A Survey of Some Research Agendas developed in response to the availability of High Frequency Financial Data Sets

By
D.E.Allen
School of Accounting, Finance and Economics
Edith Cowan University

Overview of presentation

- I shall commence by looking at some potential statistical issues that could be associated with the use of large data sets.
- Then provide a brief introduction to research in market microstructure and mention some market microstructure data bases.
- I shall consider some of the econometric issues related to the use of high frequency data sets.

Overview of presentation

- Then briefly introduce work on ACD models
- This will be followed by a review of work on realised volatility.
- This is not an exhaustive coverage of work using high frequency financial data sets but these two areas have been the focus of particular recent attention.

Some potential statistical issues related to the use of large data sets

- In the next few slides I shall briefly flag a few potential issues that Clive Granger (1998) suggested as likely to be associated with the use of large datasets.

C.W. Granger “EXTRACTING INFORMATION FROM MEGA-PANELS AND HIGH-FREQUENCY DATA” , Department of Economics, UCSD Working paper 98.01.

- Prescient though these observations may be, the nature of high frequency financial data sets has brought its own set of unique related statistical issues.

Clive Granger (1998) Extracting information from Mega-Panels and high frequency data

- Granger (1998) in a UCSD working paper suggested that
  - Economic statisticians are in the midst of a regime shift from a period of scarce data in most situations to one in which some data sets are comparatively enormous.
  - If we have a panel in which Q variables are measured for each of N regions (or agents) at each of T time periods, this gives a data set $j = 1, ..., N$; $t = 1, ..., T$ where $\underline{X}$ is a vector with $\underline{X}_{jt}$ Q components.

    A mega-panel may well have Q about 100 or so, N in the thousands and T in the millions, producing a total set of size A^{10}, for some A.

Granger (1998)

- If one moves from the classical situation of having small or data sets of limited size to having substantial amounts of data to analyze, there are three basic questions:
- (i) which of our standard procedures and concepts should be discarded?
- (ii) will some of our standard procedures evolve and then perform better with *n* large?
- (iii) what new techniques need to be developed?

Granger (1998)

- **Concepts And Procedures to Discard or Improve**
  - There are some familiar concepts and techniques that are candidates to be discarded when *n* is large.
  - *Small-sample Adjustments.*
    - There will be no need to worry about whether to use or in the sample 1/*n* 1/(*n*-1) in estimates of variance, or use R_C^2 rather than R^2
    - *Jackknife.* Or any other bias adjustment technique 0(1/n)

Granger (1998)

- *t-tests, F-tests, chi-squared tests.* Any test which has a degree or degrees of freedom that depend on *n* will now take simple asymptotic forms, with *t*- and chi-squared distributions going to normal and *F* to unity.
- *The Higher Moments.*
  - Skewness and Kurtosis have never been much use and it is doubtful if we will be interested in fitting Pearson curves. The mean might survive for old times sake but the variance may not, as it is too strongly related to the normal distribution

Granger (1998)

- *95%, 99% Confidence Intervals.*
  - In simple cases, where confidence intervals are they will be effectively zero so $O(n^{-1})$ that virtually any parsimonious parametric model will produce a very low *p*-value and will be strongly rejected by any standard hypothesis test using the usual confidence intervals. Virtually all specific null hypotheses will be rejected using present standards. It will probably be necessary to replace the concept of statistical significance with some measure of economic significance.

Granger (1998)

- *Model Selection Criteria, BIC, AIC, etc.*
  - Parsimony may or may not be a desirable property for a model but it is not required when data is plentiful. The terms in the criteria that depend on *n* will dominate its size and so searching over the number of parameters to use will be ineffective.

Granger (1998)

- *Bootstrap.*
  - The bootstrap extends the sample size in a somewhat artificial way. If *n* is large enough there is no need for such an extension. Many of the questions being tackled by the use of the bootstrap do not arise with *n* large. This will be true of many other simulation techniques. There will be little use for artificial data when real data is plentiful.

Granger (1998)

- *Bayesian Estimation.*
  - The use of the Bayesian prior in a likelihood estimation procedure is essentially an extension of the data set if the prior is correct, in some sense. As *n* becomes large the data generated component of the likelihood will eventually dominate the prior and the estimates achieved will not be affected by the choice of prior.

Granger (1998)

- **Asymptotics.**
  - For large *n* the natural reaction of a statistician will almost certainly be that various parts of asymptotic theory will become immediately relevant. One should expect that the law of large numbers, the law of the iterated logarithm and various limit theorems, particularly central limit theorems, should be applicable, for example.

Granger (1998)

- Granger cautions that in time series analysis more data of higher frequency may not given better estimates of low frequency components such as seasonalities, cycles etc.
- Notwithstanding the above there are also a further set of issues associated with the nature of high frequency financial data that Engle and Russell started to address in (1996).

Market Microstructure Research: What is it?

- High Frequency data is typically employed in market microstructure research.
- Market microstructure is the study of the trading mechanisms used for financial securities.
- The term “market microstructure” was popularised in a paper featuring this title by Garman
  - Garman, Mark, 1976, Market microstructure. *Journal of Financial Economics* 3, 257-275

Work in this area departs from the usual approaches to the theory of exchange used by economists by:

(1) making the assumption of asynchronous, temporally discrete market activities on the part of market agents and

(2) adopting a viewpoint which treats the temporal microstructure, i.e., moment-to-moment aggregate exchange behaviour, as an important descriptive aspect of such markets.

Market Microstructure Research

- Microstructure analyses typically touch on one or more of the following aspects of trade.
  - We generally assume that the security value comprises private and common components. Usually relating to relevant information sets but after trading takes place market valuations are common to every participant.
- *Mechanisms in Economic Settings*
  - Microstructure analyses are usually very specific about the mechanism or protocol, used to accomplish trade. One common and important mechanism is the continuous limit order market. The full range, though, includes search, bargaining, auctions, dealer markets, and a variety of derivative markets. These markets and mechanisms frequently may operate in parallel.

