Università degli Studi di Firenze

DOTTORATO DI RICERCA IN
"*Statistica Applicata*"

CICLO XXV

COORDINATORE Prof. Fabio Corradi

Financial Modelling with Ultra-High Frequency Data: Issues and (Possible) Solutions.

Settore Scientifico Disciplinare SECS-S/01

Dottorando
Dott. Francesco Calvori

(firma)

Tutore
Prof. Giampiero M. Gallo

(firma)

Anni 2010/2012

To Francesca

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

Introduction

This thesis is composed of three different essays, all related to the issue of volatility modelling in financial econometrics. The leading thread is the use of ultra-high frequency data (UHFD), that is, the data made available by the main exchanges around the world, which document the trading activity at the finest detail. Tick–by–tick data reflect the quotations available and the trades perfected during the day; several thousands observations are quickly accumulated in the specialized databases. The time elapsing between records are random and, as a consequence, the data are irregularly spaced. While not the first to make use of UHFD, Engle and Russell [1998] showed that detailed data on market activity show some empirical regularities which are at the base of the GARCH–type modelling of conditional variance of financial returns. The stylized facts of persistence in the series which are well–known in absolute returns carry over to other financial variables, such as durations, volume, number of trades and volatility.

Volatility analysis plays a central role in financial econometrics. Traditionally addressed as the conditional variance of asset returns, volatility was measured and modelled within the same class of models within a GARCH framework. The availability of ultra–high frequency data has fostered a stream of literature within which volatility (a parameter of a continuous time process) was considered to be estimated and its modelling to be kept as a separate stage. The literature on realized volatility (cf. for a survey Andersen and Bollerslev [1998], Andersen et al. [2003], and Hansen and Lunde [2006] among others) is nowadays vast and addresses a number of hypotheses on the type of underlying process which generates the asset data. We can therefore measure volatility and derive time series to be analysed and modelled with volatility forecasting in mind. The applications in derivatives pricing, risk management and asset allocation are the main playing

field within which the model performance has a natural metric of evaluation.

In this thesis, we backtrack on several issues, by offering a moment of critically assessing the state of the art. First of all, we offer a contribution on the issue of handling very large compressed data files. This is an overlooked issue as very little detail is given on how the vast quantity of record contained in specialized UHFD databases are processed, cleaned and transformed into manageable time series. We wanted to tackle a challenging task of manipulating zipped monthly files which typically consist of several hundreds megabytes by ticker. This task is successfully accomplished within an open-source software `TAQ_MNGR` (jointly developed with Fabrizio Cipollini and Giampiero M. Gallo). Reading the records, cleaning the data from wrong records, aggregating records with the same time stamp, and possibly performing other operations to create regularly spaced time series. The package is written in C++ with the primary aim to create a well organized database of financial time series ready for subsequent econometric analysis. The routines developed implement the UHFD handling procedures proposed by Brownless and Gallo [2006a] exploiting the functionalities of the library `zlib`, which allows us to directly manipulate compressed files (this is one of the main contributions of our software compared to the existing R-TAQ package).

We offer a contribution in the field of analysing the properties of estimators of integrated volatility in the second essay (jointly written with Cecilia Mancini). The issue is about how to measure the volatility in presence of jumps, that is of abrupt and shortly lived bursts in prices which appear as spikes in returns and typically may have an impact on subsequent volatility. The literature on the subject is extensive and various estimators are proposed (most of which uses the non-parametric approach) which allow for separating the "predictable" (Brownian) component of the volatility from the jump component due to these sudden and large movements in the price of a given asset. In this case, the contribution of our work is twofold: a complete review of the literature on the topic, pointing out the characteristics in each estimator; the second is to evaluate and compare their performance in finite samples, given that their asymptotic properties have already proved theoretically in the original papers, through a massive Monte Carlo. The results point to the Threshold Bipower Variation of Corsi et al. [2010] as one of the best procedure to disentangle the two volatility components, since

it is robust to arbitrary calibration of the method (the choice of the threshold) and has desirable small sample properties.

Finally, we offer a contribution in the field of volatility modelling, in adopting time-varying parameters when predicting the realized volatility. The main originality in this paper is the derivation of a Lagrange Multiplier test as a preliminary procedure that allows to assess whether the use of a time-varying parameters model is necessary or not. The novelty is that we frame the problem in the field of Generalized Autoregressive Score models which encompass some already existing models such as GARCH and the Multiplicative Error Models of Engle [2002] and Engle and Gallo [2006]. The new type Lagrange Multiplier test evaluates the null hypothesis of parameter stability against the alternative of GAS$(p, 0)$ effects. In a Monte Carlo experiment, we compare the performance of our $LM_{GAS(p,0)}$ test with those of two other tests for parameter stability suggested in the literature which were built on different alternative hypotheses. As a final element of originality we illustrate the modelling capability of MEMs and GASs to a series where we disentangle the two components of the Brownian contribution to volatility and the jump component on five tickers of widely traded stocks. The series were derived from the TAQ data using our software; the two series which allow us to derive the jump time series are calculated as detailed in Chapter 2 and the test shows the suitability of MEMs and GASs modelling. Model estimation and one-step ahead predicted values are derived and some model selection criteria adopted to judge upon the contribution of jumps to volatility prediction.

Chapter 1

Handling ultra high-frequency data: procedures and the `TAQ_MNGR` package

1.1 Introduction

The wide availability of high-frequency financial data has been one of the most relevant innovations in the quantitative analysis of financial markets over the last years (for a survey, see Engle and Russell [2006]). The expression "financial high-frequency data" refers to datasets containing detailed reports of all the available financial markets activity information. Engle [2000] emphasizes the relevance of these new sources of information by using the expression "ultra" high-frequency data (UHFD) in order to stress that they include the most detailed information available.

The atomic unit of information which builds up ultra high-frequency data is the "tick". The word tick comes from the practitioners jargon. Broadly speaking, a tick is made up of a time stamp and a set of information referring to some specific aspect of the market activity. The ultra high-frequency databases containing tick-by-tick information are very complex to analyse, in that:

- the number of ticks is usually huge;
- the time interval separating two consecutive ticks is random and thus irreg-

ular;

- the sequence of ticks
  - may contain some *wrong* records,
  - may not be time ordered,
  - may exhibit some anomalous behaviour as a result of particular market conditions (e.g. opening, closing, trading halts, etc);
- a tick may contain additional information which is not of interest;
- the sequence and the structure of the ticks strongly depends on the rules and procedures of the institution which produces and collects the information.

It is therefore important to understand the structure of these data in order to extract efficiently the information of interest, without distorting its content.

There are many financial markets producing tick data. The most relevant markets analysed in an ultra high-frequency perspective are the exchange rate and the equity markets. The exchange rate market is an over-the-counter (OTC) market while the equity market is organized in exchange and OTC markets. The data produced by the exchange rate market and the OTC equity market is usually collected by data providers which electronically disseminate their information, like Reuters or Bloomberg, while the information produced by the exchanges is collected by the exchanges themselves. A very important difference between these two marketplaces is that exchanges are usually regulated by special government laws. This implies that they are usually much more controlled and the information collected is much more extensive than the one collected by the OTC market.

As observed by Bollerslev [2001] in a survey on the state of the art in this area, the analysis of these data is not particularly simple, also because regulatory changes and technological advances make this field a rapidly changing one (cf., for example, the minimum price change set at USD 0.01 cent in January 2001, the automated way of updating quotes introduced in May 2003, and so on). The latter aspect is particularly delicate since some stylized facts established for some sample period may not be valid for others.

The literature is not always clear about how the time series of interest are constructed from the raw data, nor whether specific choices of preliminary data handling may have implications on statistical analysis. Notable exceptions are Bauwens and Giot [2001] who describe the trading mechanisms at work at the NYSE and provide some data handling suggestions related to the structure of the TAQ database and Vergote [2005] who is concerned by the widespread use of the 5-second rule suggested by Lee and Ready [1991] to synchronize trades and quotes without checking its robustness for the period and the asset at hand.

In this work we focus on ultra high-frequency data in the Trades and Quotes (TAQ) database available from the New York Stock Exchange (NYSE). The main issues related to managing these types of dataset are presented in this chapter and the handling procedures developed by Brownless and Gallo [2006a] are reported in details. Furthermore, the new `TAQ_MNGR` package is presented as a useful and powerful software tool to deal with UHFD.

1.2 Exchange market rules and procedures

This section briefly reviews the NYSE rules and procedures and describes TAQ, its UHF database. Referring to Brownless and Gallo [2006b], we highlight the main issues arising from a practical perspective. The interested reader can find further details in classic texts such as Hasbrouck [1992], while updated information in technical reports produced by the exchanges themselves cannot be eschewed prior to any analysis on real data. Bauwens and Giot [2001] also provide an accurate description of the rules and procedures of the NYSE as well as of other exchanges.

1.2.1 How does an exchange market work?

Trading securities on the various stock exchanges follows very complex rules and procedures which are subject to adjustments in order to adapt to technological advances and evolving regulatory needs. UHFD contain information regarding all market activity: as a result, it is not easy to synthesize the institutional features which may have relevant consequences on UHFD collection and analysis.

Agents interact with the market of a given asset through a sell or buy transac-

tion proposal which is called *order*. It is not possible to submit orders continuously on the exchange but only in a specific period of time devoted to transactions, the trading day. Orders can be classified into two broad categories: market orders and limit orders. A *market order* represent a buy or sell order of a certain number of assets of a stock at the current standing (bid or ask) price. The relevant feature of these orders is that there is certainty about transaction but uncertainty as to the actual price of the transaction. On the other hand, a *limit order* specifies a limit buy or sell price of a certain number of assets of a stock at which the transaction has to be executed. A buy limit order specifies the maximum price at which a trader is willing to buy while a sell limit order specifies the minimum price at which a trader is willing to sell. The important feature of limit orders is the uncertainty regarding the execution of the order (whether it will be executed and when) but the certainty as to the fact that, if the order were executed, it would be executed at the requested price or better.

The set of all buy and sell limit orders of a given stock forms the so called *book*, which lists the orders on the basis of their price and time of submission. The book provides a very detailed picture of the market for the asset, in fact market participants usually access information regarding the portion of the book which is "near-the-market", i.e. the list of sell and buy limit orders near the current market price, and hence more likely to be executed soon.

The current best sell and buy limit orders form the current *bid* and *ask*, that is the *quote*. The quote provides the upper and lower bound within which transactions will occur.

The market orders reaching the exchange will be executed with the current quote generating a *trade*. The priority rules used to match orders on an exchange can be very complex and specific. Special mention should be made of large orders (block transactions), which may generate many separate transactions. Generally speaking, the exchange regulations try to minimize the number of such transactions, while maintaining a time priority principle.

1.2.2 Trading on the NYSE

The NYSE is a hybrid market, in that it is both an *agency* and an *auction* market. It is an agency market since the buy and sell orders are executed through an agent, the market maker, who is called the *specialist* at the NYSE. The NYSE is also an auction market, since on the exchanges *floor* the *brokers* participate actively in the negotiation and thus contribute to the determination of the transaction price. Also note that despite the global trend towards computer automated trading systems, the "human" element has a very crucial role in the trading mechanisms, and this fact as a deep impact on the data. Rules and procedures of the NYSE are described in detail in Hasbrouck [1992], Hasbrouck et al. [1993], Madhavan and Sofianos [1998], O' Hara [1997].

Trading takes place on the floor Monday to Friday, from 9:30AM to 4:00PM. The trading floor of the NYSE is composed of a series of contiguous rooms, where the trading posts are located. All the activity linked to a given stock is done in proximity of the panels located on the trading posts. At each panel the specialist of the assigned stock and the brokers interested in trading form the so called trading crowd. The trading crowd function is to determine the transaction price of a stock through the negotiation. As a market maker, the specialist is obliged to ensure that the market is liquid enough and stable, that is s/he has to make the market at her/his own risk with the aim of making it as easy as possible for the brokers to execute their transactions. During the trading day, the specialist remains close to the trading post and the panel relative to the stock s/he is assigned to together with his clerks. The brokers, on the other hand, can trade any stock on the floor, thus they move across the various trading posts, even if they specialize just on a few stocks. Brokers can participate in the trading either in an active mode, interacting within the trading crowd, or in a passive mode, leaving her/his orders to the specialist.

The order handling mechanisms of the exchange floor are constructed around the figure of the specialist. Orders reach the specialist through either a broker or through the NYSE network. The market orders that reach the specialist wait until they are executed by the specialist. The specialist can execute them with another order, on her/his inventory, against a broker in the trading crowd or in an

another linked exchange. All orders are always executed within the current bid and ask quotations, that is to say the highest buying price and lowest selling price set by the specialist. It is important to stress that the current highest buy limit order and the lowest sell limit order in the book do not automatically become the current bid and ask. An order becomes the current quote when the specialist communicates this to the trading crowd (there are cases when the specialist is forced to do so). It is also important to emphasize that transactions are almost never automatically executed. Thus in practice, all relevant transactions are executed under the specialist supervision.

Special regulations by the NYSE apply to important events in the trading activity, namely market opening, closing and the trading interruptions. At the opening of the trading day, the specialist is obliged to present a quotation that is as close as possible to the closing price of the previous day. The specialist is helped in this operation by an automated program developed by the NYSE IT infrastructure which matches the outstanding sell and buy orders and presents the balance to the specialist. On this basis s/he will then decide how many buy or sell orders will be presented.

The closing price is set on the basis of the balance of *market-on-close* (MOC) orders, which are buy or sell orders that are executed entirely at the closing price. In case there is a non null balance of the MOC orders, the difference is executed against the current bid or ask at the closing, determining the closing price. The other orders will be executed at that price. If the size of the buy and sell MOC order is equal, the closing price is the price of the last transaction.

Under some particular circumstances, the specialist can declare a delay (opening delay) of the opening of the transactions for some stocks, or a temporary interruption (trading halt) of the transactions. The possible causes for these types of delays or halts are news pending, news dissemination, or big order balances needing to be executed. After the market breaks of October 1987 and 1989, the NYSE also introduced some market-wide "circuit-breakers" the function of which is to slow down or even stop transactions in high volatility phases. Finally the Board of Directors of the NYSE can declare in particular conditions other types of interruption of the normal trading day (snow storms, network problems, commemorations and so on).

Even if a large part of the procedures of the NYSE are automated, its dynamics are clearly strongly characterized by the interaction on the floor between the specialists and the brokers. Although almost all transactions which are executed on the NYSE come from the exchange network system, floor brokers execute a very large share of the total volume of transactions executed. There are some interesting empirical studies which have tried to further characterize the quality of brokers trading. Some estimates made by Sofianos and Werner [2000] found that the orders of the broker are on average 5 times bigger than the orders reaching the floor through the network, but, of course, technological advances and the widespread use of automated orders may have reduced the gap in recent times. Brokers will then divide an order into smaller trades, which will be executed in a time range which on average ranges from 11 to 29 minutes. This brings to mind the fact that the *on floor* information set of brokers contains lots of important information which is not available *off floor*.

Thus, these examples seem to suggest that the higher the frequency the higher attention should be devoted to the analysis of the data.

1.3 The TAQ database

The categories of data collected within the TAQ are *trades* and *quotes*. The NYSE is probably the first exchange which has been distributing its ultra high-frequency data sets since the early 1990s. In 1993 the trades, orders and quotes (TORQ) database was released (Hasbrouck [1992]) which contained a 3 month sample of data. Since 1993, the NYSE has started marketing the trades and quotes (TAQ) database. It has undergone some minor modifications through the years leading to 3 different TAQ versions (1, 2 and 3). Since 2002, order book data has also been separately available for research purposes, but will not be discussed here. Although through time many improvements have been added to the quality of the data, the TAQ data are raw, in that the NYSE does not guarantee the degree of accuracy of the data, so that further manipulations are needed for using them in research.

Quote data contain information regarding the best trading conditions available on the exchange. Table 1.1 displays a few sample records from the quote database

SYMBOL	DATE	TIME	BID	OFR	BIDSIZ	OFRSIZ	MODE	EX
ANF	20120702	9:35:00	34.43	34.63	1	4	12	B
ANF	20120702	9:35:00	34.43	34.54	1	3	12	Y
ANF	20120702	9:35:00	34.44	34.48	4	3	12	P
ANF	20120702	9:35:00	34.33	34.60	4	11	12	X
ANF	20120702	9:35:00	34.44	34.48	3	1	12	N
ANF	20120702	9:35:01	34.43	34.61	1	1	12	B
ANF	20120702	9:35:01	34.44	34.48	5	1	12	N

Table 1.1: TAQ quote records. SYMBOL is the stock symbol; DATE is the quote date in format *yyyymmdd*; TIME is the time at which the quote entered the Consolidated Quote System (CQS); BID is the bid price; OFR is the offer price; BIDSIZ is the bid size in number of round lots (100 share units); OFRSIZ is the offer size in number of round lots (100 share units); MODE is the quote condition; EX is the exchange on which the quote occurred. ANF is the ticker for the stock of Abercrombie & Fitch (clothing company).

SYMBOL	DATE	TIME	PRICE	G127	CORR	COND	EX	SIZE
ANF	20120702	9:35:00	34.4400	0	0	@	N	100
ANF	20120702	9:35:00	34.4500	0	0	F	N	100
ANF	20120702	9:35:00	34.4600	0	0	F	N	200
ANF	20120702	9:35:00	34.4800	0	0	@	D	500
ANF	20120702	9:35:03	34.4700	0	0	F	Z	200
ANF	20120702	9:35:05	34.4800	0	0	@	D	200
ANF	20120702	9:35:06	34.4800	0	0	@	N	100

Table 1.2: TAQ trade records. SYMBOL is the stock symbol; DATE is the trade date in format *yyyymmdd*; TIME is the time at which the trade entered the Consolidated Trade System (CTS); PRICE is the trade price per share; G127 is a field indicating simultaneously: a G trade (a sell or buy transaction made by a NYSE member on his own behalf); a rule 127 transaction, i.e. a transaction executed as a block position; a stopped stock; CORR is the Correction Indicator (see text); EX is the exchange on which the trade occurred; SIZE is the number of shares traded. ANF is the ticker for the stock of Abercrombie & Fitch (clothing company)

with an explanations of the various fields. The quote table fields unfortunately do not include any information on the quality of the reported data. However, the MODE field (quote condition) contains many useful information which can be used to reconstruct accurately the trading day events and some specific market conditions. Some values of this field indicate various types of trading halts that can occur during the trading day. Furthermore, the field also contains values indicating the opening and closing quotes.

Trade data contain information regarding the orders which have been executed on the exchange. Table 1.2 displays few sample records from the trade database. Some fields of the database containing information on the quality of the recorded ticks, allowing for the removal of wrong or inaccurate ticks: e.g. the CORR field (correction indicator) signals whether a tick is correct or not, and the "Z" and "G" value of COND field (sale conditions) indicate a trade reported at a later time.

1.4 Ultra high-frequency data handling

Considering the nature of UHFD, Brownless and Gallo [2006a] develop a simple and fast method for identifying single records as outliers and detail the procedure to translate the clean tick data into time series of interest for subsequent analysis. The proposed procedure involves two main steps:

1. *Data cleaning*, i.e. detecting and removing wrong observations from the raw UHFD;

2. *Data management*, i.e. constructing the time series of interest for the objectives of the analysis.

1.4.1 Data cleaning

To obtain a clean sample, the records which are not of interest have to be identified and discarded using available information. Falkenberry [2002] reports that errors are present both in fully automated and partly automated trading systems, such as the NYSE, and attributes the main determinant of errors to trading intensity.

The higher the velocity in trading, the higher the probability of errors in reporting trading information.

For the quote data, all quotes with a primary NYSE listing that did not occur on the NYSE (EX field different from N) should be eliminated. From a graphical survey carried out in Brownless and Gallo [2006a] we can notice that non-NYSE quotes have a very large spread and there are often extremely large quotes or suspicious zeros. In the paper they state that, for NYSE-listed stocks this pattern is probably due to the fact the NYSE is recognized as the leader market and thus other exchanges do not post competing quotes. Non-NYSE spreads are hence bound to be much larger than the NYSE ones. It is important to note that while discarding incorrect and delayed trade records implies removing a usually very small fractions of observations, removing non-NYSE quote records can have a dramatic impact on the reduction of the sample size. As some authors suggest (e.g. Vergote [2005] and Bohemer et al. [2007]) we also have to remove from quote data those quotations which were generated by non normal market activity, as indicated by the MODE field values 1, 2, 3, 6 or 18.

For trade data, all transactions that are not correct (CORR field different from 0) and delayed (COND field equal to Z) should be eliminated from the sample. Contrary to quote data, we prefer not to discard transaction prices that did not occur on the NYSE, though in some cases this is not advisable (see Dufour and Engle [2000]).

After this TAQ based filtration of the data of interest, the remaining tick-by-tick price series still show observations that are not consistent with the market activity.

Let $\{p\}_{i=1}^{N}$ denote an ordered tick-by-tick price series. The proposed procedure to remove outliers is

$$|p_i - \bar{p}_i(k)| < 3s_i(k) + \gamma = \begin{cases} \textit{true} & \text{observation } i \text{ is kept,} \\ \textit{false} & \text{observation } i \text{ is removed,} \end{cases} \tag{1.1}$$

where $\bar{p}_i(k)$ and $s_i(k)$ denote respectively the δ-trimmed sample mean and sample standard deviation of a neighbourhood of k observations around i and γ is a granularity parameter. The neighborhood of observations is always chosen so

that a given observation is compared with observations belonging to the same trading day. That is, the neighbourhood of the first observation of day are the first k ticks of the day, the neighbourhood of the last observation of the day are the last k ticks of the day, the neighbourhood of a generic transaction in the middle of the day is made by approximately the first preceding $k/2$ ticks and the following $k/2$ ones, and so on. The idea behind the algorithm is to assess the validity of an observation on the basis of its relative distance from a neighbourhood of the closest valid observations. The role of the γ parameter is to avoid zero variances produced by sequences of k equal prices. The percentage of trimming δ should be chosen on the basis of the frequency of outliers, the higher the frequency, the higher the δ. The parameter k should be chosen on the basis of the level of trading intensity. If the trading is not very active k should be "reasonably small", so that the window of observations does not contain too distant prices (the contrary is true if the trading is very active). The choice of γ should be a multiple of the minimum price variation allowed for the specific stock. The procedure is inevitably heuristic but it has the virtue of simplicity and effectiveness.

