-Nielsen, O., Hansen, P., Lunde, A., and Shephard, N. (2009). Realized kernels in practice: Trades and quotes. *The Econometrics Journal*, 12(3):C1–C32.

Barndorff-Nielsen, O., Hansen, P., Lunde, A., and Shephard, N. (2011). Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics*, 162(2):149–169.

Barndorff-Nielsen, O. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society. Series B*, pages 253–280.

Barndorff-Nielsen, O. and Shephard, N. (2004a). Econometric analysis of realized co-

variation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica*, 72(3):885–925.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2004). Regular and modified kernel-based estimators of integrated variance: The case with independent noise. Technical report, Economics Group, Nuffield College, University of Oxford.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6):1481–1536.

Barndorff-Nielsen, O. E. and Shephard, N. (2004b). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2(1):1–37.

Bauwens, L., Hafner, C., and Sebastien, L. (2012). *Handbook of Volatility Models and Their Applications*. Wiley.

Bollerslev, T., Law, T., and Tauchen, G. (2008). Risk, jumps, and diversification. *Journal of Econometrics*, 144(1):234–256.

Brown, R. G. (2004). *Smoothing, Forecasting and Prediction of Discrete Time Series*. Courier Corporation.

Burghardt, G., Belton, T. M., Lane, M., and Papa, J. (1994). *The Treasury Bond Basis*. Irwin Professional Publishing.

Chaboud, A. P., Chiquoine, B., Hjalmarsson, E., and Vega, C. (2014). Rise of the machines: Algorithmic trading in the foreign exchange market. *Journal of Finance*, 69(5):2045–2084.

Christensen, K., Podolskij, M., and Vetter, M. (2012). On covariation estimation for

multivariate continuous Itô semimartingales with noise in non-synchronous observation schemes. *Available at SSRN 2050642*.

Clark, P. K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, 41(1):135–55.

CME (2013 (accessed March 14, 2013)). *CME Eurodollar futures contracts*. `http://www.cmegroup.com/trading/interest-rates/stir/eurodollar.html`.

Cont, R. and Tankov, P. (2003). *Financial Modelling with Jump Processes*, volume 2. Chapman & Hall/CRC.

Corsi, F., Pirino, D., and Renò, R. (2010). Threshold bipower variation and the impact of jumps on volatility forecasting. *Journal of Econometrics*, 159(2):276–288.

Corsi, F., Zumbach, G., Muller, U. A., and Dacorogna, M. M. (2001). Consistent high-precision volatility from high-frequency data. *Economic Notes*, 30(2):183–204.

Cox, D. R. and Isham, V. (1980). *Point processes*, volume 12. CRC Press.

Cressie, N. A. (1993). *Statistics for Spatial Data, revised edition*. Wiley, New York.

Dacorogna, M., Gençay, R., Müller, U., Olsen, R., and Pictet, O. (2001). *An Introduction to High-frequency Finance*. Academic Press: San Diego, CA.

Daley, D. J. and Vere-Jones, D. (2007). *An Introduction to the Theory of Point Processes: Volumes I and II*. Springer Science & Business Media.

Davis, M. (1997). *Linear Estimation and Stochastic Control*. Chapman and Hall.

de Jong, F. and Nijman, T. (1997). High frequency analysis of lead-lag relationships between financial markets. *Journal of Empirical Finance*, 4(2):259–277.

Debelle, G. (2011). High-frequency trading in the foreign exchange market. *Study Group Report, Bank of International Settlements*.

Delbaen, F. and Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing. *Mathematische Annalen*, 300(1):463–520.

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

Dionne, G., Duchesne, P., and Pacurar, M. (2009). Intraday value at risk (IVaR) using tick-by-tick data with application to the Toronto stock exchange. *Journal of Empirical Finance*, 16(5):777–792.

Dovonon, P., Goncalves, S., and Meddahi, N. (2012). Bootstrapping realized multivariate volatility measures. *Journal of Econometrics*.

Dumas, B., Fleming, J., and Whaley, R. (2002). Implied volatility functions: Empirical tests. *The Journal of Finance*, 53(6):2059–2106.

Easley, D. and O'hara, M. (1992). Time and the process of security price adjustment. *The Journal of finance*, 47(2):577–605.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007.

Engle, R. F. and Russell, J. R. (2002). Analysis of high frequency data. Technical report, Working Paper, New York University and University of Chicago.

Epps, T. (1979). Comovements in stock prices in the very short run. *Journal of the American Statistical Association*, 74(366a):291–298.

Evans, M. and Lyons, R. (2005). Do currency markets absorb news quickly? *Journal of International Money and Finance*, 24(2):197–217.

Fan, J., Li, Y., and Yu, K. (2012). Vast volatility matrix estimation using high-frequency data for portfolio selection. *Journal of the American Statistical Association*, 107(497):412–428.

Fitzgerald, D. (1993). *Financial Futures*. Prentice Hall, New York.

Fortune, P. (1993). Stock market crashes: What have we learned from October 1987? *New England Economic Review*, March/April:3–23.