Market Microstructure Research

- Multiple characterization of prices
  - The market-clearing price, in the Walrasian tatonnement sense, rarely appears in microstructure analyses. At a single instant there may be many prices, depending on direction (buying or selling),
  - the speed with which the trade must be accomplished, sometimes on the agent's identity or other attribute, and the agent's relationship to the counterparty (as well as, of course, quantity).
  - Some prices (like bids and offers) may be hypothetical and prospective.

Market Microstructure Research

- **Liquidity**
  - Security markets are often characterized by their “liquidity.” Precise definitions only exist in the contexts of particular models.
  - The notion of liquidity impounds the usual economic concept of elasticity. In a liquid market, a small shift in demand or supply does not result in a large price change. Liquidity also refers to the cost of trading, something distinct from the price of the security being bought or sold. Liquid markets have low trading costs.
  - Liquidity also has dynamic attributes. In a liquid market, accomplishing a purchase or sale over a short horizon does not cost appreciably more than spreading the trades over a longer interval.

Market Microstructure Research

- Liquidity
  - is sometimes defined as “depth, breadth, and resiliency.” In a deep market if we look a little above the “current market price”, there is a large incremental quantity available for sale.
  - Below the current price, there is a large incremental quantity that is sought by one or more buyers. A broad market has many participants, none of whom is presumed to exert significant market power.
  - In a resilient market, the price effects that are associated with the trading process (as opposed to the “fundamental” valuations) are small and die out quickly.

Market Microstructure Research

- Transparency
  - Transparency is a market attribute that refers to how much information market participants (and potential participants) possess about the trading process.
  - Electronic markets that communicate in real time the bids and offers of buyers and sellers, and the prices of executed trades, are considered highly transparent.
  - Dealer markets, on the other hand, often have no publicly visible bids or offers, nor any trade reporting, and are therefore usually considered opaque.

Market Microstructure Research

- Typical issues addressed in market microstructure research
  - What are optimal trading strategies for typical trading problems?
  - Exactly how is information impounded in prices?
  - How do we enhance the information aggregation process?
  - How do we avoid market failures?
  - What sort of trading arrangements maximize efficiency?
  - What is the trade-off between “fairness” and efficiency?
  - How is market structure related to the valuation of securities?
  - What can market/trading data tell us about the informational environment of the firm?
  - What can market/trading data tell us about long-term risk?

Development of High Frequency data bases McInish and Wood

- Tom McInish and Robert Wood started compiling high frequency US data sets in the early 1980's.
  - Wood, McInish, and Ord (1985) "An Investigation of Transactions Data for NYSE Stocks"
  - "I had obtained transactions data for listed and Nasdaq stocks that included both trades and quotes from a consulting contract--quote data added greatly to the understanding of microstructure issues. Hence, I submitted an NSF grant proposal to construct a trade and quote database using Securities Industry Automation Corporation (SIAC) data, which included transactions for both listed and Nasdaq stocks. ISSM moved to the University of Memphis when I relocated there in 1990".
  - See **Market Microstructure Research Databases: History and Projections.(Statistical Data Included).** Robert A. WOOD. ***Journal of Business & Economic Statistics*** 18.2 (April 2000): p140.

High frequency data sets

- The TAQ Database
  - In 1993 the NYSE began releasing the TAQ database that provides trades and quotes in separate semiflat file format. An ancillary file contains ticker symbols and dates with pointers to trade and quote files, which are sorted by symbol within days. Several of the sparse data items that had been contained in the ISSM database were eliminated: Furthermore, the TAQ database does not include reliable data for shares outstanding and stock and cash distributions that had been distributed with the ISSM database. The TAQ data are not error filtered.

High Frequency Data sets

- The TORQ database contains transactions, quotes, order processing data and audit trail data for a sample of 144 NYSE stocks for the three months November, 1990 through January 1991. Made available in 1992.

SIRCA's Databases

- UWA and ECU are members of SIRCA (The Securities Industry Research Centre of the Asia Pacific) see: http://www.sirca.org.au/
  - SIRCA is a not-for-profit financial services research organisation involving twenty-six collaborating universities across Australia and New Zealand
  - Through the support of its university members, and in partnership with industry and government, SIRCA has become the region's leading specialist financial research infrastructure provider

Sirca's data sets

1. Data available via SIRCA

Securities Data

ASX Daily Data (previously known as CRD) provides historical daily trading information including total volume and value and the open, high, low and close price for each stock listed on the ASX. This data is available for the period 19 February 1990 to the present. This data also includes closing index values on a daily basis for the period commencing 31 December 1979.

ASX Intra-day Data provides historical details of all individual trades and orders placed on Stock Exchange Automated Trading System (SEATS). Each trade and order record includes details of the price, volume and disguised broker identifiers time stamped to the nearest one hundredth of a second. This data is available for the period 19 February 1990 to the present.

S&P/ASX Index Data provides details of intra-day index values and the stocks included in each index. Intra-day index values are available from 2 December 1991 to the present. Details of the stocks included in the index are available from 22 December 1993.

Company Announcements (Signal G) Data provides details of company announcements lodged with the ASX pursuant to the ASX Listing Rules. Full text announcements are available for the period 2 September 1992 to the present. Announcement headings (without full text) is also available for the period 3 December 1991 to 2 September 1992.

ASX Holdings Data (CHESS) provides details of daily changes in shareholdings. The database dates back to November 1995.

Sirca's data sets (contd)

- Reuters Data provides intra-day trade and quote information for over 240 markets around the world. Data coverage begins January 1996.

Reuters data available varies somewhat from market to market. However, the following data is generally available.