1.4.2 Data management

Simultaneous observations: In general, there are several transactions reported at the same time which were executed at different price levels. Simultaneous prices at different levels are also present in quote data. There are different explanation for this phenomenon. First of all, note that the trading of NYSE securities can also be performed on other exchanges, and thus simultaneous trades at different prices are normal. Also, a market order execution on one exchange may produce more than one transaction report in some cases. Finally, even non simultaneous trades/quotes could be all reported as simultaneous due to trade reporting approximations.

As ultra high-frequency models for the modelling of tick-by-tick data usually require one observation per time stamp, some form of aggregation has to be performed. Taking the median price could be a reasonable solution given the discrete nature of the tick-by-tick data. In case further aggregations at lower frequencies

will be performed, the method of aggregation choice becomes progressively less relevant as the difference between prices will be negligible, and simpler methods such as the last or first price of the sequence should not cause any problems. For tick-by-tick volumes or transaction counts, the natural way to aggregate observations is to substitute the simultaneous observations with the sum of the simultaneous volumes and the number of simultaneous transactions.

Irregularly spaced data: the most striking feature is that the time series are irregular, with a random time separating two subsequent observations. To turn it into a time series with discrete, equally spaced time intervals, let us consider an irregular time series $\{(t_i, y_i)\}_{i=1}^{N}$, where t_i and y_i indicate, respectively, the time and value of the ith observation, and let $\{(t_j^*, y_j^*)\}_{j=1}^{N^*}$ be the lower frequency time series that we intend to construct using an appropriate aggregation function such as

$$y_j^* = f(\{(t_i, y_i) | t_i \in (t_{j-1}^*, t_j^*]\}). \tag{1.2}$$

Some simple but useful methods which are coherent with this scheme are:

- **First:** $y_j^* = y_f$ where $t_f = \min\{t_i | t_i \in (t_{j-1}^*, t_j^*]\}$;
- **Minimum:** $y_j^* = \min\{y_i | t_i \in (t_{j-1}^*, t_j^*]\}$;
- **Maximum:** $y_j^* = \max\{y_i | t_i \in (t_{j-1}^*, t_j^*]\}$;
- **Last:** $y_j^* = y_l$ where $t_l = \max\{t_i | t_i \in (t_{j-1}^*, t_j^*]\}$;
- **Sum:** $y_j^* = \sum_{t_i \in (t_{j-1}^*, t_j^*]} y_i$;
- **Count:** $y_j^* = \#\{(y_i, t_i) | t_i \in (t_{j-1}^*, t_j^*]\}$.

In the first four methods if the set $\{t_i | t_i \in (t_{j-1}^*, t_j^*]\}$ is empty the jth observation will be considered missing. The "First", "Minimum", "Maximum" and "Last" methods can be useful for the treatment of price series (e.g. the "Maximum" and the "Minimum" are the base for the realized range; "Last" can be used for the computation of realized variance, and so on). The "Sum" method is appropriate for aggregating volumes and "Count" can be used to obtain the number of trades and/or quotes in a given interval.

As far as the construction of regular price series is concerned, Dacorogna et al. [2001] proposed some methods which are based on the interpolation at t_j^* of the previous and the next observation in series:

- **Previous point interpolation**: $y_j^* = y_p$ where $t_p = \max\{t_i | t_i < t_j^*\}$;
- **Next point interpolation**: $y_j^* = y_n$ where $t_n = \min\{t_i | t_i > t_j^*\}$;
- **Linear point interpolation**: $y_j^* = \left(1 - \frac{t_j^* - t_p}{t_n - t_p}\right) y_p + \frac{t_j^* - t_p}{t_n - t_p} y_n$.

The problem in using these methods, however, is that they might employ information not available at t_j^*. For liquid stocks, the choice of the interpolation schemes does not seem to be particularly relevant as the neighbourhood of t_j^* will be very dense of observations, and the different interpolation schemes will deliver approximately the same results of the "Last" method. On the other hand, results may be unsatisfactory for infrequently traded stocks since at certain frequencies the interval $(t_{j-1}^*, t_j^*]$ will not contain any observation and the interpolation may refer to prices recorded at some remote time. In these cases we think it more appropriate to treat the observation as missing, in order to avoid long sequences of zero or identical returns.

Opening and closing: It would be natural to expect the first (respectively, last) trade/quote of the day to be the recorded opening (respectively, closing) trade/quote. The official NYSE trading day begins at 9:30 AM and finishes at 4:00 PM, but in practice trading will start some time after the official opening and will go some time beyond the closing, which implies that de facto the actual times of the opening and closing are indeed random. In addition to this, it is common to find reports of trades and quotes in the data which clearly do not belong to the NYSE trading day, that is, transactions before 9:30 AM and after 4:00 PM. These trade and quotes records may either come from other exchanges or from the off-hours (crossing) sessions of the NYSE and are therefore discarded. Lastly, to complicate matters, it is not uncommon that the actual trading of a stock will begin later than the opening time because of opening delays; also on some special days (like, for example, those preceding a holiday) the stock exchange may close at an earlier time.

It is thus not always possible to exactly identify the opening and the closing data. The MODE field in the quote data (see Table 1.1) and the G127 and the COND fields of the trade data (see Table 1.2) contain some flags that can be used to identify the exact opening/closing trades and quotes, but, unfortunately, this piece of information is not always accurately reported.

In order to adequately capture the closing price of the day, we adopt the convention that the trading day hours span between 9:30 AM and 4:05 PM, which ensures (to a large degree) that closing prices possibly recorded with a delay are accounted for. When using fixed-time intervals such as when building 10 min returns series, the last interval will span a nominally longer period (in the example, 15 min between 3:50 and 4:05 PM). This will give the return computed as the log-difference between the closing price and the price recorded at 3:50 PM as the last observation of the day.

1.5 The `TAQ_MNGR` software

As highlighted in this chapter, dealing with the preliminary manipulation of UHFD is not such an easy task. The `TAQ_MNGR` software purpose is to help users in the construction of their own database of ready-to-use financial time series, implementing the data cleaning and managing procedures developed by Brownless and Gallo [2006a] (section 1.4).

The package is built and tested on trades and quotes files (see tables 1.2 and 1.1 respectively) provided by the Wharton Research Data Service (WRDS). For a matter of efficiency and portability, it is entirely written in `C++` and integrated in R using the `Rcpp` package and in MATLAB via `mex` interface. See Table 1.3 for a review of the available versions.

In a perspective of avoiding system overload, the WRDS web policy provides for some restrictions in terms of the largest time range and the maximum number of stock symbols that can be processed at each data request (no more than one query can be run simultaneously). Bearing these restrictions in mind, we set some simple requirements each file has to satisfy in order to be properly elaborated by the software:

	Stand-alone	R	MATLAB
Linux *(tested on Ubuntu 12.04)*	x	x	x
Windows *(tested on Windows 7)*	x	x	

Table 1.3: `TAQ_MNGR` available versions.

- the data contained in each file must belong to the same month (set the proper date range at `Step 1` of WRDS query submitting procedure – figure 1.1);
- if multiple files for the same month are saved in the same directory, each of them must contain data for different stocks (insert the proper symbols for each data request at `Step 3` – figure 1.1);
- all fields have to be included in the file (push the *"Check All"* button at `Step 4` – figure 1.1);
- select the *fixed-width text* and `G zip` as output format and compression type respectively at `Step 5` (figure 1.1).

In order to satisfy these requirements, some care must be taken in the data downloading stage. No further requirements are imposed and each file can contain (potentially) as many ticker symbols as desired since the software itself will deal with the proper splitting and storing of the downloaded files during the cleaning stage (we refer to this as the cleaning/splitting stage).

Once a data request is processed by the WRDS system, the output file is named with an alphanumeric sequence which denotes the query identification. Figure 1.2 shows how data appear after the downloading stage: each file contains one month of either trades or quotes of five among the most traded stocks on the NYSE (see table 1.4 for details about the selected ticker symbols). We suggest to use a unique folder to store the raw data files: the software itself, as we will see in more detail below, takes care of keeping track of the previously cleaned files for each prespecified output directory.

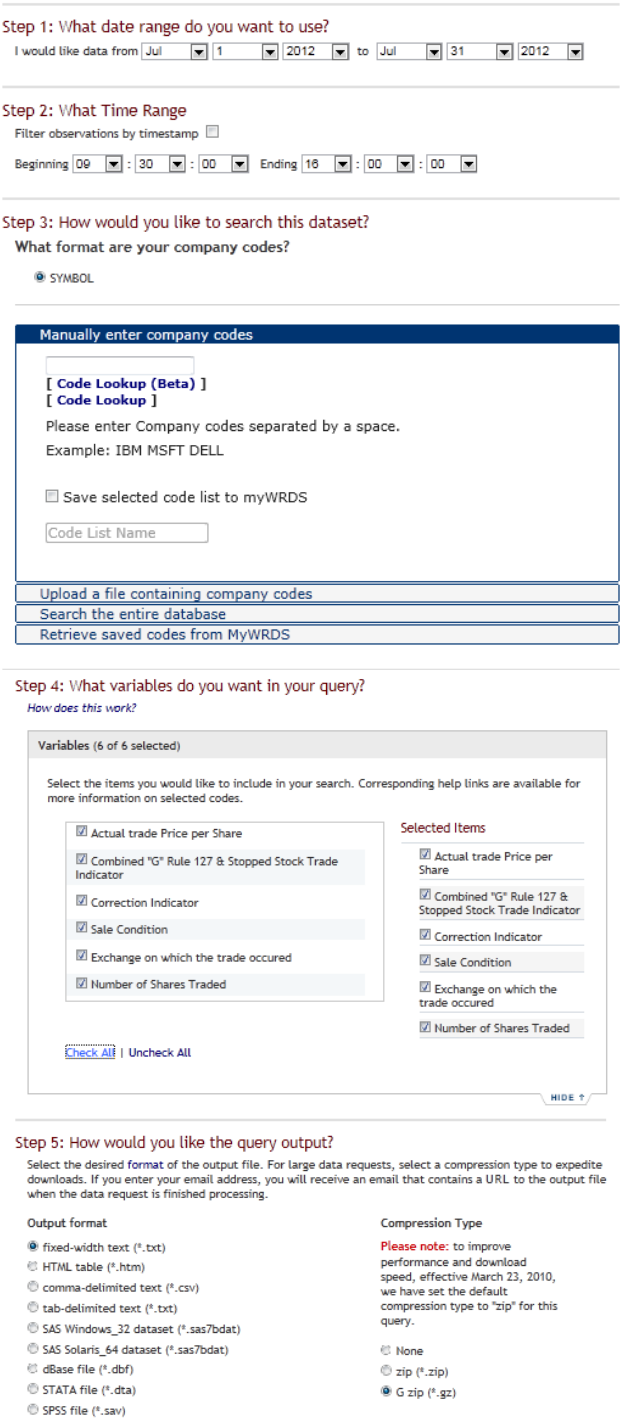

Figure 1.1: Screenshot of the WRDS query submitting webpage.

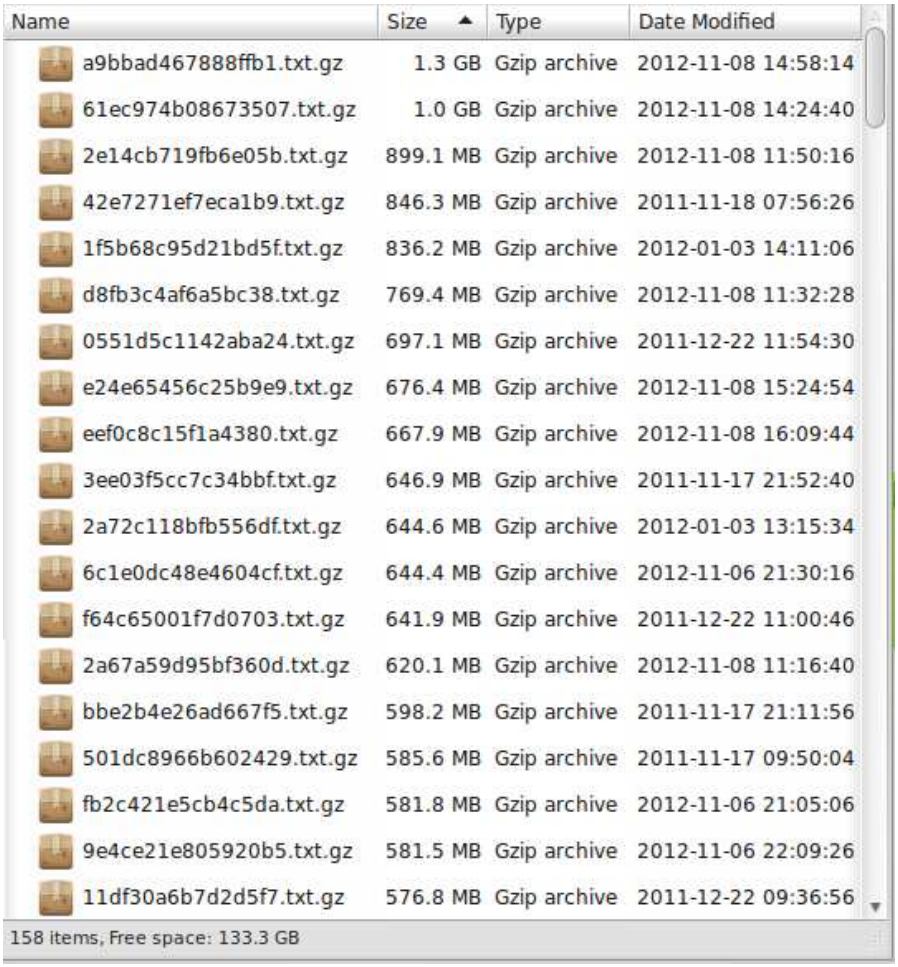

Figure 1.2: Screenshot of the directory containing the downloaded compressed raw data files.

SYMBOL	COMPANY	SECTOR	INDUSTRY
ANF	Abercrombie & Fitch Co.	Services	Apparel Stores
BAC	Bank of America Corporation	Financial	Money Center Banks
C	Citigroup, Inc.	Financial	Money Center Banks
F	Ford Motor Co.	Consumer Goods	Auto Manufacturers
GE	General Electric Company	Industrial Goods	Diversified Machinery

Table 1.4: Details about the stocks included in each downloaded TAQ file.

Observing figure 1.2, we may realize the order of magnitude (hundreds of Megabytes) of the size of each raw file in `G zip` format. This suggest the implementation of algorithms for directly processing compressed data. The proper input/output functionalities are provided by the `Gzstream` library, basically a C++ open-source wrapper for the `zlib` C-library developed by *Deepak Bandyopadhyay* and *Lutz Kettner* at the *Computational Geometry Group* at *UNC Chapel Hill*. The `Gzstream` source code is compiled in the package itself, while a pre-compiled dynamic version of the `zlib` is furnished for both Linux and Windows (most Linux distributions already include it by default). The use of these powerful libraries let the `TAQ_MNGR` package to deal efficiently with files in `G zip` format without manually decompressing them with the usual applications (a rather difficult task for "average" PC, considering the huge amount of bytes to decompress).

1.6 `TAQ_MNGR` functions in detail

`ControlFiles(dirInput, dirOutput)`

Arguments	*Description*
`dirInput`	(*character array*) name of the directory containing raw data files.
`dirOutput`	(*character array*) name of the directory where to save the file `list.txt`.

R command example: `TAQ.ControlFiles("home/user/Documents.../TAQdata", "home/user/Documents")`

This function fulfills the need of knowing the "time" content (in terms of year and month) of each raw data file, since, after the downloading stage, no hint is contained in the file name itself. In practice, this command allows to quickly check (instead of directly running the cleaning/splitting routine) if a file containing the data for a certain year/month is stored in the directory `dirInput`. The considered time range for the checking procedure spans from the most remote to the

latest date (year/month) detected among the files saved in the folder `dirInput`. The results are reported in the textual spreadsheet named `list.txt` (see figure 1.3) and stored in the directory `dirOutput`.

DATE	QUOTE	TRADE
2006 JAN	1d664cd6a1c76ac0.txt.gz	3313b65e153597fd.txt.gz
2006 FEB	f57ac35d5a7a7543.txt.gz	d6ed9b2fe9a118e9.txt.gz
2006 MAR	7fb95e61ad1299e6.txt.gz	MISSING FILE
2006 APR	d767518026f9c328.txt.gz	2f80e09ff50336b7.txt.gz
2006 MAY	MISSING FILE	f23827c0e6cd7819.txt.gz

Figure 1.3: Example of the file `list.txt`. For each considered year/month, if the file containing data for the specified year/month is detected in the `dirInput` folder, the file name is reported, otherwise the tag `MISSING FILE` appears.

`CleanTickByTick(dirInput, dirOutput, window, deltaTrim, granularity, useCleaned)`

Arguments	*Description*
`dirInput`	(*character array*) name of the directory containing raw data files.
`dirOutput`	(*character array*) name of the chosen root directory to store processed data (see figure 1.4. If it doesn't exist, it is created).
`window`	(*integer value greater than* 3) number of consecutive observations in the *evaluating window* (the k parameter of equation 1.1).
`deltaTrim`	(*real value within* [0,1)) trimming proportion (the δ parameter of the price trimmed mean $\bar{p}_i(k)$ in equation 1.1).
`granularity`	(*real value greater than* 0) positive parameter that avoids zero variances in the cleaning procedure (the γ parameter of equation 1.1).

`useCleaned` (*boolean value*) give instructions about the preliminary use of the list containing the previously cleaned files (see the function description for details).

R command example: `TAQ.CleanTickByTick("home/user... .../Documents/TAQdata","home/user/Documents/CleanedTAQData", 80, 0.10, 0.04, TRUE)`

The function implements the cleaning procedures described in section 1.4.1. This routine also keeps track of the raw files previously cleaned and stored in the given `dirOutput` directory, using a hidden textual document (stored in the directory itself) updated each time a new raw file is cleaned. The files that still have to be processed in that specific `dirOutput` folder are identified by comparing the list of raw files included in the `dirInput` directory with the list of previously cleaned files reported in the hidden file. If the variable `useCleaned` is set to `FALSE` or this function is run for the first time for that `dirOutput` folder, the preparatory checking stage is not carried out, all files in the `dirInput` directory are processed, and the hidden file (if available) is cancelled and replaced by a new one.

After this preliminary check, the cleaning/splitting stage starts. For each file to be processed:

- depending on the file type (whether trade or quote is automatically detected by the software), the proper cleaning procedure is applied;
- for each included ticker symbol, one compressed file of cleaned data is created (within the cleaned file, the field `TIME` is expressed as cumulative number of seconds since 00:00 AM, rather than with the original `hh:mm:ss` format);
- the cleaned files are stored in `dirOutput` according to the directory tree in figure 1.4 (if the folder doesn't exist, it is automatically created). Figure 1.5 shows an example of how the directories are organized after the cleaning stage.

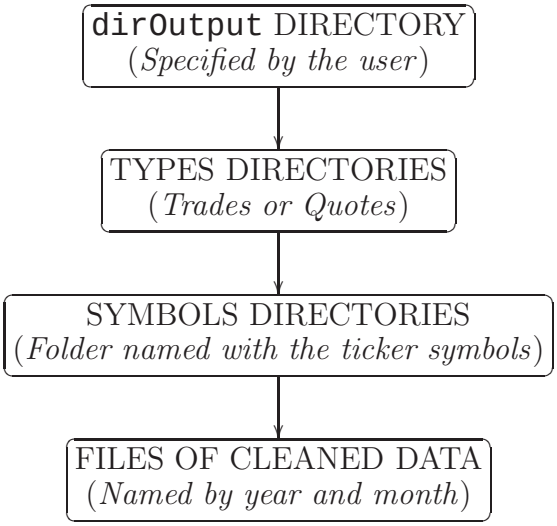

Figure 1.4: Structure of the `dirOutput` directory (where the cleaned data files are stored).

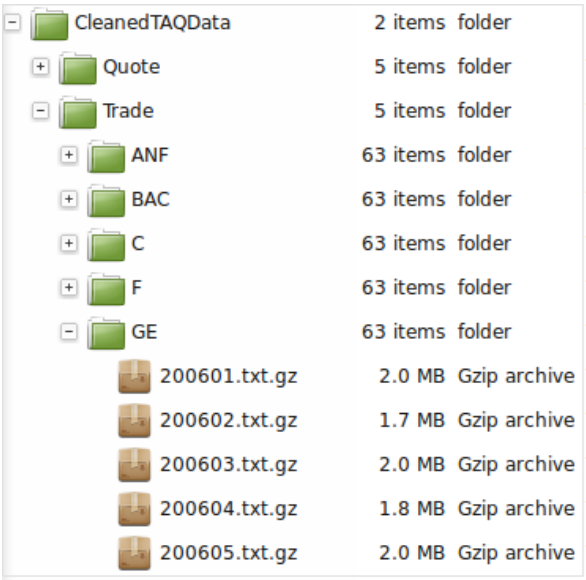

Figure 1.5: Actual example of directory tree where the cleaned data files are stored.

`Aggregate(dirInput, symbol, startDate, endDate, bin, missContinue)`

Arguments	*Description*
`dirInput`	(*character array*) name of the directory containing cleaned data files.
`symbol`	(*character array*) name of the stock symbol within `dirInput` to be aggregated.
`startDate`	(*integer value greater than* 0) lower bound (year and month) of the date range of the data intended to be aggregated (format `yyyymm`).
`endDate`	(*integer value greater than* 0) upper bound (year and month) of the date range of the data intended to be aggregated (format `yyyymm`).
`bin`	(*integer value within* [1,23700]) the aggregation frequency in seconds (for instance 23700 is the daily frequency).
`missContinue`	(*boolean value*) if `TRUE`, the aggregation function is run even if missing files (relative to cleaned data for the chosen symbol and within the specified date range) are detected. If `FALSE`, no aggregation is carried out if missing cleaned files are detected.