Foster, D. and Nelson, D. (1996). Continuous record asymptotics for rolling sample variance estimators. *Econometrica*, 64(1):139–174.

Galbraith, J. K. (2009). *The Great Crash 1929*. Mariner Books.

Gardner, E. S. (2006). Exponential smoothing: The state of the art:II. *International Journal of Forecasting*, 22(4):637–666.

Giot, P. (2005). Relationships between implied volatility indexes and stock index returns. *Journal of Portfolio Management*, 31(3):92–10.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438.

Griffin, J. and Oomen, R. (2011). Covariance measurement in the presence of non-synchronous trading and market microstructure noise. *Journal of Econometrics*, 160(1):58–68.

Griffin, J. E. and Oomen, R. C. (2008). Sampling returns for realized variance calculations: tick time or transaction time? *Econometric Reviews*, 27(1-3):230–253.

Hansen, P. R. and Lunde, A. (2006). Realized variance and market microstructure noise. *Journal of Business & Economic Statistics*, 24(2):127–161.

Harris, L. (2002). *Trading and Exchanges: Market Microstructure for Practitioners*. Oxford University Press, USA.

Hasbrouck, J. (2007). *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*. Oxford University Press, USA.

Hasbrouck, J. and Saar, G. (2009). Technology and liquidity provision: The blurring of traditional definitions. *Journal of financial Markets*, 12(2):143–172.

Hasbrouck, J. and Saar, G. (2013). Low-latency trading. *Journal of Financial Markets*, 16(4):646–679.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag.

Hayashi, T. and Yoshida, N. (2005). On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, pages 359–379.

Hayashi, T. and Yoshida, N. (2006). Nonsynchronously observed diffusions and covariance estimation. *Preprint, Kyoto University*.

Hayashi, T. and Yoshida, N. (2008). Asymptotic normality of a covariance estimator for nonsynchronously observed diffusion processes. *Annals of the Institute of Statistical Mathematics*, 60(2):367–406.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoffmann, M., Rosenbaum, M., Yoshida, N., et al. (2013). Estimation of the lead-lag parameter from non-synchronous data. *Bernoulli*, 19(2):426–461.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5–10.

Hull, J. (2009). Options, futures, and other derivatives. *Prentice Hall series in finance*.

Huth, N. and Abergel, F. (2014). High frequency lead/lag relationships - empirical facts. *Journal of Empirical Finance*, 26:41–58.

Jacod, J., Li, Y., Mykland, P. A., Podolskij, M., and Vetter, M. (2009). Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic Processes and their Applications*, 119(7):2249–2276.

Jegadeesh, N. and Pennacchi, G. G. (1996). The behavior of interest rates implied by the term structure of eurodollar futures. *Journal of Money, Credit and Banking*, pages 426–446.

King, M. R., Osler, C. L., and Rime, D. (2011). *Foreign exchange market structure, players and evolution*. Norges Bank working paper.

Korreng, M. D. (2010). UTC time transfer for high frequency trading using IS-95 CDMA base station transmissions and IEEE-1588 precision time protocol. *42nd Annual Precise Time and Time Interval (PTTI) Meeting*.

Leber, C., Geib, B., and Litz, H. (2011). High frequency trading acceleration using FPGAs. In *Field Programmable Logic and Applications (FPL), 2011 International Conference on*, pages 317–322. IEEE.

Love, R. (2005). *A Microstructural Analysis of the Effects of News on Order Flows and on Price Discovery in Foreign Exchange Markets*. PhD thesis, University of London.

Madhavan, A. (2000). Market microstructure: A survey. *Journal of Financial Markets*, 3(3):205–258.

Mastromatteo, I., Marsili, M., and Zoi, P. (2011). Financial correlations at ultra-high frequency: theoretical models and empirical estimation. *The European Physical Journal B-Condensed Matter and Complex Systems*, 80(2):243–253.

McAleer, M. and Medeiros, M. C. (2008). Realized volatility: A review. *Econometric Reviews*, 27(1-3):10–45.

Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, 8(4):323–361.

Miller, M. H. (1986). Financial innovation: The last twenty years and the next. *Journal of Financial and Quantitative Analysis*, 21(04):459–471.

Münnix, M., Schäfer, R., and Guhr, T. (2010a). Compensating asynchrony effects in the calculation of financial correlations. *Physica A: Statistical Mechanics and its Applications*, 389(4):767–779.

Münnix, M., Schäfer, R., and Guhr, T. (2010b). Impact of the tick-size on financial returns and correlations. *Physica A: Statistical Mechanics and its Applications*, 389(21):4828–4843.

Mykland, P. and Zhang, L. (2008). Inference for volatility-type objects and implications for hedging. *Statistics and Its Interface*, 1:255–278.

Mykland, P. and Zhang, L. (2012). The econometrics of high frequency data. *Statistical Methods for Stochastic Differential Equations*, pages 109–190.

Mykland, P. A. and Zhang, L. (2006). ANOVA for diffusions and Itō processes. *The Annals of Statistics*, 34(4):1931–1963.

Newey, W. and West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27.

Øksendal, B. (2003). *Stochastic Differential Equations*. Springer.