Exchange Traded Markets
For a detailed matrix of Exchange Traded For every quote change and trade the following fields are recorded:

FieldDescriptionNotes

1Reuters Instrument Code(RIC)
2Exchange
3Date
4Time
5Best bid price
6Best ask price
7Trade price
8Trade volume

Index Data
Index values are recorded every 10 to 60 seconds depending on the exchange. The following fields are recorded:

FieldDescriptionNotes
1Reuters Instrument Code (RIC)
2Date
3Time
4Current value

Research issues

- One obvious issue relate to the question of how information is impounded into price series.
  - The market can be viewed as being populated by information traders and liquidity traders.
  - Early theoretical papers utilizing this framework include Glosten and Milgrom (1985) Bid Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Agents, Journal of Financial Economics 14 71-100
  - Easley and O'Hara (1992), Time and the Process of Security Price Adjustment. The Journal of Finance 19, 69-90
  - Copeland and Galai (1983) Information Effects and the Bid-Ask Spread, Journal of Finance 38, 1457-1469
  - and Kyle (1985).
  - A review of this literature can be found in O'Hara (1995). Market Microstructure Theory, Blackwell Publishers

Research issues

- An overview of the predictions of these models is that prices adjust more quickly to reflect private information when the proportion of uninformed traders is higher. Volume is higher, transaction rates are higher when the proportion of uninformed traders is higher.
- The bid ask spread is therefore predicted to be increasing in volume and transaction rates.

Research issues

- Answers to these questions necessarily involve studying the relationship between transaction prices and market characteristics.
- The bid ask spread is often considered a measure of liquidity since it measures the cost of purchasing and then immediately reselling. The wider the spread the higher the cost of transacting. Of course spreads vary both cross sectionally across different stocks as well as temporally.
- Using high frequency data the variation in the spread can be linked to the variation in market characteristics.
- The price impact of a trade can be measured by relating quote revisions to characteristics of a trade. Again with high frequency transactions data the magnitude and possibly the direction of price adjustments can be linked to characteristics of the market.
- Volatility models for high frequency data are clearly relevant here.

Research issues

- In some cases a single asset is traded in multiple markets. An obvious question is in which market does price discovery take place?
- Do price movements in one market precede price movements in the other market?
- If the two series are cointegrated these questions can be addressed in the context of error correction models.
- Another possibility is that informed agents have multiple outlets in which they can exploit their information. For example derivative markets.
- This can be reduced to causality questions. Do price movements in one market precede price movements in the other? Do trades in one market tend to have a greater price impact across both markets?

New difficulties

- We have not yet reached the stages mapped out by Granger (1998) in the use of high frequency financial data sets.
- A whole new set of issues have emerged in relation to the typical characteristics of rthe data sets.

Market Microstructure Research

- **Econometric Issues**
  - Microstructure data are distinctive. Most microstructure series consist of discrete events randomly arranged in continuous time. Within the time series taxonomy, they are formally classified as point processes.
  - Point process characterizations are becomingly increasingly important, and I shall return to this issue later.

Market Microstructure Research

- High Frequency data
  - Microstructure data are often well-ordered. The sequence of observations in the dataset closely corresponds to the sequence in which the economic events actually happened.
  - In contrast, most macroeconomic data are time aggregated, giving rise to simultaneity and uncertainty about the directions of causal effects.
  - The fine temporal resolution of the data used, sometimes described as ultra high frequency, often supports stronger conclusions about causality

Market Microstructure Research

- **Large data sets**
  - Microstructure data samples are typically large in the sense that by most economic standards observations are exceedingly plentiful (10,000 would not be considered unusual).
  - One would not ordinarily question the validity of asymptotic statistical approximations in samples of this size. It is worth emphasizing, though, that the usual asymptotic results apply to correctly specified models, and given the complexity of trading processes some degree of misspecification is almost inevitable.
  - This has some implications I shall return to later.
  - Furthermore, despite the number of observations, the data samples are often small in terms of calendar span, on the order of days or at best months.

Characteristics of high frequency data sets

- With these new data sets come new challenges associated with their analysis.
- Modern data sets may contain tens of thousands of transactions or posted quotes in a single day time stamped to the nearest second.
- The analysis of these data are complicated by irregular temporal spacing, diurnal patterns, price discreteness, and complex often very long lived dependence.

Characteristics of high frequency data

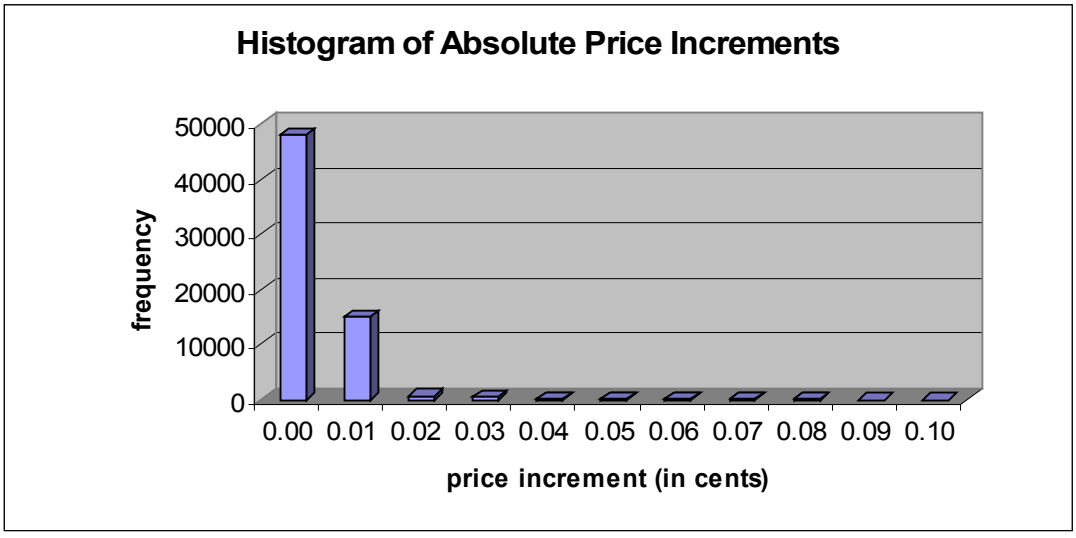
- Irregular temporal spacing
  - virtually all transactions data are inherently irregularly spaced in time. Since most econometric models are specified for fixed intervals this poses an immediate complication. A choice must be made regarding the time intervals over which to analyze the data. If fixed intervals are chosen then some sort of interpolation rule must be used when no transaction occurs exactly at the end of the interval.
  - Alternatively if stochastic intervals are used then the spacing of the data will likely need to be taken into account.
  - The irregular spacing of the data becomes even more complex when dealing with multiple series each with its own transaction rate. Here, interpolation can introduce spurious correlations due to non-synchronous trading.