R command example: `TAQ.Aggregate("home/user/Documents...../CleanedTAQData","BAC",200601, 201207, 300, TRUE)`

This function aggregates previously cleaned data according to the procedures described in section 1.4.2. In the `dirInput` folder, containing the clean data (which must be the same as one of the previously specified `dirOutput DIRECTORY` – see figure 1.4), the aggregated data files are stored following the directory tree as in figure 1.6. Figure 1.7 shows an example of how the directory structure appears after the `CleanTickByTick` and the `Aggregate` functions are run. For the chosen symbol, a file of aggregated data is stored in `G zip` format per each month within the date range specified by the user (in order to create a file of aggregated data for a given month and a specified stock symbol, both the cor-

responding trades and quotes files have to be previously cleaned). Each of them contains ready-to-use equally spaced time series at the given bin frequency. In detail the file reports: the *first*, the *minimum*, the *maximum*, the *last* of *trade prices*, *mid-quotes*, and *bid-ask spreads*, the *number of trades*, *trade volume*, *bid size*, and *ask size* relative to each bin interval in the trading day (see figure 1.8 for a portion of a file including data aggregated at `bin` = 300 seconds).

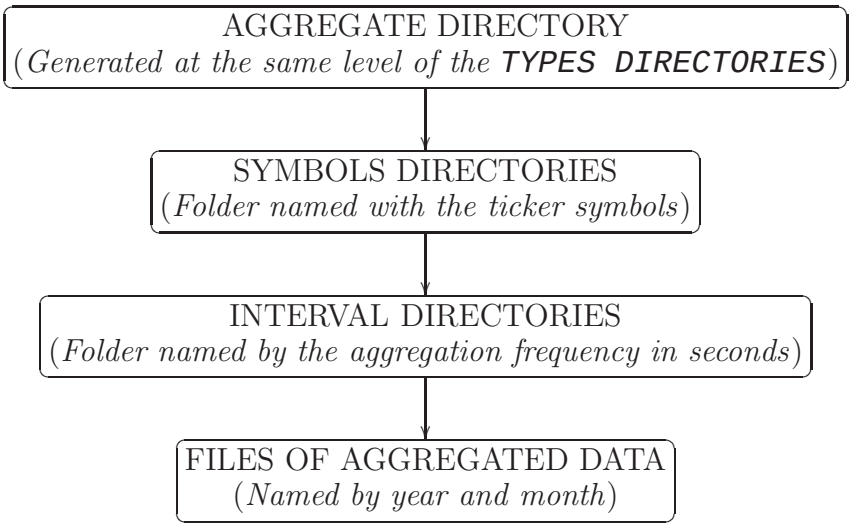

Figure 1.6: Directory structure where the aggregated data files are stored.

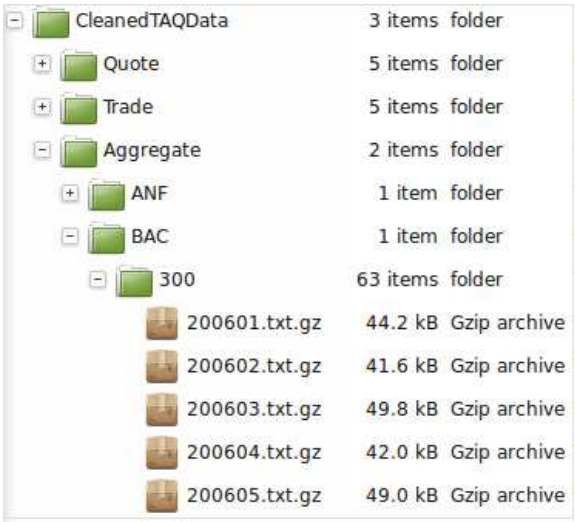

Figure 1.7: Actual example of directory tree where the aggregated data files are stored.

DATE	TIME	FIRST	MIN	MAX	LAST	SIZE	#TRADES	FSTMID	MINMID	MAXMID	LSTMID	FSTSPD	MINSPD	MAXSPD	LSTSPD	BIDSIZE	OFRSIZE
20060103	34200	46.15	46.15	46.4	46.4	7200	16	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0	0
20060103	34500	46.92	46.3	46.99	46.31	569300	352	46.33	46.305	46.375	46.315	0.06	0.01	0.09	0.01	6614	4254
20060103	34800	46.3101	46.22	46.32	46.24	227100	163	46.315	46.225	46.32	46.245	0.01	0.01	0.04	0.01	9329	9290
20060103	35100	46.24	46.23	46.31	46.29	132300	153	46.245	46.225	46.3	46.295	0.01	0.01	0.06	0.03	5086	3879
20060103	35400	46.29	46.25	46.32	46.29	185300	189	46.295	46.255	46.315	46.295	0.03	0.01	0.05	0.01	5652	9430
20060103	35700	46.29	46.25	46.38	46.25	184900	150	46.29	46.255	46.365	46.255	0.02	0.01	0.09	0.01	4319	15308
20060103	36000	46.25	46.22	46.29	46.28	148800	160	46.255	46.225	46.285	46.285	0.01	0.01	0.07	0.01	19112	11060
20060103	36300	46.29	46.26	46.36	46.27	151000	206	46.285	46.265	46.345	46.275	0.01	0.01	0.07	0.01	6422	19001
20060103	36600	46.28	46.22	46.29	46.27	164900	118	46.275	46.225	46.285	46.255	0.01	0.01	0.08	0.03	13410	4044
20060103	36900	46.27	46.2501	46.32	46.29	158700	162	46.255	46.255	46.32	46.295	0.03	0.01	0.04	0.01	16062	9626
20060103	37200	46.29	46.24	46.4	46.4	174400	176	46.285	46.255	46.425	46.425	0.01	0.01	0.07	0.07	4960	6124
20060103	37500	46.4	46.35	46.45	46.35	191900	223	46.425	46.345	46.45	46.345	0.01	0.01	0.07	0.01	3177	16027
20060103	37800	46.39	46.27	46.39	46.28	164700	123	46.345	46.27	46.35	46.275	0.01	0.01	0.04	0.01	2180	12734
20060103	38100	46.28	46.27	46.31	46.3015	191500	175	46.275	46.27	46.305	46.305	0.01	0.01	0.03	0.01	13008	15124
20060103	38400	46.3	46.2	46.31	46.26	185500	175	46.305	46.26	46.305	46.265	0.01	0.01	0.03	0.01	19952	15645
20060103	38700	46.282	46.26	46.31	46.3	156700	172	46.265	46.265	46.305	46.305	0.01	0.01	0.02	0.01	13341	13795
20060103	39000	46.31	46.3	46.33	46.31	178200	115	46.305	46.305	46.325	46.305	0.01	0.01	0.02	0.01	13148	10561
20060103	39300	46.3006	46.28	46.31	46.29	147900	124	46.31	46.285	46.31	46.285	0.02	0.01	0.03	0.01	10664	11617
20060103	39600	46.28	46.26	46.3593	46.3	109600	149	46.285	46.255	46.315	46.295	0.01	0.01	0.05	0.01	6251	9782
20060103	39900	46.298	46.22	46.31	46.22	158600	139	46.295	46.225	46.305	46.225	0.01	0.01	0.03	0.01	7629	10677
20060103	40200	46.22	46.18	46.23	46.23	156400	164	46.225	46.18	46.23	46.225	0.01	0.01	0.05	0.01	8872	10432
20060103	40500	46.22	46.15	46.23	46.18	191200	199	46.225	46.155	46.225	46.185	0.01	0.01	0.03	0.01	12341	16655
20060103	40800	46.18	46.17	46.32	46.18	261400	214	46.185	46.175	46.245	46.18	0.01	0.01	0.04	0.02	5880	11933

Figure 1.8: Sample of part of a file containing cleaned data aggregated at 5 minutes.

`Read(dirInput, symbol, import, startDate, endDate, bin, missContinue)` *Available for the R and MATLAB versions to provide an environment-specific data input command.*

Arguments	***Description***
`dirInput`	(*character array*) name of the directory containing cleaned and aggregated data files.
`symbol`	(*character array*) name of the stock symbol of the already aggregated data intended to be imported.
`import`	(*character double array*) vector containing the names of the fields intended to be imported. The names have to be chosen from the list: `"FIRST", "MIN", "MAX", "LAST", "SIZE", "#TRADES", "FIRSTMID", "MINMID", "MAXMID", "LASTMID", "FIRSTSPREAD", "MINSPREAD", "MAXSPREAD", "LASTSPREAD", "BIDSIZE", "OFRSIZE"`.
`startDate`	(*integer value greater than* 0) lower bound (year and month) of the date range of the data intended to be imported (format `yyyymm`).
`endDate`	(*integer value greater than* 0) upper bound (year and month) of the date range of the data intended to be imported (format `yyyymm`).
`bin`	(*integer value within* [1,23700]) aggregation frequency in seconds of the data intended to be imported (e.g. 300=5 minutes).
`missContinue`	(*boolean value*) if `TRUE` the importing function is run even if missing files containing aggregated data for the chosen stock symbol and within the specified date range are detected. If `FALSE` no data importing is carried out if missing aggregated files are detected.

R command example: `TAQ.Read("home/user/Documents/... .../CleanedTAQData", "BAC",c("LAST", "LASTMID", "SIZE"), 201101, 201207, 300, TRUE)`

Exploiting this function, the (cleaned and aggregated) time series is imported either in R or in `MATLAB`. The output is a matrix of data containing the fields: `DATE` (within the interval [`startDate,endDate`]), `TIME`, and the variables intended to be imported for a certain stock symbol and a given `bin`.

1.7 Conclusions

A lot of work still have to be done to improve the software. First of all, we should make the package more flexible by modifying the implemented functions in the way that they can process not only WRDS data, but also the TAQ files provided by other data providers. The second step should be the redefinition of the restrictions imposed on the file to be properly processed (i.e. we want to let the user to download multiple not-overlapping file for the same month or containing data bridging multiple months). We are also thinking to include proper routines for the computation of time series of various volatility measures.

Chapter 2

Measuring volatility in presence of jumps

2.1 Introduction

Conditional distributions of returns are a centerpiece of financial econometrics, with the idea that proper derivatives pricing, risk management and asset allocation decisions rest upon their modelling and predictability. A special role is played by the conditional variance, in view of the stylized facts that the range of variability of returns is time–varying and persistent, thus giving the idea that it can be predicted. The issue of its measurement and modelling was traditionally subsumed within the GARCH framework (see Engle [1982] and Bollerslev [1986]) and stochastic volatility (see Shephard [2005]): volatility was treated as an unobservable variable and second moments of returns were made dependent upon past squared returns with a possible asymmetric reaction to positive or negative signs of returns. While GARCH modelling has been very successful, the precision of its predictions has always been fairly disappointing, given that squared returns are a very noisy measure of volatility. Andersen and Bollerslev [1998] were the first to raise the question of alternative, more accurate and direct measures of volatility.

The *diffusion* of ultra-high frequency data (UHFD) had the potential of revolutionizing the way volatility is modelled and predicted, since recent research

shows that daily conditional volatility can be consistently estimated by exploiting UHFD. Now instead of using complicated models for unobserved volatility one can use more straightforward models for the estimated time series.

Starting from Andersen and Bollerslev [1998] therefore, a substantial stream of the literature was devoted to ways of measuring the volatility in a hypothetical continuous time process for log–prices. In the course of this chapter, we will focus on the possibility that together with a Brownian motion, a jump process may be suitable to accommodate sudden changes in prices, following specific and exceptional information arrival on the markets. This may give rise to the inclusion in the model specification of separate measures for the *Brownian* risk and the *jump* risk, as the two components of the overall risk affecting an asset, improving the forecasting accuracy of the model of interest. Therefore we focus on estimators which let us to disentangle and consistently estimate the two volatility components.

Although this field of research is rather new, the literature on this topic is large and rapidly evolving. In this chapter we review all (to the best of our knowledge) the non-parametric methods proposed in the literature to estimate the two risk components. The main purpose of it, is to carry out an extensive Monte Carlo comparative survey to investigate the small sample properties of the considered estimators, aiming at identifying the best solution.

2.2 The Basic Framework

2.2.1 Popular models used in finance

The models mostly used in finance for describing the log-price of an asset or a spot interest rate fall within the class of the Itô semimartingales. A *semimartingale* (SM hereafter) can be represented as $X_t = X_0 + A_t + L_t$, where A is a process[1] with paths of *finite variation* (fV), L is a *local martingale* (this is a little bit more general than a *martingale*, see Protter [2005]) and X_0 is the process initial condition. This model is consistent with the assumption that in the market

[1] A indicates the whole process $\{A_t\}_{t\in\mathbb{R}}$, $\mathbb{R}$ being the set of the real numbers, while A_t denotes the random variable describing the state of the process at time t.

no arbitrage opportunities are allowed (Madan [2001]). Part A includes drift components of X and jump components with finite variation, while L typically involves Brownian parts and small jumps. In particular, Itô SMs as in (2.1) below are especially used, since many of their features are now known (Jacod and Shiryaev [2003]) and are suitable for mathematical tractability. Nevertheless some authors use models lying outside the mentioned classes.

Let us consider a unidimensional Itô SM X defined on a filtered probability space $(\Omega, (\mathcal{F}_t)_{t\in[0,T]}, \mathcal{F}, X)$. X can be described by a process starting at a r.v. X_0 at time 0 and evolving in time as (Jacod [in press])

$$dX_t = a_t dt + \sigma_t dW_t + dJ_t + dM_t,\ t \in]0, T], \tag{2.1}$$

where W is a standard Brownian motion, a and σ are adapted càdlàg[1] processes and J, M are jump components describing respectively the "large" and the "small" jumps of X. With the SM representation given previously we would have $A_t = \int_0^t a_s ds + J_t, L_t = \int_0^t \sigma_s dW_s + M_t$. Assuming (2.1) means that the evolution of X is driven by a drift component $\int_o^t a_s ds$, with instantaneous positive or negative rate a_t, which is perturbed by Brownian zero mean shocks and also by the occurrence of, usually unforeseeable, jumps. The Brownian perturbations, which are *predictable*, are usually small and represent the uncertainty underlying the markets. The inclusion of the jumps has been shown to be largely consistent with various financial return series (see the references in Dobrev [2007]): large jumps reflect the arrival of surprising news, or of abnormally large orders, while for some assets small jumps seem to be necessary to better explain the observed data (see Lahaye et al. and the references therein for analyses of the determinants of the jumps in specific assets).

The key tool to model the arrival of jumps is a *random measure* $\mu(\omega, dx, ds)$ defined on $\Omega \times \mathbb{R} \times \mathbb{R}_+$ and with values in $\mathbb{N} \cup \{0\}$, $\mathbb{N}$ being the set of the natural numbers. Each jump size that the model is allowed to realize is labelled by a *mark* $x \in \mathbb{R}$, while $\mathbb{R}_+$ contains time indices. The number $\mu(\omega, B, [0, t])$ reveals how many jumps the path $X(\omega)$ did up to time t and with sizes marked

[1] having right continuous paths with left limits at each time t; *càdlàg* is the French acronym of continuè à droite limitè à gauche.

by labels within B. For instance a process described by $\int_0^t \int_{|x|>1} x\mu(\omega, dx, ds)$ is a *pure jump Lévy* process with marks coinciding with the jump sizes, which are larger than 1. More generally the jump sizes γ of an SM can depend on ω and time, e.g., a pure jump process with jumps larger than 1 can be expressed by $J_t(\omega) = \int_0^t \int_{x:|\gamma(\omega,x,s)|>1} \gamma(\omega, x, s)\mu(\omega, dx, ds)$. Denoting by $\Delta X_t = X_t - X_{t-}$ the jump size at t, on a given ω only one jump size, say the size labelled by $\bar{x}$, is realized (it holds that $\forall t, \mu(\omega, \mathbb{R}, \{t\}) \in \{0, 1\}$), thus $\Delta X_t = \gamma(\omega, \bar{x}, t) = \int_{\{t\}} \int_{\mathbb{R}} \gamma(\omega, x, s)\mu(\omega, dx, ds)$. For μ and γ the ω dependence is suppressed in what follows, for brevity.

Using $\{\int_0^t \int_{\mathbb{R}} \gamma(x, s)\mu(dx, ds), t \in [0, T]\}$ allows us only to describe pure jump Itô SMs with fV, i.e., satisfying a.s. $\sum_{s\leq T} |\Delta X_s| < \infty$, while in general the above integral might not converge. Particular fV jump processes are the *finite activity* (FA) ones, where a.s. for any finite time interval $[a, b]$ the number $N_b - N_a = \int_a^b \int_{\mathbb{R}} 1\mu(dx, ds)$ of jumps which have occurred within it is finite. When this condition is violated (i.e., on some paths we have that, up to some time instant t, $\int_0^t \int_{\mathbb{R}} 1\mu(dx, ds) = \infty$) the jump process is said to have *infinite activity* (IA). For any SM, for any fixed $\varepsilon > 0$, the jumps with size larger than ε give an FA jump process (Cont and Tankov [2004]). When infinite variation (iV) jumps can occur, then in order to make $\int_0^t \int_{\mathbb{R}} \gamma(x, s)\mu(dx, ds)$ converge, further conditions have to be imposed on the small jumps of X (e.g., the ones smaller than 1). Therefore the jump part is given by J, representing the sum of the rare FA jumps larger than 1 and M, being the sum of possibly IA jumps smaller than 1.

Various representations are possible for a unidimensional Itô SM (Jacod [in press]). For instance when X has FA jumps, $M \equiv 0$ and a.s. $J = \int_0^t \int_{\mathbb{R}} \gamma(x, s)\mu(dx, ds) = \sum_{j=1}^{N_t} \gamma_j$, where the Poisson process $N_t(\omega)$ counts the number of jumps which have occurred for path ω from 0 up to time t, and γ_j is the size of the jth jump. When $\sigma_t \equiv 0$, the model is said to be a *pure jump* process, which moves only through jumps, other than an eventual drift, as for instance the 'Carr, Geman, Madan and Yor' (CGMY) processes.

Fundamental examples included in (2.1) are the *Lévy processes*, where a and σ are constant and $\gamma(\omega, x, s) \equiv x$. In particular, when the jumps are of FA, the sizes γ_j are iid and independent of N. In this subclass we mention the Bachelier model with $a = 0$, $J = M \equiv 0$, the Black and Scholes model where $J = M \equiv 0$, and the

Merton and Kou models. On the other hand, the CGMY model is an example of Lévy process with IA jumps. Other Itô SM models outside the Lévy class are the *Brownian semimartingales* (BSMs) where $J = M \equiv 0$ but a and σ are allowed to be stochastic (*stochastic volatility models*) and the *diffusion models*, where $J = M \equiv 0$ and $a_t \equiv a(t, X_t)$ and $\sigma_t = \sigma(t, X_t)$ with deterministic functions $a(t, x), \sigma(t, x)$ ensuring that X is Markov. Models resulting by adding FA jumps to diffusions are called *jump-diffusions*. In general the Itô SM assumption allows for dependency among the jump sizes, the number of jumps which have occurred, and the other X components.

Actually, we can observe one path $X(\omega)$ on a finite and fixed time horizon $[0, T]$, and we only have a discrete record $\{X_0, X_{t_1} \dots X_{t_{n-1}}, X_{t_n}\}$ of $n + 1$ data. X is the *data generating process* (DGP). In most of the methods we exposit, the observations are *evenly spaced*, with $t_i = ih$ for a given temporal mesh h, and $n = \lfloor T/h \rfloor$, the integer part of T/h. By departing from more common use of the terms in econometrics (low frequency = weekly and above, high = daily, ultra–high = intra–daily), we indicate different magnitudes of h by using LF, respectively HF or UHF to mean that the data is sampled at *low frequency* (more or less $h > 15$ min), respectively *high frequency* ($h \in [1 \text{ min}, 15 \text{ min}]$) or *ultra high frequency* ($h < 1$ min). The huge variety of models existing in the literature allows for very different path properties (e.g., with finite/infinite variation, being continuous/discontinuous, allowing for finite/infinite jump activity), which turn out to have deeply different implications. This makes relevant the problem of model selection and of coefficient estimation.

2.2.2 Simulated models used in this chapter

We illustrate specific features of the data and of the estimators we present by using real and simulated observations. We now describe the models we simulate, which are common in the financial and econometric literature. X stands for the log-price of an asset, in Models 1 and 2 the parameters are given on a daily basis, and are such that the generated returns are expressed in percentage form, while in Model 3 they are annual and the returns are in absolute form. The number n of generated observations is either 1000 or 25200 for the different applications below.

The simulation of each path is obtained by discretizing the stochastic differential equation (SDE) giving the considered model with the Euler scheme and 1 sec temporal step; the increments of the jump parts are simulated as explained in Cont and Tankov [2004]; data is then aggregated to obtain the desired constant mesh h (which can be five minutes or something else, as specified below). For each model $X_0 \equiv 0$.

- *MODEL 1*, called HT1FJ, is taken from Huang and Tauchen [2005], and has stochastic volatility (correlated with the Brownian motion driving X) and FA jumps:

$$dX_t = 0.03dt + e^{0.125v_t} dW_t^{(1)} + dJ_t,$$

where

$$dv_t = -0.1v_t dt + dW_t^{(2)} \tag{2.2}$$

$v_0 = 0$, $W^{(1)}, W^{(2)}$ are standard correlated Brownian motions with correlation coefficient $corr(W_1^{(1)}, W_2^{(2)}) = -0.62$, and $J_t = \sum_{i=1}^{N_t} \gamma_i$ is an FA compound Poisson process with iid $\mathcal{N}(0, 0.5^2)$ sizes of jumps, where $\mathcal{N}$ indicates a Gaussian random variable (r.v.) and N has constant jump intensity $\lambda = 0.014$. The volatility factor v is a mean reverting Ornstein–Uhlenbeck process tending to push σ towards its *steady state*, $e^0 = 1$.