Pigorsch, C., Pigorsch, U., and Popov, I. (2012). Volatility estimation based on high-frequency data. In Duan, J.-C., Härdle, W. K., and Gentle, J. E., editors, *Handbook of Computational Finance*, pages 335–369. Springer.

Podolskij, M., Vetter, M., et al. (2009). Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps. *Bernoulli*, 15(3):634–658.

Poon, S.-H. and Granger, C. (2003). Forecasting volatility in financial markets. *Journal of Economic Literature*, 41:478–539.

Precup, O. and Iori, G. (2007). Cross-correlation measures in the high-frequency domain. *European Journal of Finance*, 13(4):319–331.

Puntanen, S. and Styan, G. (2005). Historical introduction: Issai Schur and the early development of the Schur complement. *The Schur Complement and Its Applications*, pages 1–16.

Rao, C. R. (2009). *Linear Statistical Inference and Its Applications*. John Wiley & Sons.

Rencher, A. C. and Christensen, W. F. (2012). *Methods of Multivariate Analysis*. John Wiley & Sons.

Renò, R. (2003). A closer look at the Epps effect. *International Journal of Theoretical and Applied Finance*, 6(01):87–102.

Revuz, D. and Yor, M. (2004). *Continuous Martingales and Brownian Motion*. Springer.

Robert, C. Y. and Rosenbaum, M. (2012). Volatility and covariation estimation when microstructure noise and trading times are endogenous. *Mathematical Finance*, 22(1):133–164.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *The Journal of Finance*, 39(4):1127–1139.

Schoutens, W. (2003). *Lévy processes in Finance*. Wiley.

Scott-Quinn, B. (2012). *Commercial and Investment Banking and the International Credit and Capital Markets*. Palgrave Macmillan.

Shephard, N. (2015). Martingale unobserved component models. In Koopman, S. J. and Shephard, N., editors, *Unobserved Components and Time Series Econometrics*, chapter 10, pages 218–249. Oxford University Press.

Tankov, P. and Cont, R. (2004). *Financial Modelling with Jump Processes*. Chapman & Hall/CRC.

Taylor, J. W. (2004). Smooth transition exponential smoothing. *Journal of Forecasting*, 23(6):385–404.

Thompson, G. (2002). Optimal trading of an asset driven by a hidden markov process in the presence of fixed transaction costs. *Research papers in management studies - University of Cambridge, Judge Institute of Management Studies*.

Todorov, V. and Tauchen, G. (2011). Volatility jumps. *Journal of Business & Economic Statistics*, 29(3):356–371.

Tóth, B. (2008). *From multi-agent modeling to microscopic market dynamics: A statistical physics approach*. PhD thesis, Budapest University of Technology and Economics.

Tóth, B. and Kertész, J. (2009a). Accurate estimator of correlations between asynchronous signals. *Physica A: Statistical Mechanics and its Applications*, 388(8):1696–1705.

Tóth, B. and Kertész, J. (2009b). The Epps effect revisited. *Quantitative Finance*, 9(7):793–802.

Tóth, B., Tóth, B., and Kertész, J. (2007). Modeling the Epps effect of cross correlations in asset prices. *arXiv preprint arXiv:0704.3798*.

Trader, A. (2014a). HFT in my backyard (i). `http://www.amsterdamtrader.com/2014/09/hft-in-my-backyard.html`. Accessed: 2014-09-30.

Trader, A. (2014b). HFT in my backyard (iii). `http://sniperinmahwah.wordpress.com/2014/10/02/hft-in-my-backyard-iii`. Accessed: 2014-09-30.

Trigg, D. and Leach, A. (1995). Exponential smoothing with an adaptive response rate. *Management science: an anthology*, 2(1):425.

Veraart, A. and Winkel, M. (2010). Time change. In Cont, R., editor, *Encyclopedia of Quantitative Finance*, pages 1812–1816. Wiley.

Voev, V. and Lunde, A. (2007). Integrated covariance estimation using high-frequency data in the presence of noise. *Journal of Financial Econometrics*, 5(1):68–104.

Wang, Y. and Zou, J. (2010). Vast volatility matrix estimation for high-frequency financial data. *The Annals of Statistics*, 38(2):943–978.

West, M. and Harrison, P. (1989). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag.

Whaley, R. E. (2000). The investor fear gauge. *Journal of Portfolio Management*, 26(3):12–17.

Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli*, 12(6):1019–1043.

Zhang, L. (2011). Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics*, 160(1):33–47.

Zhang, L., Mykland, P., and Aït-Sahalia, Y. (2005). A tale of two time scales: determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472):1394–1411.

Zhou, B. (1995). Estimating the covariance matrix from unsynchronized high frequency financial data. Technical report, Alfred P. Sloan School of Management, Massachusetts Institute of Technology. `http://hdl.handle.net/1721.1/47553`.

Zhou, B. (1996). High-frequency data and volatility in foreign-exchange rates. *Journal of Business & Economic Statistics*, 14(1):45–52.

Zimmermann, H., Drobetz, W., and Oertmann, P. (2002). *Global Asset Allocation: New Methods and Applications*. John Wiley & Sons.