Characteristics of high frequency data

- Discreteness
  - All economic data is discrete. When viewed over long time horizons he variance of the process is usually quite large relative to the magnitude of the minimum movement.
  - In the case of transaction data, however, this is not the case and for many data sets the transaction price changes take only a handful of values called ticks.
  - There may be market rules that restrict prices to fall on a pre-specified set of values. Price changes must fall on multiples of the smallest allowable price change called a tick.
  - In a market for an actively traded stock it is generally not common for the price to move a large number of ticks from one transaction to another.

Characteristics of high frequency data

- Histogram of absolute price movements in cents for NAB spanning Jan 2 2004 to March 31 2004 (63 trading days and 64 561 durations). Clearly most transactions involve no price change. (I am grateful to Dr Zdravetz Lazarov for providing these statistical summaries)

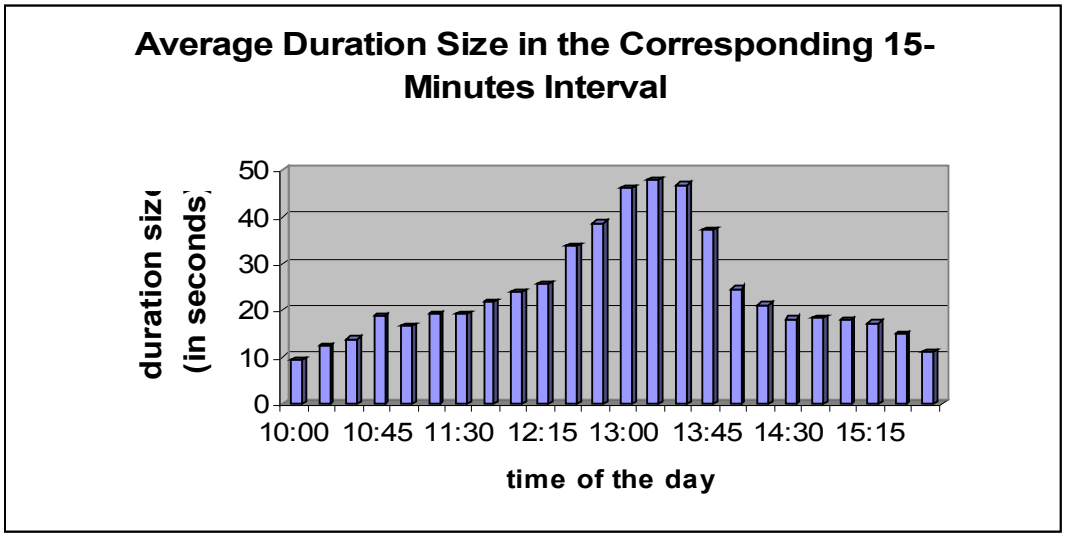


Characteristics of high frequency data

- Diurnal patterns
  - Intraday financial data typically contain very strong diurnal or periodic patterns.
  - For most stock markets volatility, the frequency of trades, volume, and spreads all typically exhibit a U-shaped pattern over the course of the day. For an early reference see McInish and Wood (1992).
  - Volatility is systematically higher near the open and generally just prior to the close.
  - Volume and spreads have a similar pattern.
  - The time between trades, or durations, tend to be shortest near the open and just prior to the close.
  - This was first documented in Engle and Russell (1998).
  - A slide for NAB displaying these effects follows

Characteristics of high frequency data

- There is an obvious diurnal pattern to durations through the day. Shorter durations at the open and the close. See pattern for NAB below.



Characteristics of high frequency data

- Temporal dependence
  - high frequency financial returns data typically display strong dependence. The dependence is largely the result of price discreteness and the fact that there is often a spread between the price paid by buyer and seller initiated trades.
  - This is typically referred to as bid-ask bounce and is responsible for the large first order negative autocorrelation.

Characteristics of high frequency data

- Correlations of transaction prices for NAB, reflects the impact of bid/ask bounce

transaction price changes

Lag	Autocorrelation	Q-stat	p-value
1	-0.316	6425.0	0.000
2	-0.179	8499.2	0.000
3	0.000	8499.2	0.000
4	-0.001	8499.2	0.000
5	0.000	8499.2	0.000
6	0.000	8499.2	0.000
7	0.001	8499.3	0.000
8	-0.001	8499.3	0.000
9	0.000	8499.3	0.000
10	0.001	8499.3	0.000
11	0.000	8499.3	0.000
12	0.001	8499.4	0.000
13	-0.001	8499.5	0.000
14	0.001	8499.5	0.000
15	0.000	8499.5	0.000
16	-0.001	8499.6	0.000
17	0.001	8499.7	0.000
18	-0.002	8500.0	0.000