- *MODEL 2*, called HT2F, also is taken from Huang and Tauchen [2005]. It has continuous paths and the stochastic volatility is determined by two factors. This allows σ to have an erratic behaviour such as to produce a rugged appearance of the price series and consequently high values of the returns $\Delta_i X = X_{t_i} - X_{t_{i-1}}$. Thus this model represents an interesting challenge for the methods aiming at recognizing returns where jumps have occurred.

$$dX_t = 0.03dt + sexp[-1.2 + 0.04v_t^{(1)} + 1.5v_t^{(2)}]dW_t^{(0)},$$

$$dv_t^{(1)} = 0.00137v_t^{(1)}dt + dW_t^{(1)}, \quad dv_t^{(2)} = -1.386v_t^{(2)}dt + [1 + 0.250v_t^{(2)}]dW_t^{(2)},$$

where $v_0^{(1)} = v_0^{(2)} = 0$, $corr(W^{(0)}, W^{(1)}) = corr(W^{(0)}, W^{(2)}) = -0.3$, and the function $sexp(x) = e^x I_{\{x \le k\}} + e^k \sqrt{k - k^2 + x^2}/\sqrt{k}\ I_{\{x > k\}}$, with $k =$

$\ln\sqrt{10000/252}$, is built in order to ensure that the coefficients of the SDE for X satisfy growth conditions such that a solution of the system exists and the Euler scheme works well.

- *MODEL 3*, called Gauss-CGMY, is taken from Carr et al. [2002], with the parameter values estimated for the SPX index (futures on the S&P500 index).

$$dX_t = 2 \times 10^{-10} dW_t + dM_t,$$

where M is a CGMY process, which is an IA jump process. Also this model represents a difficult challenge for the methods we present for estimating IV_T, because it contains a negligible Brownian part, thus in finite samples the jumps can be easily confused with the movements of a BM.

The first row of figure 2.1 displays the log-prices of the SPX for the period 1994/12/23–1999/3/9 and one path of simulated data for the three models presented above. The second row shows the corresponding returns series. This figure suggests that the three DGPs produce realistic time series of log-prices, the more realistic one being probably HT1FJ.

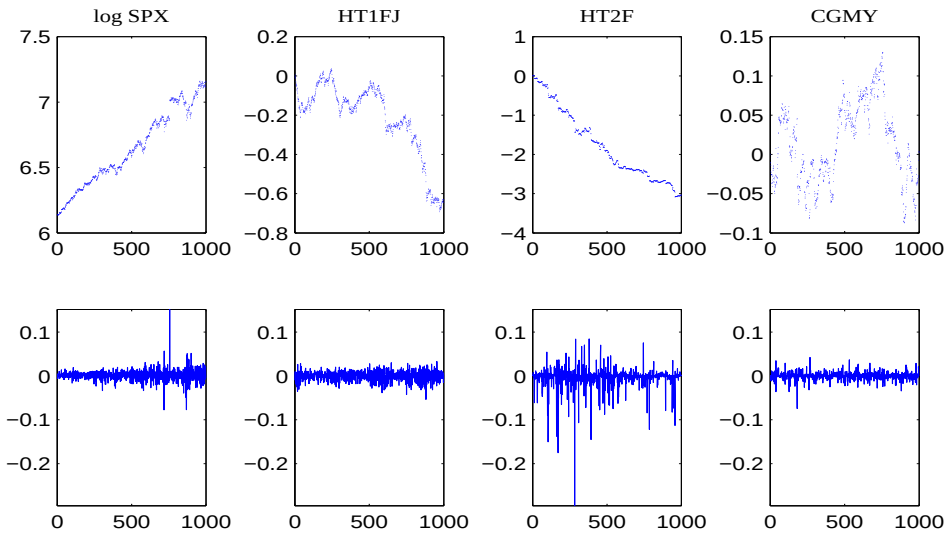

Figure 2.1: Realized paths of log-price (first row) and log-return $X_{t_i} - X_{t_{i-1}}$ (second row) for 1000 daily observations (four years) of SPX (first column) and simulated data from the models HT1FJ (second column), HT2F (third column) and Gauss-CGMY (third column).

2.2.3 Realized variance and quadratic variation

The *realized variance* of $X = \log S$ is defined as $RV_n = \sum_{i=1}^{n}(\Delta_i X)^2$ and is a way of measuring the risk of S, the price of an asset. In fact let us think of the case where $J + M \equiv 0$, then as $h \to 0$, $RV_n \xrightarrow{P} IV_T = \int_0^T \sigma_t^2 dt$, the *integrated variance* up to time T, where σ modulates how W impacts on X. More generally, assuming X as in (2.1) with $M \equiv 0$, we have

$$RV_n \xrightarrow{P} [X, X]_T = IV_T + [J, J]_T, \tag{2.3}$$

where $[J, J]_T = \sum_{s \leq T}(\Delta X_s)^2 = \int_0^T \int_{\mathbb{R}} \gamma^2(x, s)\mu(dx, ds)$ is the sum of the squared jump sizes realized by X on ω within the time horizon $[0, T]$ and modulates how the jump risk impacts on X. This notion can be easily extended to include the IA jump component since for any SM, while the sum of the jumps need not converge, the sum of their squares does so, a.s.

$[X, X]_T$ is called the *quadratic variation* (QV) of X and has been adopted, since 1998, in the literature as a measure of the impact of the overall risk underlying X (Andersen and Bollerslev [1998]). Before then, risks were quantified by some parameters within a specific model for X (e.g., by the constant σ in the Black and Scholes model; by σ and the standard deviation σ_J of the jump sizes in the Merton model). $[X, X]_T$ is defined path by path, so that it can be assessed simply by looking at one trajectory of X and has the advantage of being a model-free notion. If for instance X is a simple Poisson process N, then the N_T jumps which have occurred up to T all have size 1, and the sum of their squares is the same N_T, i.e., $[N, N]_T = N_T$. If X is a Lévy process then $[X, X]_T = \sigma^2 T + \int_0^T \int_{\mathbb{R}} x^2 \mu(dx, ds)$. Many properties of the quadratic variation of an SM can be found in Cont and Tankov [2004]; Jacod and Shiryaev [2003].

In the literature, $\sqrt{\sum_{i=1}^{n}(\Delta_i X)^2}$ is usually called *realized volatility*, while by *volatility* we indicate the, possibly stochastic, coefficient σ of (2.1). As RV_n is a consistent estimator of QV, $\sqrt{RV_n}$ estimates $\sqrt{QV}$. The smaller h, the closer we expect $\sqrt{RV_n}$ to be to $\sqrt{QV}$. The plot of $\sqrt{RV_n}$ of an asset against the step h between the observations is called the *signature plot* (SP, see figure 2.2, bottom panels) and is used in the econometric literature to visualize at which fine scale h the estimated values $\sqrt{RV_n}$ stabilize. The mesh h at which stability begins (e.g.,

$h = 10$ min in figure 2.2, left-bottom panel) is considered a proper mesh at which the estimate of $\sqrt{QV}$ is reliable.

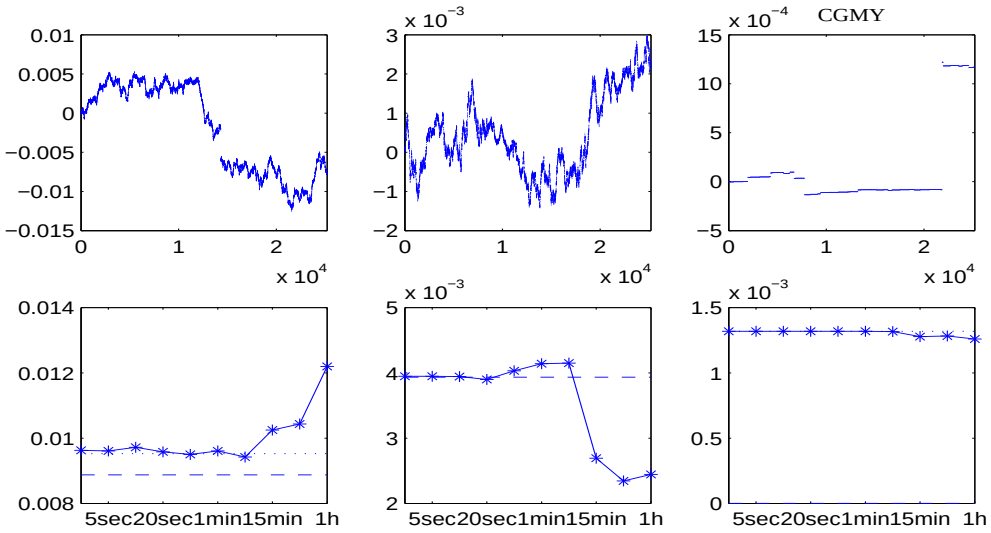

Figure 2.2: First row: simulated one-day paths of models HT1FJ, HT2F and Gauss-CGMY for the evolution of the log-price of an asset, built from 25,200 one-second observations; second row: relative signature plots of $\sqrt{RV_n}$ at the frequencies of $h = 1$ hour; 30, 15, 5, 1 min and 30, 20 ,10, 5, 1 sec.

2.2.4 Importance of disentangling.

$\sum_i (\Delta_i X)^2$ is a measure of the global risk affecting an asset. However Brownian risk and jump risk are amplified by different coefficients. In particular, the price of a derivative product written on a jumping asset needs different risk premia for the two components, and capturing and measuring IV_T and $[J, J]_T$ separately allows premium assessment and a more precise hedge of risk (Wright and Zhou [2007], Brownlees and Gallo [2010]). Secondly, a separate estimation of IV_T serves for model selection purposes (e.g., for testing the necessity of a jump component in a model, as in Barndorff-Nielsen and Shephard [2006]). Thirdly, including separate measures for the contributions of IV_T and $[J, J]_T$ in econometric models for the future evolution of X significantly improves the accuracy of volatility forecasts, especially in periods following the occurrence of jumps (Andersen et al. [2007]; Corsi et al. [2010]).

2.2.5 Further notation

$\sigma.W$ indicates the stochastic integral process of σ wrt W; given the representation (2.1), we indicate $D_t = \int_0^{\cdot} a_s ds + \sigma.W$; $IV_t = \int_0^t \sigma_s^2 ds$, while for brevity IV_T is also indicated by IV; given two deterministic functions f and g, $f(h) \sim g(h)$ means that both $f = O(g)$ and $g = O(f)$ as $h \to 0$; given two random functions $f(\omega, h), g(\omega, h)$, $f = o_P(1)$ means that f tends to zero in probability as $h \to 0$, $f = O_P(1)$ means that f is bounded in probability, $f \sim g$ means that both $f = O_P(g)$ and $g = O_P(f)$; when not ambiguous we suppress the n-dependence of the estimators we present below, e.g., RV_n is also indicated by RV; *rhs* stands for *right hand side*.

2.3 Disentangling Browninan and Jump Risk: Estimators of Integrated Variance

If we can observe a process only every h units of time, data will appear discrete and thus as a succession of jumps. However there are different degrees of discontinuity and some discrete data is compatible with continuous sample paths, some is not (Aït-Sahalia [2002]). Many methods have been proposed for disentangling IV from the jumps contribution in the QV of a given asset model. When we assume a parametric model for the evolution of an asset (for instance a Heston model, or CGMY model), we are uncertain whether the DGP in fact belongs to the selected class of models. Thus non-parametric approaches are particularly important, as they are applicable to many sub-classes of SMs. Anyway, even non-parametric methods usually need some assumptions restricting the model classes the DGP is allowed to belong to. Trying to generalize as much as possible the framework where it is possible to obtain a consistent estimator $\hat{IV}_n$ and to evaluate its efficiency (the speed at which it converges) is thus not merely the passion of mathematicians for theoretical disquisitions, but serves a genuine need to limit the risk of excluding models for our DGP. For instance the question is open as to whether FA jumps are sufficient to describe the evolution of an asset price or IA jumps have to be considered. Most models in finance including jumps consider an FA jump component. However Lee and Hanning [2010] and

Aït-Sahalia and Jacod [in press] show that for some assets the inclusion of an IA jump component could be sensible. Aït-Sahalia and Jacod [in press] proposes a test to check finiteness versus infiniteness of jump activity. We cannot exclude in general the presence of IA jumps at the moment, so it is important to check whether a proposed estimator of IV is robust to them or not. We concentrate only on non-parametric methods allowing of considering rather general stochastic coefficients. Some methods (such as, e.g., threshold, wavelet, outlyingness hard rejection weighting) aim to recognize the jump occurrences so to eliminate the jump contaminated returns and to apply methods proper for a BSM. On the other hand other methods (such as the *MultiPower Variations*, MPVs) keep all the returns but aim to dampen the jumps impact.

We concentrate on estimating IV in the presence of jumps, however in practice also the estimation of $IP_T(p) = \int_0^T \sigma_s^p ds$ with $p \neq 2$ and of spot volatility σ_s is sometimes needed. For instance, for a given estimator $\hat{IV}_n$, through a CLT, it is typically shown that the asymptotic variance of the estimation error $\hat{IV}_n - IV$ involves the *Integrated Quarticity* $IQ = IP_T(4) = \int_0^T \sigma_s^4 ds$ (see, e.g., (2.4)). For any $p > 0$, we can use the MPVs, or the *Threshold power variations*, which are treated below.

Non-parametric estimation of spot volatility in the presence of jumps in the returns (even allowing σ to jump) given discrete observations has been done using: kernels (Bandi and Nguyen [2003]; Bandi and Renó [2010]; Mancini and Renó [in press]) or more general delta sequences (Mattiussi and Renó [2010]); MPVs within local windows (Ngo and Ogawa [2009]); *Threshold realized variance* (TRV) in local windows (Aït-Sahalia and Jacod [2009]); and a duration-based technique (Andersen et al. [2009b]).

RV is the privileged measure of IV in the absence of jumps (see, e.g., Andersen et al. [2003]) as the sum of the squared increments represents a natural way to measure the variability of X, it is simple to be computed and efficient (in the Cramer–Rao inequality lower bound sense, see Aït-Sahalia and Jacod [2008] for the efficiency rates in a Lévy model framework and Jacod and Protter [1998] for BSMs), as it is the maximum likelihood estimator of QV in the parametric framework. As, by (2.3), in the presence of jumps RV mixes up IV information and jump information, modifications of it in many different directions have been

proposed to filter the jumps out.

We repeat that we consider here only the case of equally spaced observations as most of the proposed estimators of IV adopt such an assumption. In what follows we have tried to unify the notation from the different cited papers, so the names we give to the important variables do not always correspond to the names in the original articles.

2.3.1 Bipower and Multipower Variation

The estimator of IV most widely used is the realized *Bipower Variation*

$$BPV_n(T) = \frac{\pi}{2} \sum_{i=2}^{n} |\Delta_i X||\Delta_{i-1} X|,$$

introduced in Barndorff-Nielsen and Shephard [2004]. When not ambiguous, we will suppress the time T dependence. If no jumps occur within $]t_{i-1}, t_{i+1}]$ and $\Delta_i X$, $\Delta_{i+1} X$ are iid Gaussian $\mathcal{N}(0, \sigma^2 h)$ with constant σ, then $E[|\Delta_i X||\Delta_{i+1} X|] = \frac{2}{\pi}\sigma^2 h$. Therefore, each term in the above sum, if divided by h, is an unbiased estimator of the local spot variance $\sigma^2_{t_i} \equiv \sigma^2$ and the sum of the terms $|\Delta_i X||\Delta_{i-1} X|$ approaches in probability $\sum_{i=1}^{n} \sigma^2_{t_i} h$, a Riemann sum converging to IV when $h \to 0$. If the jumps have FA, for sufficiently small h, within each $]t_{i-1}, t_{i+1}]$ at most one single jump will occur, affecting only one of the two returns $\Delta_i X$, $\Delta_{i+1} X$. Suppose the jump occurs within $]t_{i-1}, t_i]$, then the multiplication of $|\Delta_i X|$ with the adjacent $|\Delta_{i+1} X| = O_P(\sqrt{h})$ will dampen the jump's impact. As a.s. for the given ω only $N_T(\omega)$ jumps occur within $[0, T]$, the total contribution of the terms involving jumps is $O_P(N_T(\omega)\sqrt{h})$ and tends to zero in probability as $h \to 0$. In fact in Barndorff-Nielsen and Shephard [2006] it is shown that BPV_n converges to IV in probability, when the jumps have FA and the drift and volatility processes (a, σ) are jointly independent of W (Theorem 2). If moreover $J \equiv 0$, it is also shown that (Theorem 3) conditionally on (a, σ) we have

$$\frac{BPV_n - IV}{\sqrt{h}\sqrt{\vartheta_{BPV} IQ}} \xrightarrow{d} \mathcal{N}(0, 1), \tag{2.4}$$

where $\vartheta_{BPV} = \frac{\pi^2}{4} + \pi - 3 \approx 2.609$. Denoting by $AVar_{\hat{IV}}$ the asymptotic variance of $(\hat{IV}_n - IV)/\sqrt{h}$, the above CLT means that the estimation error $BPV_n - IV$ tends to zero as $h \to 0$ with the same order of $\sqrt{h}$ and that $AVar_{BPV}$ is given by $\vartheta_{BPV} IQ$. Unfortunately, the magnitude of ϑ_{BPV} indicates an inefficiency of BPV in estimating IV even in the absence of jumps, as an efficient estimator would have $AVar = 2\ IQ$. Even if small, the inefficiency of BPV can be quite important in the applications (see the figures in Mancini [2009]), especially when the data is given at HF or LF. However, in practice returns at UHF undergo *microstructure noises* (see below), which invalidate the consistency of all the estimators presented in this section. Further, for some assets or commodities, observations are not so frequently available. In all the estimation methods we mention here $AVar_{\hat{IV}}$ has the form $\vartheta_{\hat{IV}} IQ$, with a proper constant term $\vartheta_{\hat{IV}}$, so the efficiency comparisons are simply done through the $\vartheta_{\hat{IV}}$s.

Further measures of the variation of $\sigma.W$ are the **Multipower Variations** (mentioned in Barndorff-Nielsen and Shephard [2006] and subsequently developed):

$$MPV_n^{r_1 \dots r_k} = \prod_{i=1}^{k} \mu_{r_i}^{-1} h^{1 - \frac{\sum_i r_i}{2}} \sum_{i=k+1}^{n} |\Delta_i X|^{r_1} |\Delta_{i-1} X|^{r_2} .. |\Delta_{i-k} X|^{r_k},$$

where $\mu_{r_i} = E(|u|^{r_i}) = 2^{r_i/2} \Gamma(\frac{r_i+1}{2}) / \Gamma(1/2)$ and $u = \mathcal{N}(0,1)$ When the returns are iid Gaussian then each scaled summand gives an unbiased estimator of $|\sigma|^{\sum_i r_i}$. In Woerner [2006] it is shown that $MPV_n^{r_1 \dots r_k}$ consistently estimates $\int_0^T |\sigma_s|^{\sum_{j=1}^k r_j} ds$ under path regularity and independence from W of the drift and volatility coefficients, as soon as $r_i > 0$ for all $i = 1 \dots k$ and $\max_i r_i < 2$. Without the latter condition, the big jumps would not be sufficiently dampened any more, and the normalization by $h^{1 - \sum_i r_i/2}$ would cause the explosion of MPV_n to infinity. Note that the integral $IP_T(\sum_{j=1}^k r_j)$ can be estimated using in turn the multipower variations, for instance the multipower mostly used to estimate IQ is $MPV_n^{4/3,4/3,4/3}$.

The results in Vetter [2010] show that $\max_i r_i < 1$ is a necessary condition to obtain that an MPV estimator of IV converges at rate $\sqrt{h}$. That forces us to use at least three powers. An upper estimation bias is caused when a large jump is

present, as the return is not really vanishing in finite samples. That bias would be even worse in the presence of jumps in contiguous returns, which however has a low probability of happening.

A further drawback found in finite samples is that MPVs are sensitive to the presence of zero returns arising from stale quotes and rounding to a discrete price grid (Andersen et al. [2009a]). In practice zero returns are quite numerous at UHF if the considered asset is not very liquid, as, e.g., US bonds. The occurrence of a zero return in $]t_{i-1}, t_i]$ affects two consecutive terms in BPV_n, as such destroying information about σ.

Much effort has been expended to estimate IV, seeking to improve the efficiency, finite sample performance, robustness even to IA jumps and to noise. Extensions of the BPV in different directions (in the present framework) are for instance the Threshold-BPV, the MinRV, the MedRV, the Bipower Type estimator and the Bipower range, all of which are treated below.

2.3.2 Threshold realized variance

An alternative estimator is the *Threshold Realized Variance* (also called *truncated RV*) of Mancini (Mancini [2001, 2009])

$$TRV_n(T) = \sum_{i=1}^{n} (\Delta_i X)^2 I_{\{(\Delta_i X)^2 \leq r_{t_i}(h)\}},$$

where $r_t(h) = c_t R(h)$, c_t is a stochastic process which is a.s. bounded and bounded away from zero, and $R(h)$ is any deterministic function of the step h between the observations such that $\lim_{h\to 0} R(h) = 0$ and $\lim_{h\to 0} (h \log \frac{1}{h})/R(h) = 0$. $r_t(h)$ is called *threshold*.

The intuition of why TRV_n approaches IV is based on the result of Lévy (Karatzas and Shreve [1999], Theorem 9.25) stating that a.s. the absolute value of the maximal increment $\Delta_i W$ on $[0, T]$ tends to zero at the same speed as the deterministic function $\sqrt{2h \ln \frac{1}{h}}$, as $h \to 0$. An analogous property holds for the maximal absolute increment of $\int a_s ds + \int \sigma_s dW_s$ (Mancini [2009], note that the drift part increments tend to zero at the higher speed h). For any t the threshold function tends a.s. to zero, but more slowly than the function

$2h \ln \frac{1}{h}$, consequently, if for small h, we find that the squared increment $(\Delta_i X)^2$ is larger than $r_{t_i}(h)$, which in turn is higher than $2h \ln \frac{1}{h}$, then it is likely that some jumps have occurred. In fact we have that in the presence of only FA jumps with $P\{\Delta N_t \neq 0, \Delta X_t = 0\} = 0$ then a.s. for a sufficiently small h, depending on the selected ω, we have

$$\forall i = 1 \dots n, \quad I_{\{(\Delta_i X)^2 \leq r_{t_i}(h)\}}(\omega) = I_{\{\Delta_i N = 0\}}(\omega). \tag{2.5}$$

Therefore TRV_n a.s. for sufficiently small h excludes from RV_n the finitely many returns where a jump occurred, allowing estimation of IV. In the presence of also IA jumps, then for all $\delta > 0$ a.s. for sufficiently small h the threshold $r_{t_i}(h)$ cuts off all the jumps of J and the jumps of M larger, in absolute value, than $\delta + 2\sqrt{r_{t_i}(h)}$. However as a.s. $\max_i r_{t_i}(h) \to 0$, since δ is arbitrary, then every jump of M is cut off.

The main advantage of TRV is its efficiency when the jumps have fV, as $\vartheta_{TRV} = 2$. Further, TRV is immune from the zero returns issue. The second advantage of the threshold technique is that, in the presence of only FA jumps, a.s. for h sufficiently small identification of the location of the jumps is possible with high precision (Mancini [2004]). Estimation of the jump sizes and evaluation of the convergence rate is done in Mancini [2009]. On the contrary, using MPVs, the identification of jump locations and sizes is not straightforward. Such a property of the threshold method has important consequences. For example, after removing the time intervals where some jumps occurred, we can adapt known estimation methods for diffusion processes also to jump-diffusion processes (Mancini and Renó [in press]).

Regarding the choice of the *threshold*, a further remark needs to be done. In principle the choice of the terms $r_{t_i}(h)$ would have to depend on a preliminary rough estimate of $\sigma_{t_{i-1}}$, since the magnitude of σ has an impact on determining whether $(\Delta_i X)^2 \leq r_{t_i}(h)$ or not, and thus on the TRV_n finite sample performance. The fact that the estimator depends on unobservable characteristics of the DGP is a typical issue in non-parametric estimation. Further, each model has its own finite sample optimal threshold. However, in principle, in order to construct TRV we need some preliminary information about the same σ. The formal study of

methods for optimal threshold selection in a given model is the object of further research (Mancini [2008]), however it is common now to use power functions ch^α as mentioned above. In figure 2.3 the sensitivity of TRV_n to the choice of c in $r(h) = ch^{0.99}$ is checked on simulations of *Model 1* (see section 2.2.2) for fixed observation steps of $h = 5$ min and 1 sec. On the other hand, allowing $r_t(h)$ to be stochastic as specified above permits us to use an adaptive threshold accounting for volatility persistence: in Mancini and Renó [in press] (h = 1 day) $r_{t_i}(h)$ is 9 times a GARCH forecast of the future $\sigma^2_{t_i} h$ made at time t_{i-1}; in Corsi et al. [2010] an iterative estimation of σ_{t_i} is implemented (see below). Another common choice is $r_t(h) = 9BPV_n h^{0.98}/T$ for all $t \in [0, T]$ (see, e.g., Jacod and Todorov [2009]). In Table 2 we call BTV, *Bipower Threshold Variation*, the threshold estimator with the choice $r_t(h) = 9BPV_n h^{0.99}/T$.

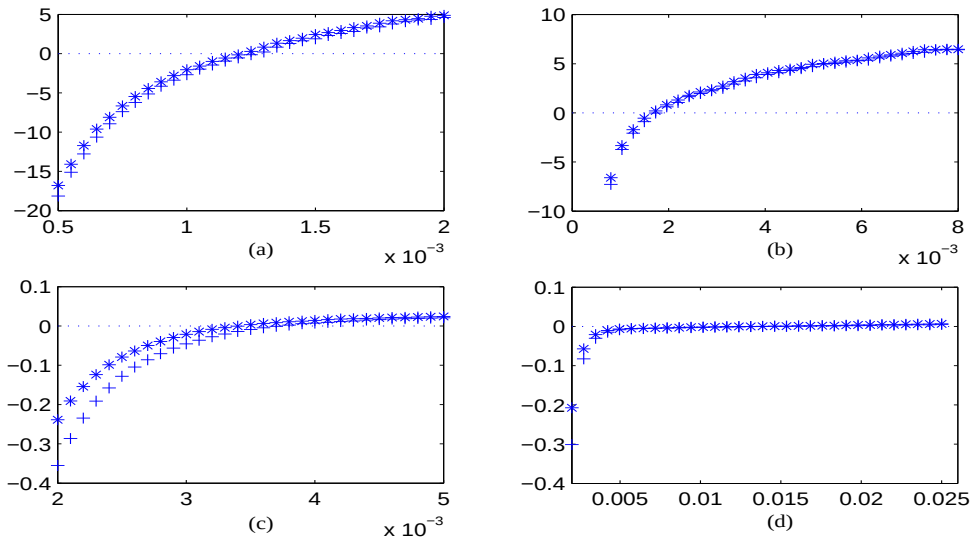

Figure 2.3: Mean of the relative percentage estimation bias $100(\hat{IV} - IV)/IV$ for TRV and TBV as the c within the thresholds $r(h) = ch^{0.99}$ (*) and $r(h) = ch^{0.999}$ (+) varies. (a) TRV with $h = 5$ min, (b) TBV_n with $h = 5$ min, (c) TRV with $h = 1$ sec, (d) TBV with $h = 1$ sec. For any c in the x-axis, 1000 daily paths of Model 1 are simulated, where J is constrained to have one jump each day. The daily estimation bias is then averaged.

2.3.3 Threshold-Bipower Variation

By combining the Threshold and Bipower criteria we improve both the robustness to the threshold selection and the finite sample efficiency of BPV_n in estimating IV in the presence of jumps. Corsi et al. [2010] define

$$TBV_n = \frac{\pi}{2} \sum_{i=1}^{n-1} |\Delta_i X| I_{\{(\Delta_i X)^2 \leq r_{t_i}(h)\}} |\Delta_{i+1} X| I_{\{\Delta_{i+1} X \leq r_{t_{i+1}}(h)\}} \tag{2.6}$$

under general predictable drift coefficient a, càdlàg volatility σ, $M \equiv 0$, iid J jump times and sizes. The threshold function can be stochastic and different in different intervals $]t_{i-1}, t_i]$, as soon as it satisfies the boundedness requirements mentioned above. The entire class of the MPVs is generalized by replacing the returns by "thresholded" returns. In practical applications, for h fixed at the level of 5 min, $r_{t_i}(h) = c^2 \hat{\sigma}^2_{t_i} h$ is chosen, where $c = 3$ and the spot σ is recursively estimated by averaging 22 "thresholded" returns in a bilateral local window containing t_i and excluding the returns at times t_{i-1}, t_i, t_{i+1} so to avoid having an eventual jump occurring close to t_i impacting on $\hat{\sigma}^2_{t_i}$. In figure 2.3 we implement both TRV and TBV with threshold $ch^{0.99}$ and we show how the robustness to the choice of c increases by using TBV in place of TRV.

On the other hand, assuming FA jumps and $r_t(h) \equiv R(h)$, as a.s. asymptotically TBV_n cuts off all the discontinuities, the CLT for TBV_n gives convergence rate $\sqrt{h}$ and $AVar = \vartheta_{TBV} IQ$ with $\vartheta_{TBV} = \vartheta_{BPV}$, as for BPV_n in the no-jumps case. In Corsi et al. [2010] a finite sample performance comparison is reported among TRV, BPV, *staggered* BPV, TBV and a corrected version of TBV. It is shown that in fact for the purposes of estimation of IV, the less biased estimator is TBV. Very interestingly the authors show that using the latter measure of IV significantly improves volatility forecast power over the Andersen et al. [2007] approach. In the latter paper, contrarily to Corsi et al. [2010], the authors conclude that jumps have a negative or not significant impact in determining the future volatility, however to measure the contribution of IV they use the not efficient BPV.

2.3.4 Other methods

All methods below, apart from the Generalized Range (GR), similarly as the threshold method, are based on the sum of properly selected or weighted squared returns. For all the methods asymptotic properties are shown (theoretically and/or numerically) only in the presence of FA jumps in returns.

In Christensen et al. [2010], a **Quantile based** approach is proposed. The data sample is divided into n/m groups of m adjacent returns; $k+1$ percentage numbers $\lambda_1 \dots \lambda_{k+1} \in (1/2, 1)$ are chosen and the estimator of IV is defined by

$$QRV_n = \sum_{j=1}^{k+1} \alpha_j \frac{m}{n} \sum_{i=1}^{n/m} \frac{(\Delta_{\ell_{i,\lambda_j}} X)^2/h + (\Delta_{\bar{\ell}_{i,\lambda_j}} X)^2/h}{\nu(m, \lambda_j)},$$

where the selected returns are empirical quantiles of the ith group of returns, $\Delta_{\ell_{i,\lambda_j}} X$ being the $(\lambda_j \cdot m)$th order statistic and $\Delta_{\bar{\ell}_{i,\lambda_j}} X$ the $(m - \lambda_j m + 1)$th order statistic of the returns in group i. Each scaling factor $\nu(m, \lambda_j)$ is the expectation of the term at its numerator when all the normalized returns $\Delta_j X/\sqrt{h}$ are Gaussian, k is fixed and α_j are optimally chosen weights to minimize the AVar. Large returns not consistent with Gaussian realizations are ignored as soon as all λ_j are < 1, which enables the estimator to be robust both to large jumps and to the presence of data outliers. For small observation step h, each block contains at most one jump, the volatility within the block is approximately constant (in the assumed framework) and each term $\left[(\Delta_{\ell_{i,\lambda_j}} X)^2/h + (\Delta_{\bar{\ell}_{i,\lambda_j}} X)^2/h\right]/\nu(m, \lambda_j)$ estimates the scaled return variance over the ith block. Special cases in this class of estimators are *MinRV* and *MedRV*. In the presence of FA jumps for any fixed m the estimator is consistent as $n \to \infty$ and, assuming σ to be an Itô continuous SM, it converges at speed $\sqrt{h}$ with ϑ_{QRV} which is close to 2.4 when $k = 4$ and the λ_j are in the range $[0.8, 0.95]$. Efficiency is improved by allowing for higher k, or for overlapping blocks, but it deteriorates when m increases. A finite sample performance comparison on simulated data among QRV, RV, BPV, TRV and MedRV can be found in the paper, where about two observations per minute are used.

Andersen et al. [2009a] propose **MinRV** and **MedRV**, defined by

$$MinRV_n = \frac{\pi}{\pi - 2}\frac{n}{n-1}\sum_{i=1}^{n-1} min(|\Delta_i X|, |\Delta_{i+1} X|)^2,$$

$$MedRV_n = \frac{\pi}{6 - 4\sqrt{3} + \pi}\frac{n}{n-2}\sum_{i=2}^{n-1} median(|\Delta_{i-1} X|, |\Delta_i X|, |\Delta_{i+1} X|)^2,$$

where in each case the first factor makes each summand an unbiased estimator of the spot variance $\sigma^2_{t_i}$ if the returns involved in the block are iid Gaussian (thus with the same σ). As, for sufficiently small h, jumps in contiguous intervals $]t_{i-1}, t_i]$ will not occur, the returns affected by large jumps are completely ignored, which improves the estimators finite sample performances over the BPV. The consistency of both estimators is proved in the presence of FA jumps, as soon as σ is a.s. bounded away from zero. Under the further assumption (Itô σ) a CLT holds with $\vartheta_{MinRV} = 3.81$, $\vartheta_{MedRV} = 2.96$. Intuitively, the better efficiency of MedRV is due to the fact that large returns contain more information about σ than small returns, but they are thrown away in MinRV. Further, in finite samples, contrarily to MedRV, MinRV still suffers from a similar exposure to zero returns as BPV.

Realized Outlyingness Weighted Variation. Boudt et al. [2010] modify RV by considering weighted returns, with small or zero weight given to local outliers, i.e., returns which are extreme in amplitude with respect to the neighbours within a local window. The following estimator

$$ROWV_n = c_w \sum_{i=1}^{n} w(d_{i,h})(\Delta_i X)^2$$

is defined, where: w is a weight function, the preferred one is the *hard rejection function* $w_{HR,\beta}(z) = I_{\{z \leq k_\beta\}}$, k_β being the β quantile of the χ^2 law, the recommended β value being 0.999; c_w is a correction factor explicitly given in terms of w and ensuring consistency in the considered framework; $d_{i,h} = \Delta_i X^2/(h\ \hat{\sigma}^2_{t_i})$ measures the distance of the squared normalized return from 0; $\hat{\sigma}^2_{t_i}$ is a first step *Least Trimmed Squares* (LTS) estimator of spot $\sigma^2_{t_i}$ obtained using the returns in a local window around $\Delta_i X$, so that $d_{i,h}$ would be asymptotically χ^2 distributed

if the underlying model was Gaussian with constant σ.

The convergence rate is shown to be $\sqrt{h}$ when $J \equiv 0$ and no leverage is present. The AVar constant ϑ_{ROWV} is explicitly given and tabulated in the web appendix in terms of the weight function w: ϑ_{ROWV} is decreasing in β, $\vartheta_{ROWV,w_{HR},\beta=0.999} = 2.152$ and $\vartheta_{ROWV,w_{HR},\beta=1} = 2$ is minimal. However $\beta = 1$ is an infeasible choice in the presence of jumps, because in that case $ROWV$ would coincide with RV, which is not consistent to IV. The extension of this estimator to the multivariate asset context has nice properties, which motivates proposing the outlyingness measure. In the bivariate case the finite sample performance on simulated data is good already at HF, which in principle allows us to bypass the distortions connected with the presence of microstructure noise in the data.

Range Bipower Variation. Christensen and Podolskij [2010] modify BPV by replacing each $\Delta_i X$ by the maximum return between two any time instants of a predefined time grid within $]t_{i-1}, t_i]$. More precisely the available n observations are divided into N blocks of m data, so that each $]t_{i-1}, t_i]$ contains m observed returns. For the i^{th} block

$$s_{X,i,h,m} = \max_{t_\ell, t_u:\ t_i \le t_\ell < t_u \le t_{i+1}} |X_{t_\ell} - X_{t_u}|$$

is taken. Considering the range $s_{W,0,1} = \max_{s,u:\ 0\le s<u\le 1} |W_s - W_u|$ of a BM continuously observed on $[0,1]$ for financial modelling purposes goes back to the eighties (see the references in Dobrev [2007]). We consider here the feasible discretized version. With constant σ and $X = \sigma W$ we have $E[s^r_{\sigma W,i,h,m}] = \sigma^r h^{r/2} \lambda_{r,m}$, with $\lambda_{r,m} = E[s^r_{W,0,1,m}]$ (see Christensen and Podolskij [2007]), and the *Realized Range Variation* $RRV_n = \lambda^{-1}_{2,m} \sum_{i=1}^N s^2_{X,i,h,m}$ consistently estimates the IV of an Itô BSM X, for any m, taking $N \to \infty$, and the convergence rate is $\sqrt{N} = \sqrt{\frac{n}{m}}$. The constant $\lambda_{2,m}$ does not have an explicit form, but can be obtained numerically with arbitrary precision, a plot of $\lambda_{2,m}$ letting m vary is displayed in Christensen and Podolskij [2007]. However in the presence of FA jumps RRV_n is shown to tend to $IV + \lambda^{-1}_{2,m}[J,J]_T$ (Christensen and Podolskij [2010]), so in order to estimate IV

the following *Range based Bipower Variation* is computed:

$$RBV_{n,m} = \frac{1}{\lambda_{1,m}^2} \sum_{i=1}^{N-1} s_{X,i,h,m} s_{X,i+1,h,m}. \tag{2.7}$$

This is shown to be consistent, for any m, as $N \to \infty$, while a CLT is shown to hold only in the absence of jumps. As $n \to \infty$ and m tends to a finite integer c or to c=$+\infty$, the convergence rate is shown to be $\sqrt{n/m}$, which is strictly less than $\sqrt{n}$ if $m > 1$. The $\vartheta_{RBV,c} = \vartheta_{BPV} = 2.6091$ if $m = c = 1$, while it decreases in c and equals 0.3631 if $c = \infty$ (see figure 1 in Christensen and Podolskij [2010]). However $c = +\infty$ corresponds to the unfeasible situation where we can observe continuously X over $[0, T]$. In sum, as we refine the partition within each $]t_{i-1}, t_i]$ (c increases) $\vartheta_{RBV,c}$ decreases, but the convergence rate decreases as well.

Dobrev [2007] proposes the following alternative **Generalization of the realized Range** of X. The n observed returns are here not necessarily evenly spaced, and we set $h = \max_i |t_i - t_{i-1}|$. A fixed number $k < n$ of returns is selected so as to reach the maximal variation

$$GR_{X,n} = \max_{0 \le t_1 \le t_2 \cdots \le t_{2k} \le T} \sum_{i=1}^{k} |X_{t_{2i}} - X_{t_{2i-1}}|,$$

where the times $t_1 \dots t_k, t_{k+1} \dots t_{2k}$ vary within the grid of the n available returns. This apparently strange formulation in fact allows for nonadjacent returns ($X_{t_{2i}} - X_{t_{2i-1}}$ is adjacent to $X_{t_{2(i-1)}} - X_{t_{2(i-1)-1}}$ only if $t_{2(i-1)} = t_{2i-1}$). Only k terms are considered above, but in order to establish where the maximum is attained, all the n absolute returns have to be checked. IV is asymptotically correctly estimated by

$$\frac{GR_{X,n}^2}{E[GR_{B,n}^2]},$$

where the scaling factor $E[GR_{B,n}^2]$ is the expectation of the generalized range of a BM. No CLT is given, so we do not know the convergence speed, however a performance comparison on simulated data is done with RBV, BPV and an averaged BPV, and the GR estimator seems to be able to compete.

Rate	Estimator	Efficiency	Estimator	Efficiency	Estimator	Efficiency
$\sqrt{n}$	TRV	2.00	$\text{QRV}_{k=5}$	2.19	$\text{QRV}_{\text{overl.}}$	2.32
	$\text{QRV}_{k=4}$	2.42	TBV	2.61	MedRV	2.96
	$\text{RV}^{\star}$	2.00	$\text{ROWV}^{\star}$	2.00	$\text{BPV}^{\star}$	2.61
$\sqrt{n/m}$	$\text{RBV}^{\star}_{m\to\infty}$	0.36	$\text{RBV}^{\star}_{m\to 1}$	2.61		

Table 2.1: Asymptotic efficiency comparison of the presented estimators. The AVar of each estimator equals ϑ IQ. For GR no CLT is available. $\star$ means that a CLT for the considered method is available only in the absence of jumps. For $\text{QRV}_{\text{overl}}$ we considered $k = 4$ and for ROWV w_{HR} with $\beta = 1$.

2.4 The Comparative Monte Carlo Experiment

Table 2.2 compares the indicated estimators, which are implemented on the three simulated models. For each model, 1000 paths are generated and the mean relative percentage estimation error $100(\hat{IV} - IV)/IV$ is reported, when either 5 min or 1 sec observations are used within a one day time horizon. In parentheses the empirical standard error of each such bias is reported. As mentioned, RV is not consistent for IV in the presence of jumps, but it is important to understand how wrong the estimate is when RV is not properly used.

For TRV the chosen threshold is $ch^{0.99}$ and c is calibrated for all three models so that it minimizes the absolute value of the estimation error in Model 1 (if $h = 5$ min then $c = 0.0012$, if $h = 1$ sec $c = 0.0034$; these c values correspond to about 1/8 and 1/3 of the average spot volatility 0.0098. On the contrary, for TRV°c is optimally chosen in each model (in Model 2: if $h = 5$ min $c = 0.12$, if $h = 1$ sec, $c = 0.4$; in Model 3: if $h = 5$ min, $c = 9.00 \times 10^{-9}$; if $h = 1$ sec, $c = 2.34 \times 10^{-6}$). For TBV the chosen threshold is $c^2\hat{\sigma}^2_{t_i}h$, as described after (2.6). Also in this case c is chosen optimally only for Model 1 (if $h = 5$ min then $c = 4.1$, if $h = 1$ sec $c = 12.6$), while for TBV°it is optimally chosen in each model (in Model 2: if $h = 5$ min, $c = 10$; if $h = 1$ sec, $c = 7$; in Model 3: if $h = 5$ min, $c = 0.276$; if $h = 1$ sec, $c = 2.89 \times 10^{+5}$). We recall that BTV is the special case of TRV with $r_t(h) = 9\ BPV_n(T)\ h^{0.99}/T$ and here $T = 1$.