Characteristics of high frequency data

- The correlations of mid quote prices for NAB. Remove impact of bid/ask bounce

mid-quote price changes

Lag	Autocorrelation	Q-stat	p-value
1	-0.009	4.756	0.029
2	0.000	4.757	0.093
3	0.000	4.757	0.190
4	0.000	4.760	0.313
5	-0.001	4.780	0.443
6	0.000	4.780	0.572
7	0.000	4.785	0.686
8	0.000	4.795	0.779
9	0.000	4.796	0.852
10	0.000	4.799	0.904
11	-0.001	4.825	0.939
12	0.000	4.826	0.964
13	0.000	4.829	0.979
14	0.000	4.830	0.988
15	0.000	4.830	0.993
16	0.000	4.831	0.997
17	-0.002	4.993	0.998
18	0.002	5.155	0.999
19	0.000	5.168	0.999
20	0.000	5.177	1.000

Characteristics of high frequency data

- Autocorrelations of absolute mid quote price changes

absolute mid-quote price cahnes

Lag	Autocorrelation	Q-stat	p-value
1	0.009	4.968	0.026
2	0.000	4.969	0.083
3	0.000	4.976	0.174
4	0.000	4.978	0.290
5	0.001	5.019	0.414
6	0.000	5.029	0.540
7	0.000	5.029	0.656
8	0.000	5.029	0.754
9	0.000	5.040	0.831
10	0.000	5.043	0.888
11	0.000	5.043	0.929
12	0.000	5.054	0.956
13	0.000	5.065	0.974
14	0.000	5.066	0.985
15	0.000	5.072	0.992
16	0.000	5.073	0.995
17	0.002	5.222	0.997

Characteristics of high frequency data

- Absolute mid point price changes scaled by square root of the duration

Lag	Autocorrelation	Q-stat	p-value
1	0.020	25.973	0.000
2	0.002	26.127	0.000
3	0.001	26.219	0.000
4	0.001	26.245	0.000
5	0.002	26.462	0.000
6	0.002	26.643	0.000
7	0.001	26.674	0.000
8	0.000	26.688	0.001
9	0.001	26.773	0.002
10	0.001	26.795	0.003
11	0.001	26.837	0.005
12	0.001	26.977	0.008
13	0.000	26.991	0.012
14	0.001	27.063	0.019
15	0.000	27.070	0.028
16	0.000	27.086	0.041
17	0.005	28.449	0.040
18	0.007	31.366	0.026
19	0.001	31.497	0.036
20	0.001	31.540	0.048

Characteristics of high frequency data

- Other factors leading to dependence in price changes include traders breaking large orders up into a sequence of smaller orders in hopes of transacting at a better price overall. These sequences of buys or sells can lead to a sequence of transactions that move the price in the same direction. Hence at longer horizons we sometimes find positive autocorrelations.

Characteristics of high frequency data

- Autocorrelations in durations

duration

Lag	Autocorrelation	Q-stat	p-value
1	0.153	1509.2	0.000
2	0.134	2664.6	0.000
3	0.130	3746.9	0.000
4	0.136	4938.9	0.000
5	0.123	5913.5	0.000
6	0.125	6913.4	0.000
7	0.122	7870.3	0.000
8	0.124	8856.2	0.000
9	0.116	9722.6	0.000
10	0.110	10506.0	0.000
11	0.113	11325.0	0.000
12	0.116	12190.0	0.000
13	0.119	13098.0	0.000
14	0.123	14075.0	0.000
15	0.119	14982.0	0.000
16	0.115	15830.0	0.000

Characteristics of high frequency data

- Autocorrelations in volumes traded

volume

Lag	Autocorrelation	Q-stat	p-value
1	0.034	72.6	0.000
2	0.043	192.4	0.000
3	0.021	221.8	0.000
4	0.011	229.2	0.000
5	0.015	244.4	0.000
6	0.016	261.6	0.000
7	0.018	283.4	0.000
8	0.008	287.2	0.000
9	0.011	295.1	0.000
10	0.014	307.1	0.000
11	0.025	348.1	0.000
12	0.027	396.8	0.000
13	0.041	505.7	0.000
14	0.010	512.1	0.000
15	0.010	519.1	0.000
16	0.009	524.6	0.000
17	0.010	531.4	0.000
18	0.011	538.8	0.000
19	0.010	544.7	0.000
20	0.010	551.5	0.000

Econometric issues

- The following section draws heavily on a survey by Engle and Russell (2004) “Analysis of High Frequency Financial Data”

- They first suggested financial econometric models capable of capturing some of these characteristics.

- formally, let t_1, t_2, ..., t_i, ... denote a sequence of strictly increasing random variables corresponding to event arrival times such as transactions. Jointly, these arrival times are referred to as a point process.
- We can use a counting process N(t) which is simply the number of event arrivals that have occurred at or prior to time t. This will be a step function with unit increments at each arrival time.

Econometric issues

- If there is additional information associated with the arrival times then the process is referred to as a marked point process.
- Hence, if the marks associated with the ith arrival time are denoted by an M-dimensional vector y_i then the information associated with the ith event is summarized by its arrival time and the value of the marks $[t_i, y_i]$.

Econometric issues

- This information might relate to volumes, spreads etc.,
- Depending on the issue being investigated; either the arrival time, the marks, or both may be of interest. Often hypotheses can be couched in the framework of conditional expectations of future values.
- If we denote the filtration of arrival times and marks at the time of the ith event arrival by

  $bt_i = \{t_i, ti_{-1}, ..., t_0\}$ and

  $by_i = \{y_i, yi_{-1}, ...,y_0\}$ respectively.

  The probability structure for the dynamics associated with a stationary, marked point process is can be completely characterized and conveniently expressed as the joint distribution of marks and arrival times given the filtration of past arrival times and marks:

$$f\left(t_{N(t)+1}, y_{N(t)+1} \middle| \widehat{t}_{N(t)}, \widehat{y}_{N(t)}\right) \qquad (1)$$

Econometric issues

- It may be the case that the waiting time until the next mark is important.

$$f_t\left(t_{N(t)+1}|\widehat{t}_{N(t)},\widehat{y}_{N(t)}\right)=\int f\left(t_{N(t)+1},y|\widehat{t}_{N(t)},\widehat{y}_{N(t)}\right)dy \qquad (2)$$

- This is simply a point process where the arrival times may depend on the past arrival times and the past marks.