For QRV°the blocked version is used, the number m of returns per block is optimally chosen in each model (in Model 1: if $h = 5$ min, $m = 84$; if $h = 1$

	5 min			*1 sec*		
	HT1FJ	HT2F	CGMY	HT1FJ	HT2F	CGMY
RV	28.543	0.015	3.03e+7	28.160	-0.006	3.04e+7
	(51.495)	(22.438)	(8.20e+7)	(47.467)	(1.319)	(8.21e+7)
BPV	7.840	-2.707	6.34e+5	0.629	-0.008	248.449
	(21.081)	(25.190)	(5.37e+6)	(1.243)	(1.546)	(5.64e+3)
TRV°	0.003	-0.004	0.462	-3.14e-4	-0.006	0.125
	(15.843)	(22.419)	(49.614)	(0.931)	(1.319)	(50.122)
TRV	0.003	-9.892	2.68e+6	-3.14e-4	-4.770	5.65e+4
	(15.843)	(24.962)	(3.00e+6)	(0.931)	(15.114)	(4.50e+4)
BTV	-0.836	-19.795	5.77e+4	-2.013	-19.193	-99.730
	(16.099)	(21.731)	(5.75e+5)	(0.939)	(10.593)	(4.496)
TBV°	0.009	-2.707	4.467	-2.00e-5	-0.008	8.266
	(17.748)	(25.190)	(816.980)	(1.034)	(1.546)	(2.60e+3)
TBV	0.009	-7.378	5.24e+3	-2.00e-5	-0.008	-99.949
	(17.748)	(24.637)	(3.82e+4)	(1.034)	(1.546)	(0.002)
QRV°	-4.556	-27.852	1.64e+4	0.014	-1.267	10.109
	(16.119)	(19.126)	(1.02e+5)	(1.003)	(1.441)	(2.04e+3)
MinRV	3.002	-2.334	3.00e+5	-0.003	3.462e-4	-9.604
	(21.873)	(30.439)	(2.40e+6)	(1.246)	(1.856)	(2.19e+3)
MedRV	3.304	-2.968	5.00e+5	0.021	0.003	96.350
	(19.400)	(26.183)	(3.67e+6)	(1.118)	(1.637)	(3.92e+3)
ROWV°	0.387	1.301	-48.234	-0.045	-0.384	-99.967
	(16.922)	(22.726)	(566.590)	(0.940)	(1.393)	(2.08e-8)
RBV*	-0.523	-2.865	5.95e+5	0.466	-0.170	247.883
	(32.755)	(25.149)	(3.15e+6)	(1.241)	(1.544)	(5.63e+3)
GR°	9.381	0.729	1.14e+6	0.413	-0.033	3.80e+3
	(19.389)	(68.872)	(2.79e+6)	(1.125)	(15.661)	(9.48e+3)
GR	9.934	-18.507	1.24e+6	0.561	-16.783	4.99e+3
	(20.421)	(18.497)	(3.04e+6)	(1.266)	(10.265)	(1.24e+4)

Table 2.2: Mean relative percentage estimation error $100(\hat{IV} - IV)/IV$ of the estimators in the three simulated models. Model 1 contains FA jumps, Model 2 contains no jumps, Model 3 has IA jumps.

sec, $m = 2520$; in Model 2: if $h = 5$ min, $m = 84$; if $h = 1$ sec, $m = 200$; in Model 3: if $h = 5$ min, $m = 84$; if $h = 1$ sec, $m = 8$), $\lambda = [0.8, 0.85, 0.9, 0.95]$ and the optimal α is calculated for each model numerically using the algorithm described in (Christensen et al. [2010]). For instance in Model 1, if $h = 5$ min, $\alpha = [0.1821, 0.1353, 0.2141, 0.4685]$. ROWV°is computed using $w_{HR,99.9\%}$. The factor c_ω equals 1.013. For the first step estimation of spot variance the MAD (Median Absolute Deviation) method is used, in place of LTS, because it is much more simply implemented, while Boudt confirmed (in a private conversation) that this change only marginally influences the results. The local window length parameter λ is optimally calibrated in each model (in Model 1: if $h = 5$ min, $\lambda = 1/4$; if $h = 1$ sec, $\lambda = 1$; in Model 2: if $h = 5$ min, $\lambda = 1/42$; if $h = 1$ sec, $\lambda = 1/126$; in Model 3: if $h = 5$ min, $\lambda = 1/7$; if $h = 1$ sec, $\lambda = 1$). For RBV* the number m of increments per window is chosen in each model so as to strike a balance between bias and efficiency, since, in this case, the efficiency decreases as m increases (in Model 1: if $h = 5$ min, $m = 12$; if $h = 1$ sec, $m = 1$; in Model 2: if $h = 5$ min, $m = 1$; if $h = 1$ sec, $m = 15$; in Model 3: if $h = 5$ min, $m = 2$; if $h = 1$ sec, $m = 1$). The implementation of GR has been done using a properly designed algorithm which is different from the one described in Dobrev [2007]. However Dobrev kindly provided the results obtained with his algorithm on our simulated data, and these coincide with ours. For GR°the number k of increments included in the estimation procedure is optimally chosen in each model (in Model 1: if $h = 5$ min, $k = 44$; if $h = 1$ sec, $k = 12600$; in Model 2: if $h = 5$ min, $k = 1$; if $h = 1$ sec, $k = 14$; in Model 3: if $h = 5$ min, $k = 83$; if $h = 1$ sec, $k = 12600$), while for GR, k is set for all the three models to values giving acceptable results in terms of the bias and efficiency of the estimator (if $h = 5$ min, $k = 30$ and if $h = 1$ sec, $k = 6000$).

The optimal choices we considered are only possible in a simulation framework and not with empirical data, but allow us to evaluate the potential precision the estimators could reach. We report the estimation errors also when $h = 1$ sec in order to assess the performance of the estimators: this is not feasible with empirical data, due to the high relevance of the microstructure noise in the 1 sec returns, but it gives useful indications in ideal conditions.

2.5 Noisy data

Estimation of IV is done with $h \to 0$. In practice, the smaller the step, the better the estimate. Having access to UHFD let us to consider extremely small h. However we have to consider the bias induced by *microstructure noise*. A strand

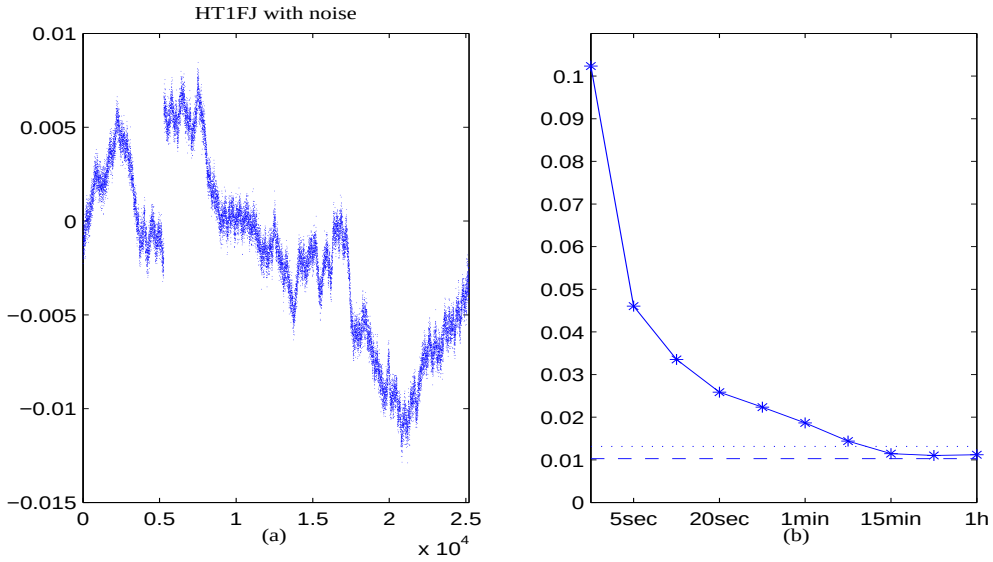

Figure 2.4: (a) one day simulated path of HT1FJ model with iid Gaussian noise $\mathcal{N}(0, (4.5 * 10^{-4})^2)$, the variance being estimated from the J&J series, and n=25'200; (b) relative SP of $\sqrt{RV_n}$.

of the literature still keeps SM price models (such as, e.g., HT1FJ), motivated by the hypothesis of the absence of arbitrage opportunities in efficient markets. However the behavioural discrepancy $\epsilon = X - Y$ between the observed path X and the model path Y should also accounted for. Y is called the *efficient* log-price, ϵ is called *microstructure noise* process and represents the discrepancy between the observed and the efficient price process. The main causes of this discrepancy are, among others, price discreteness, infrequent trading, and negative first order autocorrelation induced in the returns series by the bid-ask bounce. We are interested in the IV of Y, Y however is unobservable. Simulated noisy prices (figure 2.4 (a)) seem to display a local strip (which is more evident for higher values of the noise variance), similar to the bid–ask spread, indicating an interval where the efficient price is placed.

Various approaches have been developed in the literature to model the ϵ behaviour. The most common framework considers *additive iid noise* (also called *additive white noise*). Even if not completely realistic (as pointed out in Hansen and Lunde [2006]), such an assumption is accepted at moderate observation frequencies and in general as a first approximation to the real phenomenon. For a comprehensive discussion of the noise models and the effect of noise on the inference for the underlying process see also Li and Mykland [2007].

X is not an SM anymore in that as $h \to 0$, while the increment $|\Delta_i Y|$ of any SM tends to zero in probability, we have for small $\varepsilon > 0$ that $P\{|\Delta_i X| > \varepsilon\}$ is close to one, due to the fact that $P\{|\Delta_i \epsilon| > \tilde{\varepsilon}\}$ does, for small $\tilde{\varepsilon} > 0$, e.g., under the additive iid noise assumption, for the commonly used model distributions (Gaussian or uniform) for ϵ_{t_i}. On the other hand, when we observe prices at LF they seem much more compatible with an SM DGP. figure 2.1 shows that observing the SPX prices only once per day over four years, the true price path and the simulated paths have similar characteristics and produce similar returns. So microstructure effects have a substantial impact on data only at UHF.

When including iid centred noises, denoted $\sigma_\epsilon^2 = Var[\epsilon_{t_i}]$, because

$$\sum_i (\Delta_i Y)^2 \sim \sum_i (\Delta_i X)^2 + \sum_i (\Delta_i U)^2 \sim [X, X]_T + 2n\sigma_\epsilon^2 \tag{2.8}$$

explodes to infinity as $h \to 0$, RV_n is not adequate anymore to estimate QV (and thus to estimate IV when X is continuous). This is reflected in figure 2.4 (b), showing the SP of $\sqrt{RV_n}$ of simulated log-prices generated by the HT1FJ model plus an independent Gaussian iid noise process.

Specific devices have been designed to robustify estimators of IV in the presence of both jumps and noise. The most common approach is *sparse sampling*, which avoids sampling too frequently. Criteria for optimal frequency selection are given in Bandi and Russell [2008] in a BSM framework and are under study in Mancini [2011, working paper] in the presence of jumps. The optimal frequency depends on the specific asset we need to analyse, however five minutes is commonly considered as a watershed: for data observed at lower frequencies, noise is usually considered negligible.

2.6 Conclusions

We reviewed all the methods in the literature we are aware of about non-parametric estimation of IV within the framework of general Itô SM models evolving in continuous time but which are observable only discretely. The inconsistencies arising in the application of the estimation methods to simulated data show how difficult the asset prices modelling issue is. Trying to account for realistic features (e.g., microstructure noise) immediately gives rise to increasing technical difficulties in reaching the theoretical results needed for the practical application of the methods. Deepening our comparative study of the different estimators performances under different assumptions about the price DGP would be very interesting.

This being said, the stream of literature on integrated variance estimation assumes a continuous time data generating process with various sources of movements and of noise and derives theoretical properties for the estimators based on fairly abstract theoretical conditions. While the results are elegant, the main issue remains as per the empirical relevance of the large host of suggestions when compared with observed intra–daily data. To this end, the software developed in Chapter 1 will allow us to compute all existing measures while controlling for the main parameters of interest (e.g. frequency of observation). Whether jumps are isolated or autocorrelated is a separate issue which could cast light on the very question of the hypotheses which are assumed for the data generating process, possibly suggesting other assumptions more in line with the observed behavior of financial time series.

Chapter 3

GAS Models and a New Test for Parameter Instability

3.1 Introduction

To summarize what we have done so far, Chapter 1 provided us with the tools to manage and manipulate the tick–by–tick data from the exchanges, while Chapter 2 gave us the opportunity to investigate the properties of all non parametric estimators of integrated variance. Having devoted the main focus of previous contributions to the measurement of volatility, it is now appropriate to turn our attention to volatility modelling issues with the goal of inserting an explicit component characterized as "jumps". We thus start from considering the dynamic behavior of the Realized Variance (section 2.2.3), as a measure of the overall risk associated to a certain asset, we employ a time series parametric approach which let us to include time variation in a selection of model parameters. Given this framework, different possible model specifications and testing procedures are available depending on how the parameters are assumed to evolve over time. We focus on the Generalized Autoregressive Score (GAS) models developed by Creal et al. [2012]. We propose a new test for parameter instability (see Calvori et al. [2012]) which generalizes the ARCH Lagrange Multiplier (LM) test of Engle [1982] to settings beyond the time–varying variance of a Gaussian conditional distribution and is useful for detecting if some component of the model (expressed

by a parameter) is fixed or really time–varying.

In this chapter we first introduce the GAS basic specification as in Creal et al. [2012], and then we give details about our new test, denoted by $LM_{GAS(p,0)}$, which is a Lagrange Multiplier test for the null hypothesis of constant parameters against the alternative of GAS$(p, 0)$ effects. Furthermore we introduce two familiar tests built against different type of parameter instability under the alternative: the one proposed by Andrews [1993] and the other developed by Müller and Petalas [2010]. Finally an extensive Monte Carlo experiment is carried out to compare small sample properties of the three tests.

3.2 Generalized Autoregressive Score (GAS) models

We use the following notation. $y_t \in \mathbb{R}^m, (t = 1, \ldots, n)$ is the (observable) variable of interest; n denotes the sample size, $\beta_t \in B \subset \mathbb{R}^k$ is a vector of time-varying parameters driving the dynamic of y_t; $\theta \in \Theta \subset \mathbb{R}^\ell$ is a vector of static parameters; B and Θ denote the parameter spaces of the time-varying and static parameter vectors, respectively.

3.2.1 Basic specification

Cox [1981] classifies time-varying parameter models into two classes: observation-driven and parameter-driven specifications. GAS models are a class of observation driven time series models. In an observation driven framework (see section 3.4.2 for the definition of the parameter driven framework), the time-varying parameter β_t is a *predictable* stochastic process, since it is driven by deterministic functions of lagged dependent variables and contemporary or lagged exogenous variables. One of the main advantages of this framework is that the likelihood can be easily evaluated through the prediction error decomposition leading to simple estimation and inference procedures.

The main challenge in the observation driven framework is to formulate properly the function of the observations that drives the time-varying parameter. The GAS approach uses the (scaled) conditional density score as a driver for β_t and

can thus be used in every model with a parametric conditional density for y_t. Intuitively, the score defines a steepest ascent direction for improving the model's local fit in terms of the likelihood at time t given the current value of the parameter β_t. The GAS framework encompasses as special cases the normal GARCH model of Engle [1982] and Bollerslev [1986], the ACD and ACI models of Engle and Russell [1998] and Russell [2001], the MEM model of Engle and Gallo [2006] and Cipollini et al. [2012], and some of the models for Poisson counts in Davis et al. [2003], among others.

We assume y_t be generated by the conditional density $p(y_t|\beta_t, \mathcal{F}_t;\ \theta)$, where $\mathcal{F}_t$ represent the information available at time t (lagged dependent variables, lagged values of the time-varying parameter, lagged and contemporary exogenous variables). In the GAS(p, q) framework, the time-varying parameter evolves according to

$$\beta_{t+1} = \omega + \sum_{i=1}^{p} A_i s_{t-i+1} + \sum_{j=1}^{q} C_j \beta_{t-j+1}, \tag{3.1}$$

where the vector ω and the matrices $A_i, (i = 1, \ldots, p)$ and $C_j, (j = 1, \ldots, q)$ are static parameters,

$$s_t = S_t \cdot \nabla_t, \qquad \nabla_t = \frac{\partial \ln p(y_t|\beta_t, \mathcal{F}_t;\ \theta)}{\partial \beta_t}, \qquad S_t = S(t, \beta_t, \mathcal{F}_t; \theta), \tag{3.2}$$

and $\theta = (\gamma', \omega', a', c')'$ where γ include all constant parameters besides ω, $a = \text{vec}(A_1, \ldots, A_p)$ and $c = \text{vec}(C_1, \ldots, C_q)$. One possible choice for the scaling function S_t is

$$S_t = \mathcal{I}_{t|t-1}^{-1} = E[\nabla_t \nabla_t' | \mathcal{F}_{t-1}],$$

where $\mathcal{I}_{t|t-1}$ denotes the conditional information matrix of the tth observation (see Creal et al. [2012] for more details).

3.2.2 Two simple examples

Example 1: Gaussian time-varying mean. Assuming

$$p(y_t|\beta_t, \mathcal{F}_t;\ \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_t - \mu_t)^2}{2\sigma^2}\right], \tag{3.3}$$

with $\mu_t = \beta_t$ denoting the time-varying mean and σ the constant standard deviation, we can easily derive the GAS$(1,1)$ updating equation (using $\mathcal{I}_{t|t-1}^{-1}$ to scale the score):

$$\beta_{t+1} = \omega + A_1(y_t - \beta_t) + C_1\beta_t, \tag{3.4}$$

given that

$$\nabla_t = \frac{y_t - \beta_t}{\sigma^2}, \qquad S_t^{-1} = \mathcal{I}_{t|t-1} = \frac{1}{\sigma^2}. \tag{3.5}$$

Example 2: GARCH models. Assuming

$$p(y_t|\beta_t, \mathcal{F}_t;\ \theta) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{y_t^2}{2\sigma_t^2}\right), \tag{3.6}$$

where $\beta_t = \sigma_t^2$ is the time-varying variance, the corresponding GAS$(1,1)$ model with $S_t = \mathcal{I}_{t|t-1}^{-1}$

$$\beta_{t+1} = \omega + A_1(y_t^2 - \beta_t) + C_1\beta_t, \tag{3.7}$$

since

$$\nabla_t = \frac{y_t^2 - \beta_t}{2\beta_t^2}, \qquad S_t^{-1} = \mathcal{I}_{t|t-1} = \frac{1}{2\beta_t^2}. \tag{3.8}$$

This is equivalent to the standard GARCH$(1,1)$ dynamic of σ_t^2 given by

$$\beta_{t+1} = \lambda_0 + \lambda_1 y_t^2 + \zeta_1\beta_t, \tag{3.9}$$

(see Bollerslev [1986] for details). For further examples refer to Creal et al. [2012].

3.3 An LM test for GAS$(p,0)$ effects

The GAS model described in the previous section is sufficiently general to accommodate the dynamics in most parametric, fully specified models. Estimating a GAS specification may not always be sensible, particularly if there is little dynamic in the parameter over time. Therefore, it may be useful to first perform a mis-specification test based on the null hypothesis of constant parameters. In a GAS$(p,0)$ specification we can easily formalize the null hypothesis of parameter stability by setting all the elements of the vector a equal to zero, so that the time-varying parameters β_t will be constant and equal to ω for every t. Under

H_0, the model is usually easy to estimate suggesting an LM test.

We introduce the following notation. Let

$$\mathcal{L}_t = \ln p(y_t|\beta_t, \mathcal{F}_t; \theta) \tag{3.10}$$

be the likelihood at time t, and let $\mathcal{L}_{\gamma,t}$, $\mathcal{L}_{\omega,t}$, and $\mathcal{L}_{a,t}$ denote its derivatives with respect to γ, ω, and a, respectively. Finally, define

$$\bar{s}_{p,t} = \text{vec}(s_t, \ldots, s_{t-p+1}).$$

The Lagrange Multiplier test for $H_0 : a = 0$ is given by

$$LM_{GAS(p,0)} = n \cdot \left[\frac{1}{n}\sum_{t=1}^{n}\begin{pmatrix}\hat{\mathcal{L}}_{\gamma,t}\\ \hat{\mathcal{L}}_{\omega,t}\\ \hat{\mathcal{L}}_{\omega,t}\otimes\hat{\bar{s}}_{p,t-1}\end{pmatrix}\right]' \hat{D}_n^{-1} \left[\frac{1}{n}\sum_{t=1}^{n}\begin{pmatrix}\hat{\mathcal{L}}_{\gamma,t}\\ \hat{\mathcal{L}}_{\omega,t}\\ \hat{\mathcal{L}}_{\omega,t}\otimes\hat{\bar{s}}_{p,t-1}\end{pmatrix}\right], \tag{3.11}$$

with

$$\hat{D}_n = \frac{1}{n}\sum_{t=1}^{n}\begin{pmatrix}\hat{\mathcal{L}}_{\gamma,t}\\ \hat{\mathcal{L}}_{\omega,t}\\ \hat{\mathcal{L}}_{\omega,t}\otimes\hat{\bar{s}}_{p,t-1}\end{pmatrix}\begin{pmatrix}\hat{\mathcal{L}}_{\gamma,t}\\ \hat{\mathcal{L}}_{\omega,t}\\ \hat{\mathcal{L}}_{\omega,t}\otimes\hat{\bar{s}}_{p,t-1}\end{pmatrix}', \tag{3.12}$$

where the hat means that a quantity is evaluated at the maximum likelihood estimates under H_0, $\hat{D}_n$ is a matrix that converges to the expected value of the outer product of the gradient, and the equality $\hat{\mathcal{L}}_{a,t} = \hat{\mathcal{L}}_{\omega,t} \otimes \hat{\bar{s}}_{p,t-1}$ follows from the fact that under the null the derivative of the conditional log-density of y_t with respect to β_t is the same as that with respect to ω:

$$\hat{\nabla}_t = \hat{\mathcal{L}}_{\omega,t}. \tag{3.13}$$

Following Davidson and MacKinnon [1990], we can also obtain the $LM_{GAS(p,0)}$ test statistic as the explained sum of squares of the auxiliary OLS regression

$$1 = (\hat{\mathcal{L}}'_{\gamma,t}, \hat{\mathcal{L}}'_{\omega,t}, \hat{\mathcal{L}}'_{\omega,t} \otimes \hat{\bar{s}}'_{p,t-1})\alpha_{\text{LM}} + \text{residual}, \tag{3.14}$$

where α_{LM} is a vector of auxiliary regression parameters. The regression inter-

pretation of the GAS test makes it easy to compute in standard packages. The only quantities needed are the numerical scores of the conditional density at each time t. These are easily obtained either analytically or otherwise via standard numerical differentiation. Under standard regularity conditions, the GAS LM test converges to a χ^2 random variable with $\dim(a)$ degrees of freedom.

The $LM_{GAS(p,0)}$ test has a very intuitive interpretation. By looking at the test statistic in equation (3.11) or (3.14), we see that the key term is $\hat{\mathcal{L}}_{\omega,t} \otimes \hat{\tilde{s}}_{p,t-1}$ (the sample averages of both $\hat{\mathcal{L}}_{\gamma,t}$ and $\hat{\mathcal{L}}_{\omega,t}$ are zero by construction). Recalling the equality (3.13), we can state that the elements of this vector are $\text{vec}(\hat{S}_{t-i}\hat{\mathcal{L}}_{\omega,t-i}\hat{\mathcal{L}}'_{\omega,t})$ for $i = 1, \ldots, p$. The easiest case to inspect is that of unit scaling, i.e., $S_t = \mathbb{1}_k$. The LM test against the GAS$(p, 0)$ alternative then checks whether there is any autocorrelation in the score of the likelihood with respect to the intercept ω of the GAS transition equation. Differently said, it checks whether there is any (higher order) autocorrelation in the score of the likelihood of the *static* model $\beta_t \equiv \omega$ with respect to β_t.

Even though the new $LM_{GAS(p,0)}$ test has been derived with the GAS alternative in mind, we expect it to have also power against different forms of parameter dynamics. The same holds for the two tests, one against structural breaks and the other against parameter driven parameter dynamics, which are presented in the next section. Which of these three approaches performs best under different conditions is the question to which we turn in our Monte Carlo survey.

3.4 Two Alternative Testing Frameworks

In this section we present two different testing methodologies for parameter instability, besides the new procedure proposed in this chapter. Each of these frameworks has been designed with a particular alternative in mind and is employed in our Monte Carlo comparative survey to evaluate how the $LM_{GAS(p,0)}$ test works in finite sample.

3.4.1 The Andrews test

Andrews [1993] proposes a framework for testing for parameter instability in general nonlinear parametric models. His tests are designed against alternatives with a one-time structural change in (a subset of) the parameters. The tests are based on partial-sample GMM (PS-GMM) estimators and can be of the supremum Wald, Lagrange multiplier (LM), or likelihood ratio (LR) type. Versions of these tests that use weighted averages rather than the supremum of the tests over all possible break points are proposed by Ploberger et al. [1989] and Andrews and Ploberger [1994].

Let $\pi \in (0,1)$ and let $\lfloor \pi n \rfloor + 1$ denote the breakpoint of the parameter β_t, where $\lfloor x \rfloor$ denotes the integer part of $x \in \mathbb{R}$. The null and alternative hypothesis for the Andrews test are given by

$$\mathrm{H}_0 : \beta_t = \bar{\beta}_0 \quad \forall t \geq 1 \text{ and some } \bar{\beta}_0 \in B \subset \mathbb{R}^k, \tag{3.15}$$

$$\mathrm{H}_1 : \bigcup_{\pi \in \Pi} \mathrm{H}_{1,n}(\pi) \text{ for some } \Pi \subset (0,1), \tag{3.16}$$

with

$$\mathrm{H}_{1,n}(\pi) : \beta_t = \begin{cases} \bar{\beta}_1(\pi) & \text{for } t = 1, \ldots, \lfloor \pi n \rfloor \\ \bar{\beta}_2(\pi) & \text{for } t = \lfloor \pi n \rfloor + 1, \ldots, n, \end{cases} \tag{3.17}$$

for constants $\bar{\beta}_1(\pi), \bar{\beta}_2(\pi) \in B$. Though the test is designed for a single structural break at unknown date, it is by now well-known that the test also has good power properties against a range of other, more general alternatives; see for example Hansen [2001].

3.4.2 The Müller-Petalas test

In parameter driven time-varying parameter models, the parameter β_t is a stochastic process that is subject to its own source of error. Important examples of this class of models are the structural time series models of Harvey [1989], the stochastic volatility models as reviewed in Shephard [2005], the stochastic conditional duration model of Bauwens and Veredas [2004], and the stochastic copula model

of Hafner and Manner [2012]. This additional randomness in β_t on top of the randomness in y_t conditional on β_t makes these models typically hard to estimate. Except for some specific cases, such as linear-Gaussian state space models and discrete-state hidden Markov models, see Durbin and Koopman [2012] and Hamilton [1989] respectively, the likelihood function is typically not available in closed form, and its evaluation requires approximation and/or simulation methods.

An elegant and general set-up to test for parameter instability of the above form is provided by Müller and Petalas [2010]. Their approach encompasses non-linear and non-Gaussian models with moderately time-varying parameters. The restriction to moderate time-variation follows from a local asymptotic argument they use: in a local asymptotic framework they show that the inference problem of unstable parameters can be addressed by considering a linear Gaussian state space model where the observations are replaced by the likelihood scores of the static model.

The (potentially) time-varying parameter is assumed to evolve over time according to the equation

$$\beta_t = \beta^* + \nu_t, \qquad \text{for all } t = 1, \ldots, n, \tag{3.18}$$

with ν_t being a stochastic process representing the deviation at time t from the parameter baseline value β^*. Given equation (3.18), the hypothesis of interest for the Müller-Petalas test are

$$\begin{aligned} H_0 &: \nu_t = 0 \quad \text{for all } t = 1, \ldots, n, \\ H_1 &: \nu_t \neq 0 \quad \text{for some } t. \end{aligned} \tag{3.19}$$

Such an approach turns out not to be only asymptotically optimal against the alternative of (local) parameter driven dynamics in β_t, but against a much wider range of (local) alternative dynamics.

3.5 The Monte Carlo Experiment

In this section we describe the Monte Carlo experiment carried out to investigate the small sample behaviour of the considered tests. Our aim is to evaluate the tests empirical power of the tests under different specifications of parameter instability. We present the details of each simulated model, as well as the results of our comparative survey.

3.5.1 The basic set-up

Depending on the different approach used to simulate the time-varying parameter path, we classify the data generating processes (DGPs) employed in our Monte Carlo survey in three main categories: regime switching models (with two levels and equally spaced switching points), models with structural breaks (breaks size and changing points are random), and state-space models. For each specification of the DGP, we perform $N = 1000$ replications of time series of $n = 2000$ observations on which we implement the $LM_{GAS(1,0)}$, the $LM_{GAS(5,0)}$, the Andrews [1993], and the Müller and Petalas [2010] (MP) tests.

Let $n_b \in \mathbb{N}$ denote the fixed number of switch/break points occurring in a simulated series, and define

$$N_b = \begin{cases} \frac{n_b}{2} - 1 & \text{if } n_b \text{ is even,} \\ \lfloor \frac{n_b}{2} \rfloor & \text{if } n_b \text{ is odd.} \end{cases} \tag{3.20}$$

To simulate *regular* regime-switching models, we assume β_t to evolve over time according to

$$\beta_t = \begin{cases} \Delta & \text{for } \lfloor \frac{2j+1}{n_b+1} n \rfloor < t \leq \lfloor \frac{2(j+1)}{n_b+1} n \rfloor \text{ for every } j = 0, \ldots, N_b, \\ 0 & \text{otherwise,} \end{cases} \tag{3.21}$$

with Δ denoting the difference (in absolute value) between the two regimes.

To simulate models with structural breaks, we define $\Xi_{n_b} = \{\lfloor \pi_i n \rfloor + 1 : i = 1, \ldots, n_b\}$ as the set containing the simulated breakpoints in the time-varying parameter path, where π_i is a random variable with uniform distribution on the

interval $(0, 1)$. The equation describing the time dynamics of β_t in the structural breaks DGP is

$$\beta_t = \sum_{\tau=1}^{t} I_{\{\tau \in \Xi_{n_b}\}} v_\tau, \qquad \text{for } t = 1, \ldots, n, \tag{3.22}$$

with

$$I_{\{\tau \in \Xi_{n_b}\}} = \begin{cases} 1 & \text{if } \tau \in \Xi_{n_b}, \\ 0 & \text{otherwise,} \end{cases}$$

and v_τ being a Gaussian random variable (representing the size of the structural breaks) with zero mean and variance (constant over time) σ_v^2.

Regarding the state-space framework, we assume that the time-varying parameter follows the usual (zero-mean) AR(1) updating state equation:

$$\beta_t = \phi \beta_{t-1} + \eta_t, \tag{3.23}$$

where $\eta_t \sim$ iid $\mathcal{N}(0, \sigma_\eta^2)$.

For each one of the three β_t dynamics, we specify three different models (see table 3.1 for details) to simulate the time series of the observations y_t. Figure 3.1

Time-varying mean	$y_t = \beta_t + \epsilon_t,$	$\epsilon_t \sim$ iid $\mathcal{N}(0, \sigma_\epsilon^2)$
Time-varying variance	$y_t = e^{\frac{1}{2}\beta_t} \epsilon_t,$	$\epsilon_t \sim$ iid $\mathcal{N}(0, \sigma_\epsilon^2)$
Time-varying correlation	$y_t \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_t \\ \rho_t & 1 \end{bmatrix} \right),$	$\rho_t = \tanh \beta_t$

Table 3.1: Models employed to simulate the observations y_t. σ_ϵ^2 denote the constant variance of the error term ϵ_t. The models are parametrized in the way that β_t can vary over the whole real line.

shows simulated series for the time-varying mean DGP in the three dynamics for β_t.

For the regime-switching and the structural breaks approach we investigate the evolution of the empirical power functions as the number of switch/break points increases. In practice, for each fixed number of switch/break points, we evaluate the empirical power function by varying either Δ or σ_v^2 (see equations (3.21) and (3.22) respectively) depending on the assumptions about the DGP. For the state-space framework, the power function is evaluated in two different

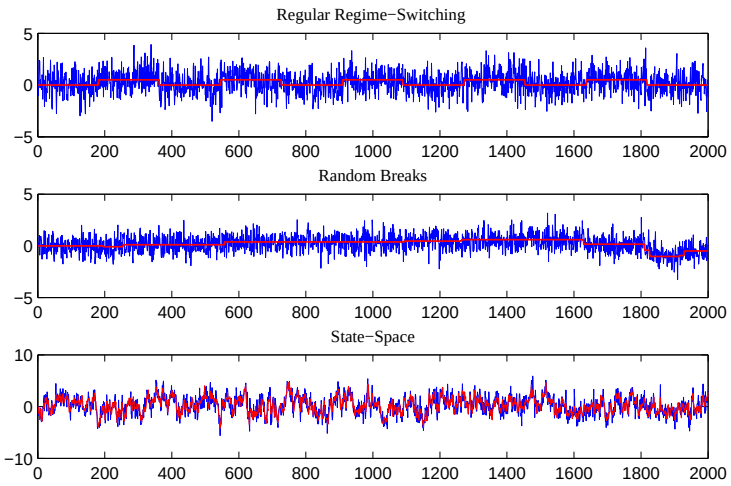

Figure 3.1: Three simulated series of the time-varying mean model. The blue and the red line represents the y_t and the β_t simulated paths respectively.

settings: varying σ_η^2 with constant $\phi = 0.9$, varying both ϕ and $\sigma_\eta^2 = 0.15(1-\phi^2)$, in the way that the unconditional variance of β_t remains equal to 0.15 .

3.5.2 Results

In this section, we present the results of the Monte Carlo experiment. For every DGP, we implement 5% significance level test. The asymptotic distribution of the GAS test statistic is standard, while for the critical values of the other two tests we refer to the corresponding literature.

Let us provide the main findings that can be obtained by observing the results of our simulation analysis. We start from the *regular* regime–switching dynamic for β_t (see figures 3.4–3.6). In the case of one switching point in the simulated parameter path, the Andrews test can be considered the benchmark in terms of power function, since the author proofs its (local asymptotic) optimality for testing against the alternative of a unique structural break with unknown change point.

A further remark about the Andrews test has to be done. Theoretically, to compute the test statistic, we should consider all the potential changing points

in our sample, calculate the statistics considered (Wald, LM or LR-like) for each point, and take the supremum of the built sequence. However, to reduce the computational burden, a grid on the possible changing points needs to be defined in practice. The test performance depends rather considerably on how fine this grid is taken to be, since the statistic computed without taking into account all possible break points, is always less than or equal to the actual supremum, so that the test tends to accept the null hypothesis more than it should do, reducing the power. In our simulation set-up, given the critical values (obtained in Andrews [1993] by Monte Carlo simulations) for a certain level of significance, we calibrate the grid in order to make the test observed size as close as possible to the test's nominal size for that specific critical value.

Focusing our attention on the top left graph of figure 3.4, which shows the empirical power functions for the Gaussian DGP with time–varying mean with only one switch in the middle of the sample, we notice that the MP and the Andrews tests have overlapping power functions, which rapidly increase toward one. On the other side, both the $LM_{GAS(1,0)}$ and the $LM_{GAS(5,0)}$ tests, despite showing to have power against this type of alternative, have more difficulty in capturing the parameter instability for small values of Δ. This can be explained by recalling that testing for GAS$(p, 0)$ effects can be viewed as a joint test for zero autocovariances until lag p of the conditional density score with respect to ω. In general, since the score at time t is evaluated under the null, it depends on the observations through y_t only (no lagged observations are involved), hence the *dependence structure* of the generated time series influences the *dependence structure* of the score. In the case we are considering (Gaussian model with time-varying mean) the score with respect to ω at time t is a linear function of y_t, so that we can directly focus on the observations autocorrelation to explain the $LM_{GAS(p,0)}$ test mechanism. If Δ is small with respect to the standard deviation of the error term, the *white noise* component prevails, while if Δ is large enough, the probability of drawing observations below the overall mean before the changing point and above the overall mean after the changing point is higher, generating highly autocorrelated observations and consequently a highly autocorrelated score. See figure 3.2 for a graphical explanation of this concept.

Skimming trough the graphs in figure 3.4, we can investigate the tests be-

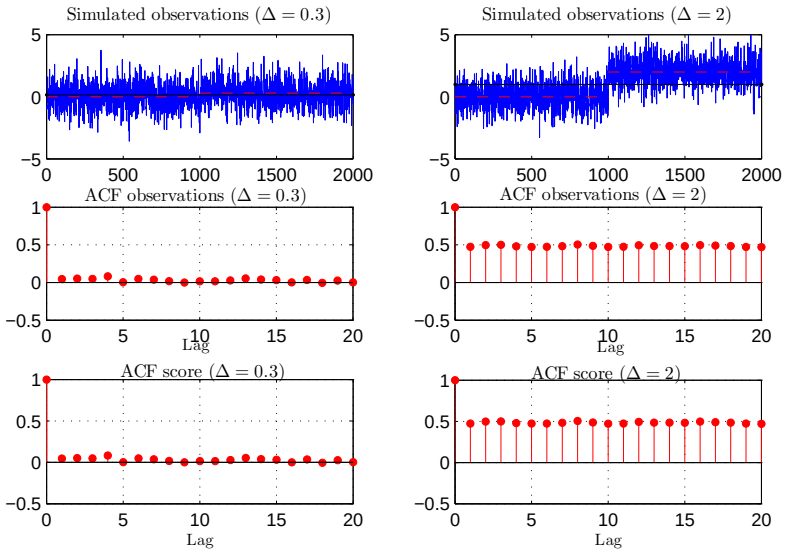

Figure 3.2: Simulated series for regime-switching (time-varying mean) with one changing point (the blue, the dashed red and the black line represents y_t, β_t, and the overall mean respectively), and sample autocorrelation functions (ACF) of the observations and the score series, for different values of Δ.

haviour when the number of change points in the simulated DGP increases. The power functions of the Andrews and MP tests have a flatter shape, while both the $LM_{GAS(1,0)}$ and the $LM_{GAS(5,0)}$ tests highlight a remarkable robustness, with power functions that remain stable as the number of switching points increase. The decreasing performances of the Andrews test are in some sense implicit in the structure of the test statistic, developed against the alternative of a unique break. In fact, the distance between the two regimes (computed by using the PS-GMM procedure on the observations before and after a potential switching point respectively) is underestimated if the number of these changing points is greater then one, given that at least one of the parts into which the sample is split includes at least one break and this gives underestimation of the higher level or overestimation of the lower one. Regarding the MP test, its loss of power comes from the assumption behind the test of a slow moving time-varying parameter dynamic. In fact, in presence of multiple regular switches, the estimated parameter path does note have the necessary *boost* to move across levels when the switching points become closer and closer as their number increases (see figure

3.3). On the other side, the sample autocorrelation function of the observations (and consequently of the score with respect to ω) remains pretty stable as the number of changing points increases, explaining the $LM_{GAS(p,0)}$ robustness.

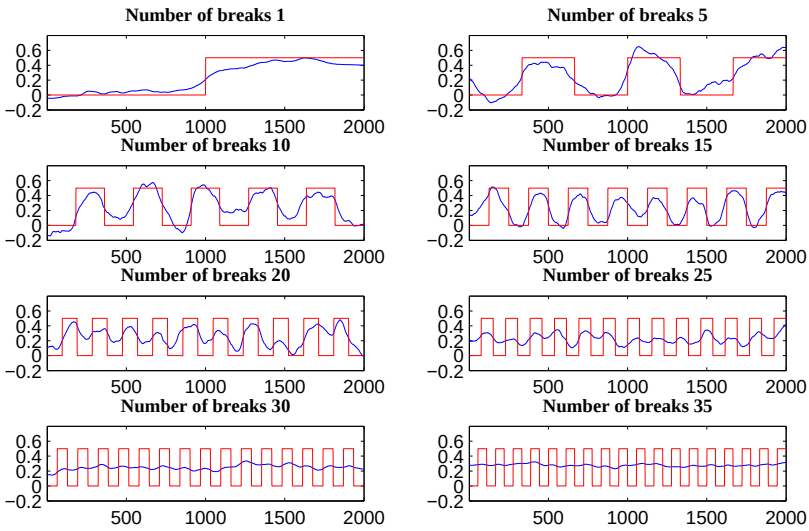

Figure 3.3: Evolution of the parameter path estimated by using the Müller-Petalas procedure (blue line) as the number of switching points in the simulated parameter path (red line) increases.

Above considerations, can be easily extended, with the necessary adjustments, to the results obtained for the DGPs with regular regime switching dynamics for β_t. In fact, by observing the graphs from figure 3.4 to figure 3.6, we can notice that the tests empirical power functions follow substantially the same pattern across DGPs.

Moving to the random breaks class of DGPs, we recall that the variance of the break size (denoted by σ_v^2 in equation (3.22)) is reported on the horizontal axis of each graph. Skimming from figure 3.7 to figure 3.9, the power functions show a similar behaviour.

For the one break case, the three tests are, in terms of performance, in the same rank of the regime–switching case with the difference that the power functions never reach the unit value within the range considered for the variance, and seem to tend to a value smaller then one. In practice, when the changing point

happens at a time too close to 1 or n there is not enough information to catch to parameter instability, even for high values of the break size variance. As the number of random breaks increases, the simulated time-varying parameter path tends to a univariate Gaussian random walk, assuming a more and more erratic behaviour, which makes the detection of the parameter instability almost trivial. The graphs confirm this finding, since all power functions move quickly toward one.

Considering now the the state-space framework, on the horizontal axis of the graphs of figure 3.10 we report the values of the variance σ_η^2 of the state equation error term in the top row and the values of the autoregressive coefficient ϕ of the same equation in the bottom row (see equation (3.23)). In this case the two tests for GAS effects work better for almost all the alternative hypothesis. A further interesting remark is about the top row of figure 3.10, namely the optimality of the MP test for rather small values of σ_η^2. This is expected since the MP test is designed for the alternative of parameter driven parameter instability with persistent and small magnitude variation (*local alternatives*).

In conclusion, the proposed $LM_{GAS(p,0)}$ test, designed for observation driven alternatives, exhibits higher power for alternatives that display regular regime switches or non-local parameter driven time-variation. For parameter driven time variation close to the null or for infrequent structural changes, the test of Müller and Petalas [2010] shows the best performance. Hence these two testing procedures should represent the right starting point of a modelling approach with (potentially) time-varying parameters.

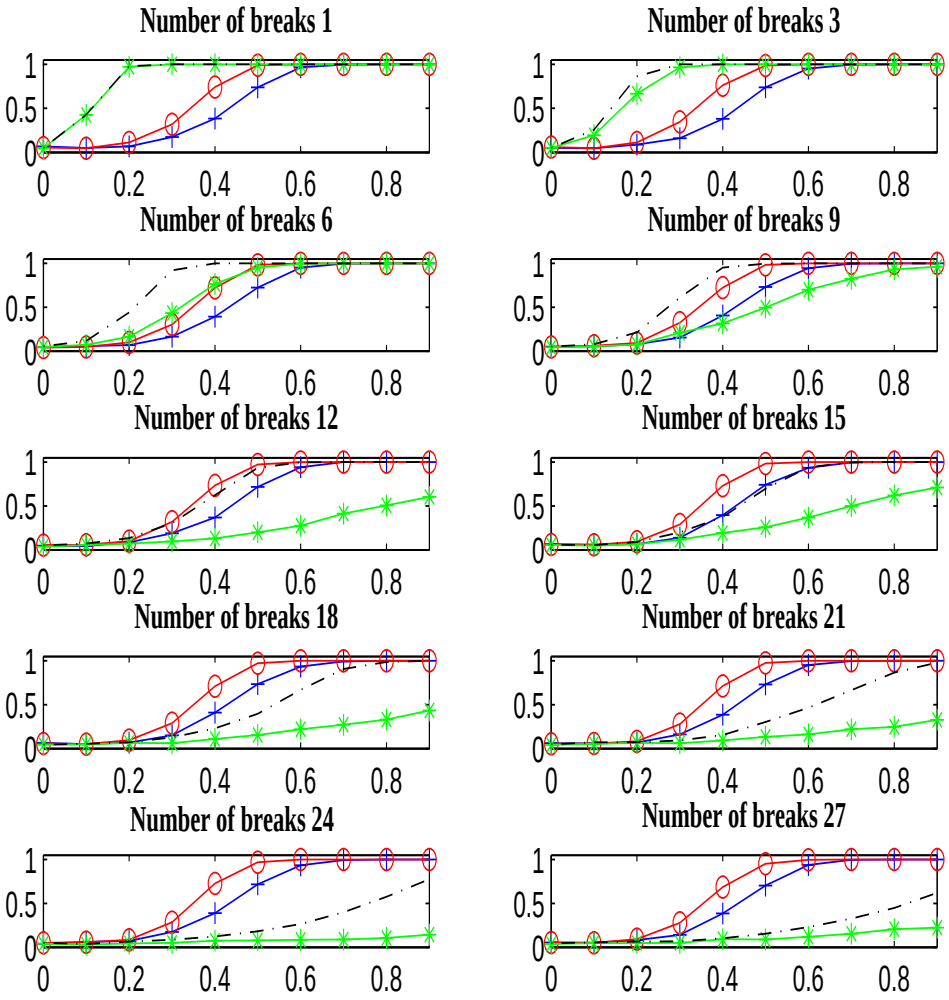

Figure 3.4: Empirical power functions of the $LM_{GAS(1,0)}$ (blue, $-+$), the $LM_{GAS(5,0)}$ (red, $-o$), the Andrews test (green, $-*$), and the Müller-Petalas test (black, $-\cdot$) for the regime-switching time-varying mean model. The horizontal axis shows the values of Δ (see equation (3.21)).