Econometric issues

- On the other hand, other economic hypotheses may involve the dynamics of the marks such as models for the spread, or prices.
- We might wish to model or forecasting the value for the next mark, regardless of when it occurs, given the filtration of the joint process. This is given by

$$f_y\left(y_{N(t)+1}|\widehat{t}_{N(t)},\widehat{y}_{N(t)}\right)=\int f\left(t,y_{N(t)+1}|\widehat{t}_{N(t)},\widehat{y}_{N(t)}\right)dt \qquad (3)$$

Econometric issues

- Here, the information set is updated at each event arrival time and we refer to such models as event time or tick time models of the marks.
- Yet another alternative approach is to model the value of the mark to be at some future time t + τ (τ > 0) given the filtration at time t. That is

$$g\left(y_{N(t+\tau)} | \widehat{t}_{N(t)}, \widehat{y}_{N(t)}\right) \qquad (4)$$

Econometric issues

- Here the conditional distribution associated with the mark over a fixed time interval is the object of interest.
- A final approach taken in the literature is to study the distribution of the length of time it will take for a particular type of event, defined by the mark, to occur. For example, one might want to know how long will it take for the price to move by more than a specified amount, or how long will it take for a set amount of volume to be transacted. This can be expressed

$$g\left(t+\tau_{\min}|\widehat{t}_{N(t)},\widehat{y}_{N(t)}\right) \tag{5}$$

Econometric issues

– Simple point processes

- A point process is referred to as a simple point process if as a time interval goes to zero, the probability of multiple events occurring over that time interval can be made an arbitrarily small fraction of the probability of a single event occurring. A convenient way is to use a homogenous Poisson process to characterise the intensity

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr\left(N(t + \Delta t > N(t)\right)}{\Delta t} \tag{6}$$

Econometric issues

- A point process that evolves with after-effects can be conveniently described using the conditional intensity function which specifies the instantaneous probability of an event arrival conditional upon filtration of event arrival times. That is, the conditional intensity is given

$$\lambda(t|N(t),t_{i-1},t_{i-2},...,t_0)=\lim_{\Delta t\to 0}\frac{P\left(N(t+\Delta t)>N(t)|N(t),t_{N(t)},t_{N(t)-1},...,t_0\right)}{\Delta t} \quad (7)$$

Econometric issues

- The conditional intensity function associated with any single waiting time has traditionally been called a hazard function in the econometrics literature.
- Here, however, the intensity function is defined as a function of t across multiple events,

The ACD Model

- Engle and Russell [1996] propose the Autoregressive Conditional Duration (ACD) model which is particularly well suited for high frequency financial data.
- This parameterization is most easily expressed in terms of the waiting times between events.
- Let $x_i = t_i - t_{i-1}$ be the interval of time between event arrivals which will be called the duration.
- The distribution of the duration is specified directly conditional on the past durations.

The ACD Model

- The ACD model is then specified by two conditions. Let ψi be the expectation of the duration given the past arrival times:

$$E\left(x_i|x_{i-1},x_{i-2},...,x_1\right)=\psi_i\left(x_{i-1},x_{i-2},...,x_1;\right)=\psi_i \tag{8}$$

- Let

$$x_i=\psi_i\varepsilon_i \tag{9}$$

- Where ε_i are i.i.d.

The ACD Model

- The baseline hazard is given by

$$\lambda_0 = \frac{p(\epsilon;\phi)}{S(\epsilon;\phi)} \tag{10}$$

where $S_0(\epsilon;\phi) = \int_{\epsilon}^{\infty} p(u;\phi)du$ is the survivor function.

The intensity function is given by

$$\lambda(t|N(t), t_{i-1}, t_{i-2}, ..., t_0) = \lambda_0 \left(\frac{t - t_{N(t)-1}}{\psi_{N(t)}} \right) \frac{1}{\psi_{N(t)}}$$

The ACD Model

- Engle and Russell (1998) suggest the following linear parameterization for the expectation:

$$\psi_i = \omega + \sum_{j=1}^{p} \alpha_j x_{i-j} + \sum_{j=1}^{q} \beta_j \psi_{i-j} \tag{12}$$

- Since the conditional expectation of the duration depends on p lags of the duration and q lags of the expected duration this is termed an ACD(p, q) model.

The ACD Model

- Popular choices for the density $p(\epsilon; \varphi)$ include the exponential and the Weibull distributions suggested in Engle and Russell (1998).
- These models are termed the Exponential ACD (EACD) and Weibull ACD (WACD) models respectively.
- The exponential distribution has the property that the baseline hazard is monotonic.
- The Weibull distribution relaxes this assumption and allows for a hump-shaped baseline intensity.

ACD Models

- The ACD(p, q) specification in (12) appears very similar to a ARCH(p, q) models of Engle (1982) and Bollerslev (1986) and indeed the two models share many of the same properties.
- The durations xi follow an ARMA(max(p, q), q).
- Let $\eta_i \equiv x_i - \psi_i$ which is a martingale difference by construction. then

$$x_i = \omega + \sum_{j=1}^{\max(p,q)} (\alpha_j + \beta_j) x_{i-j} - \sum_{j=1}^{q} \beta_j \eta_{i-j} + \eta_i$$

ACD Models

- If α(L) and β(L) denote polynomials in the lag operator of orders p and q respectively then the persistence of the model can be measured by α (1) + β (1) . For most duration data this sum is very close to (but less than) one indicating strong persistence but stationarity.

- The most basic application of the ACD model to financial transactions data is to model the arrival times of trades. In this case it denotes the arrival of the i^{th} transaction and xi denotes the time between the i^{th} and $(i-1)^{th}$ transactions. Engle and Russell (1998) propose using an ACD(2,2) model with Weibull errors to model the arrival times of IBM transactions.
- Engle, Robert and J. Russell, 1998, Autoregressive Conditional Duration: A New Model for Irregularly Spaced Data, Econometrica 66, 5, 1127-1162

ACD Models

- Engle and Russell (1998) suggest including an additional term to account for a diurnal pattern so that the ith duration is given by:

  $x_i = \varphi_i \psi_i \varepsilon_i$

  Now, ψi is the expectation of the duration after partialing out the deterministic pattern and is interpreted as the fraction above or below the average

ACD Models

- There is a close correspondence between ACD models and GARCH models
- we can use standard GARCH software to perform QML estimation of ACD models. This is accomplished by setting the dependent variable equal to the square root of the duration and imposing a conditional mean equation of zero. The resulting parameter values provide consistent estimates of the parameters used to forecast the expected duration.

ACD Models

- There are non-linear forms of the ACD model corresponding to the EGARCH model

- Bauwens and Giot (2000), "The Logarithmic ACD Model: An Application to the bid-ask quotes process of three NYSE stocks", Annales d'conomie et de Statistique, 60, 117-149

  - Suggest a linear model using logs,

$$\ln(\psi_i) = \omega + \sum_{j=1}^{p} \alpha_j \epsilon_{i-j} + \sum_{j=1}^{q} \beta_j \ln(\psi_{i-j})$$

ACD Models

- Engle and Lange (2001) combine the use of price durations with the cumulative signed volume transacted over the price duration to measure liquidity. For each of the US stocks analyzed, the volume quantity associated with each transaction is given a positive sign if it is buyer initiated and negative sign if it is seller.
- This signed volume is then cumulated for each price duration. The cumulative signed volume, referred to as VNET, is the total net volume that can be transacted before inducing a price move hence providing a time varying measure of the depth of the market.
- Engle, Robert and J. Lange, 2001, Predicting VNET; A Model of the Dynamics of Market Depth, Journal of Financial Markets, V4N2, 113-142

Another major use of high frequency financial data

- I shall have to terminate discussion of ACD models and move onto another use of high frequency data.
- This involves the direct use of intraday financial data to better estimate volatility.
- The following slides draw heavily on M. McAleer and M. Medeiros, “Realised Volatility: A Review”, forthcoming *Econometric Reviews* 2007.

Realised volatility

- McAleer and Medeiros note that the search for an adequate framework for the estimation and prediction of the conditional variance of financial assets returns has led to the analysis of high frequency intraday data.
- Merton (1980) noted that the variance over a fixed interval can be estimated arbitrarily, although accurately, as the sum of squared realizations, provided the data are available at a sufficiently high sampling frequency.

Realised volatility

- Andersen and Bollerslev (1998) showed that ex post daily foreign exchange volatility is best measured by aggregating 288 squared five-minute returns.
- The five-minute frequency is a trade-off between accuracy, which is theoretically optimized using the highest possible frequency, and microstructure noise that can arise through the bid-ask bounce, asynchronous trading, infrequent trading, and price discreteness, among other factors

Realised volatility

- Based on the theoretical results of Barndorff-Nielsen and Shephard (2002), Econometric analysis of realised volatility and its use in estimating stochastic volatility models, *Journal of the Royal Statistical Society B*, 64, 253 –
- Andersen, Bollerslev, Diebold and Labys (2003) Modeling and forecasting realized volatility, *Econometrica*, 71, 529 – 626.

  and Meddahi (2002), A theoretical comparison between integrated and realized volatility, *Journal of Applied Econometrics*, 17, 479 – 508.
- several recent studies have documented the properties of realized volatilities constructed from high frequency data.

Realised volatility

- Consider a simple discrete time model in which the daily returns of a given asset are typically characterized as

$$r_t = h_t^{1/2} \eta_t$$

- Where η_t is a sequence of independently and normally distributed random variables with zero mean and unit variance,

Realised volatility

– Suppose that, in a given trading day *t*, the logarithmic prices are observed tick-by-tick. Consider a grid containing all observation points, and set pti, i=1,..n to be the *i*th price observation during day *t*, where n is the total number of observations at day *t*. Furthermore, suppose that

$$\Lambda_t = \left\{ \tau_0, \ldots, \tau_{n_t} \right\}$$

Realised volatility

$$r_{t,i} = h_{t,i}^{1/2} \eta_{t,i}$$

where $\eta_{t,i} \sim \mathrm{NID}\left(0, n_t^{-1}\right)$

$$r_{t,i} = p_{t,i} - p_{t,i-1}$$

is the *i*th intra-period return of day *t* such that

$$r_t = \sum_{i=0}^{n_t} r_{t,i}$$

$$h_t = \frac{1}{n_t} \sum_{i=1}^{n_t} h_{t,i}$$

Realised volatility

Define the information set

$$\Im_{t,i} \equiv \Im\left\{p_{a,b}\right\}_{a=-\infty,b=0}^{a=t,b=i}$$

as the σ-algebra generated by all the information to the *i*th tick in day *t*. Therefore,

$\Im_{t,0}$ is the information set available prior to the start of day *t*. It follows that

$$\mathrm{E}\left(r_t^2 \mid \Im_{t,0}\right) = h_t$$

$$\mathrm{V}\left(r_t^2 \mid \Im_{t,0}\right) = 2h_t^2$$

The *realized variance* is defined as the sum of all available intraday high frequency squared returns given by

Realised volatility

$$RV_t^{(\text{all})} = \sum_{i=0}^{n_t} r_{t,i}^2$$

And the squared daily return can be written as

.

$$r_t^2 = \left(\sum_{i=0}^{n_t} r_{t,i} \right)^2 = \sum_{i=0}^{n_t} r_{t,i}^2 + 2 \sum_{i=0}^{n_t - 1} \sum_{j=i+1}^{n_t} r_{t,i} r_{t,j}$$

Realised volatility

- Such that

$$\mathrm{E}\left(r_t^2 \mid \Im_{t,0}\right) = \mathrm{E}\left(\sum_{i=0}^{n_t} r_{t,i}^2 \middle| \Im_{t,0}\right) + 2\mathrm{E}\left(\sum_{i=0}^{n_t-1}\sum_{j=i+1}^{n_t} r_{t,i}r_{t,j} \middle| \Im_{t,0}\right) = \mathrm{E}\left(RV_t^{(all)} \mid \Im_{t,0}\right) + 2\mathrm{E}\left(\sum_{i=0}^{n_t-1}\sum_{j=i+1}^{n_t} r_{t,i}r_{t,j} \middle| \Im_{t,0}\right)$$

If the intraday returns are uncorrelated, then

$$\mathrm{E}\left(r_t^2 \mid \Im_{t,0}\right) = \mathrm{E}\left(RV_t^{(\mathrm{all})} \mid \Im_{t,0}\right) = h_t$$

As a result, two unbiased estimators for the average day-*t* return variance exist, namely the squared day-*t* return and the realized variance. However, it can be shown that

Realised volatility

$$\mathrm{V}\left(RV_t^{(all)} \mid \Im_{t,0}\right) = \frac{2}{n_t}\sum_{i=0}^{n_t}\frac{h_{t,i}^2}{n_t} < \frac{2}{n_t}\left(\sum_{i=0}^{n_t}\frac{h_{t,i}}{\sqrt{n_t}}\right)^2 = \mathrm{V}\left(r_t^2 \mid \Im_{t,0}\right)$$

as

$$\mathrm{E}\left[\left(\sum_{i=1}^{n_t} h_{t,i}\eta_{t,i}^2\right)^2 \middle| \Im_{t,0}\right] = \frac{3}{n_t^2}\sum_{i=0}^{n_t} h_{t,i}^2 + \frac{2}{n_t^2}\sum_{i=1}^{n_t-1}\sum_{j=i+1}^{n_t} h_{t,i}h_{t,j}$$

Realised volatility

- In short, the average daily returns variance can be estimated more accurately by summing the squared intraday returns rather than calculating the squared daily return. Moreover, when returns are observed at any arbitrary frequency, it is possible to estimate the average daily variance free of measurement error as

$$\lim_{n_t \to \infty} \mathrm{V}\left(RV_t^{(all)} \mid \Im_{t,0} \right) = 0$$

Realised volatility

- The only requirement on the dynamics of the intraday return variance for the above to hold is that

$$\sum_{i=1}^{n_t} h_{t,i}^2 \propto n_t^{1+c}$$

where $0 \leq c < 1$

This result motivates a number of empirical papers

Realised volatility

- The theoretical foundations of these results are derived from a continuous time framework that is based on the theory of quadratic variations.
- Things get more complicated when the existence of market microstructure noise is acknowledged.

Realised volatility

- There are several sampling schemes that can be used, as follows:
  - The most widely used sampling scheme is *calendar time sampling* (CTS), where the intervals are equidistant in calendar time, they could be at every 5 or 15 minutes intervals.
  - On potential problem follows from the fact that there may not be a trade at the required point in time.

Realised volatility

- Another sampling alternative is *transaction time sampling* \where prices are recorded every *m*th transaction.
- The third sampling scheme is known as *business time sampling* (BTS), where the sampling times are chosen such that .
- The last sampling alternative is called *tick time sampling* where prices are recorded at every price change.
- The table on the next slide from McAleer and Medeiros (2006) tabulates the properties of different estimators given various combined assumptions about price processes, noise structures etc

Table 1: Properties of Methods for Estimating the Daily Integrated Variance

Method	Unbiased [1]	Consistent	Price Model	Noise (Time Dependence)	Noise and Efficient Prices
TTSE ZMA (2005)	Yes	Yes	Diffusion	IID	Independent
TTSE AMZ (2005b)	Yes	Yes	Diffusion	Dependent	Independent
Kernel Hansen and Lunde (2006)	Yes	No	Diffusion	Dependent/ IID	Dependent/ Independent
Kernel, BHLS (2006)	Yes	Yes	Diffusion	Dependent/ IID	Dependent/ Independent
Sparse Sampling	No	No	Diffusion	IID	Independent
Optimal frequency selection, Bandi and Russell (2005a, 2006b)	No	No	Diffusion	IID	Independent
Optimal frequency selection Oomen (2006)	No	No	Pure Jump	Dependent	Independent
Kernel Oomen (2005)	Yes	No	Pure Jump	IID	Independent
All available data	No	No	Diffusion	IID	Independent
MA filter HLL (2006)	Yes	Yes	Diffusion	IID	Independent

[1] In Table 1 we consider large sample bias. Some of the estimators, such as TTSE, are biased in small samples but not asymptotically.

Realised volatility

- Clearly the choice of how to sample high frequency data when estimating realised volatility is a complex issue.
- Many problems remain unresolved.
- I have barely scratched the surface of this particular literature and recommend McAleer and Medeiro's (2006) excellent survey.

Conclusion

- It seems that we are no where near to a situation suggested by Granger (1998) in which large data sets bring convenient simplifications to statistical and econometric analysis.
- Indeed, high frequency financial data sets bring their own unique difficulties which are often attached to market microstructure features which reflect the ways in which financial markets operate.
- In this presentation I have briefly examined a couple of areas of research which make use of these extensive data sets and some of the statistical and econometric issues that arise from the use of this type of data.
- This work is new, and particularly in the case of realised volatility, still evolving.