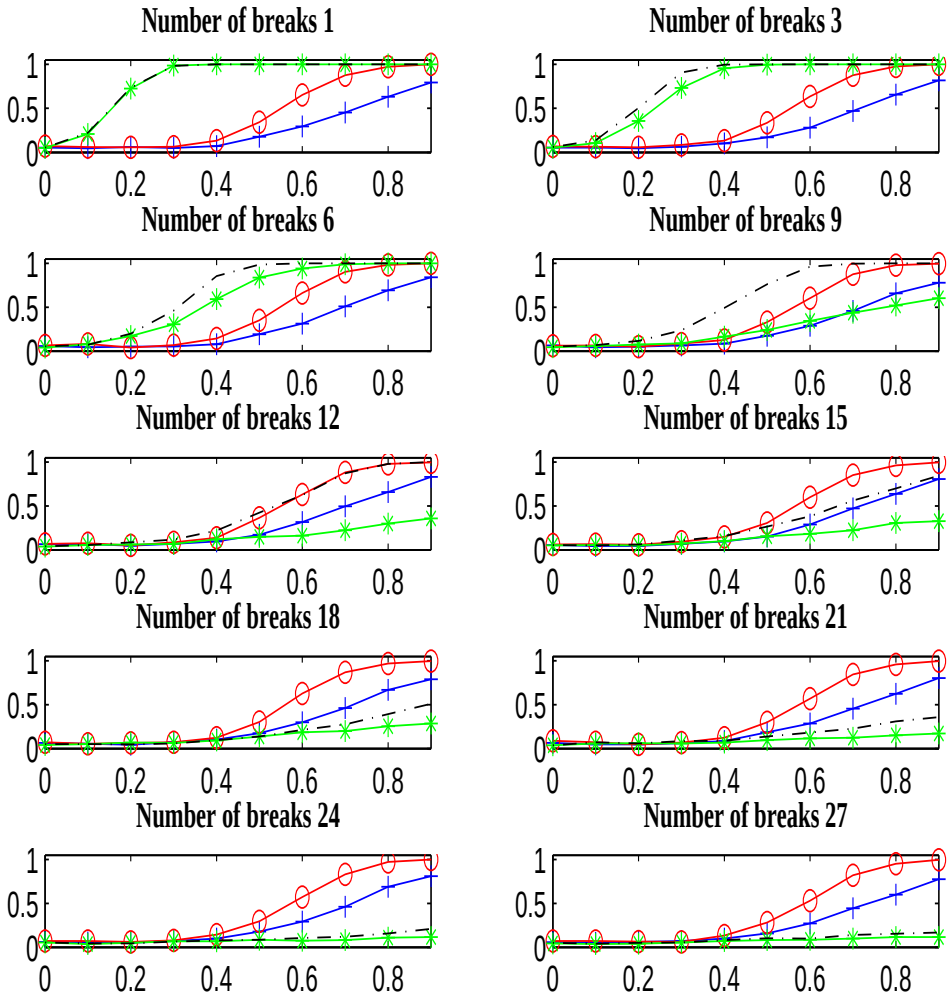

Figure 3.5: Empirical power functions of the $LM_{GAS(1,0)}$ (blue, $-+$), the $LM_{GAS(5,0)}$ (red, $-o$), the Andrews test (green, $-*$), and the Müller-Petalas test (black, $-\cdot$) for the regime-switching time-varying variance model. The horizontal axis shows the values of Δ (see equation (3.21)).

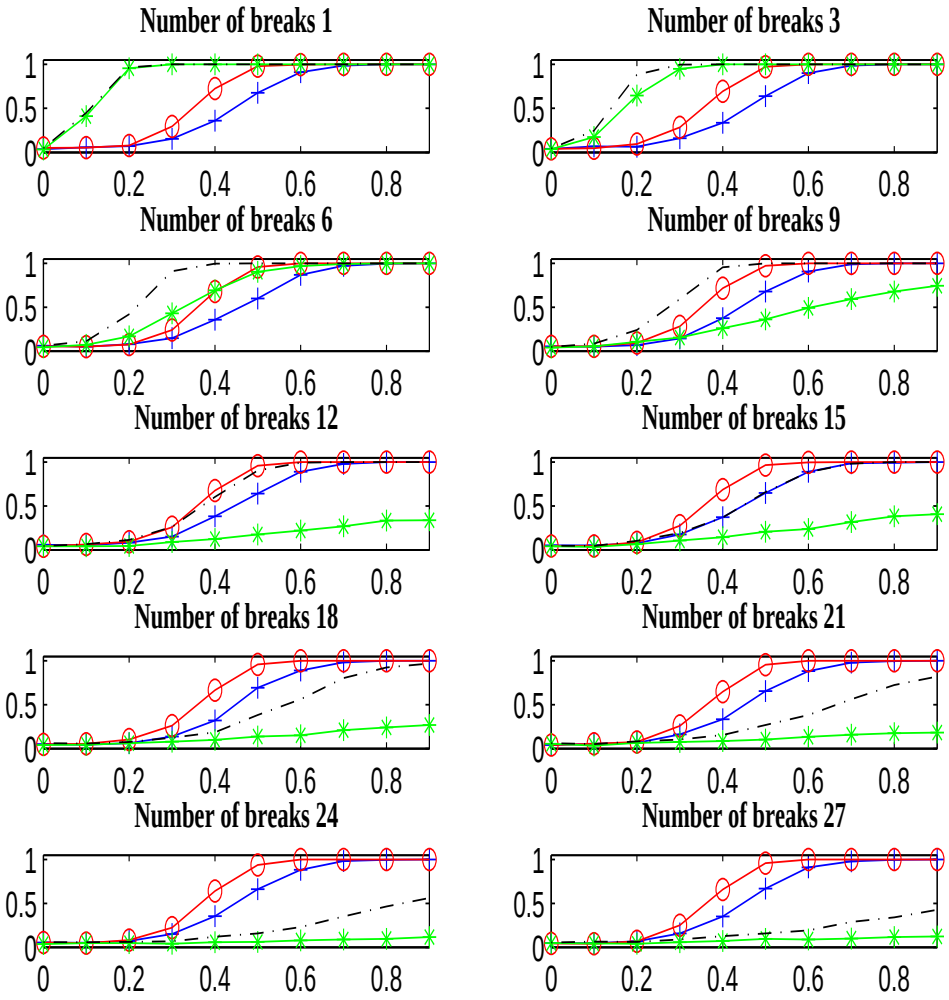

Figure 3.6: Empirical power functions of the $LM_{GAS(1,0)}$ (blue, $-+$), the $LM_{GAS(5,0)}$ (red, $-o$), the Andrews test (green, $-*$), and the Müller-Petalas test (black, $-\cdot$) for the regime-switching time-varying correlation model. The horizontal axis shows the values of Δ (see equation (3.21)).

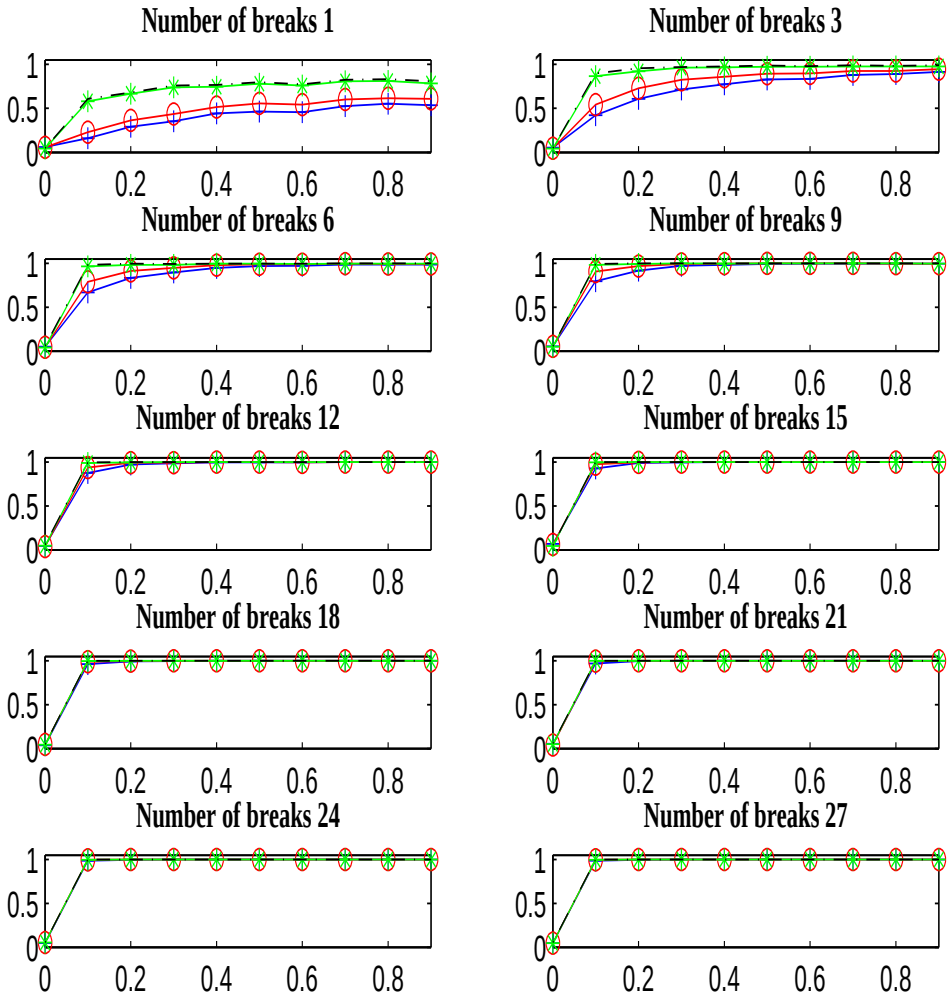

Figure 3.7: Empirical power functions of the $LM_{GAS(1,0)}$ (blue, $-+$), the $LM_{GAS(5,0)}$ (red, $-o$), the Andrews test (green, $-*$), and the Müller-Petalas test (black, $-\cdot$) for the random breaks time-varying mean model. The horizontal axis shows the values of σ_v^2 (see equation (3.22)).

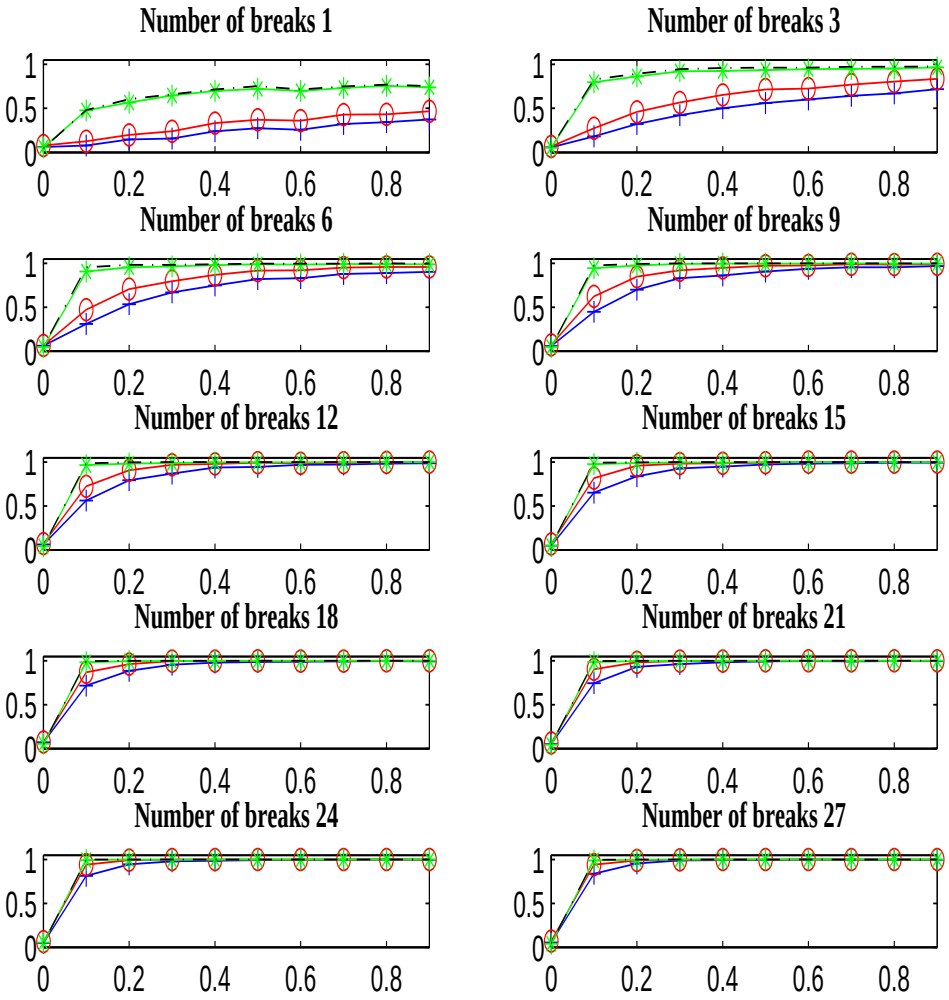

Figure 3.8: Empirical power functions of the $LM_{GAS(1,0)}$ (blue, $-+$), the $LM_{GAS(5,0)}$ (red, $-o$), the Andrews test (green, $-*$), and the Müller-Petalas test (black, $-\cdot$) for the random breaks time-varying variance model. The horizontal axis shows the values of σ_v^2 (see equation (3.22)).

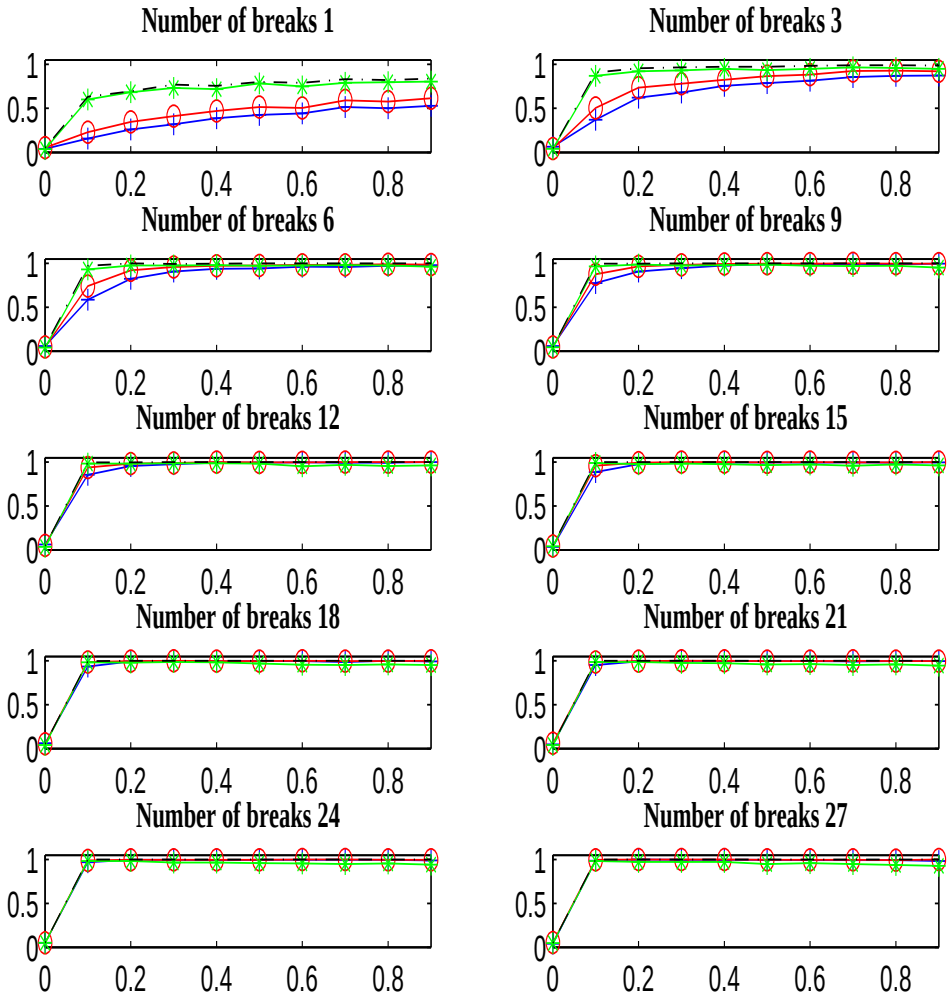

Figure 3.9: Empirical power functions of the $LM_{GAS(1,0)}$ (blue, $-+$), the $LM_{GAS(5,0)}$ (red, $-o$), the Andrews test (green, $-*$), and the Müller-Petalas test (black, $-\cdot$) for the random breaks time-varying correlation model. The horizontal axis shows the values of σ_v^2 (see equation (3.22)).

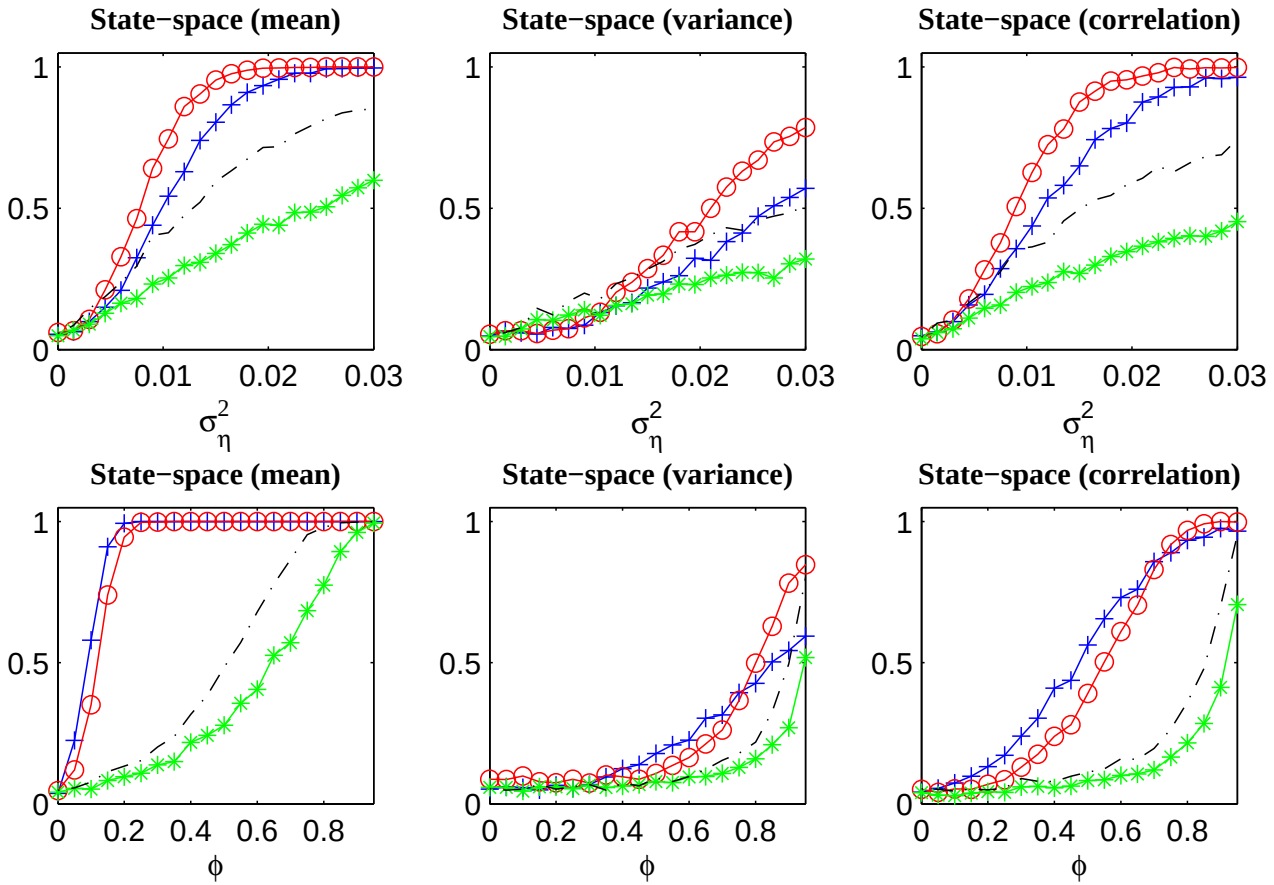

Figure 3.10: Empirical power functions of the $LM_{GAS(1,0)}$ (blue, $-+$), the $LM_{GAS(5,0)}$ (red, $-o$), the Andrews test (green, $-*$), and the Müller-Petalas test (black, $-\cdot$) for the state-space frameworks. The horizontal axis of the first and the second row of graphs shows the values of σ_η^2 and ϕ respectively (see equation (3.23)).

3.6 Empirical Application

3.6.1 Model specification

In this section we present, for illustrative purposes, two different approaches for modelling daily Realized Variance defined in section 2.2.3). Both the considered model specifications are then extended to include different measures for the continuous and the jump component (see section 2.2.4).

To build the time series of interest, we employ almost 7 years of UHFD for the five stocks detailed in table 1.4 spanning from January, 2006 to July, 2012. We first exploit the tools of the software `TAQ_MNGR`, described in section 1.5, to clean the data and build equally spaced price series at 5 minutes. The 5 minutes frequency is properly chosen to avoid the bias induced on the volatility measures by microstructure noise (see section 2.5). The obtained 5 minutes series are then employed to estimate RV_t, TC_t, and TJ_t ($t = 1, \ldots, n$), where RV_t, TC_t, and TJ_t denote respectively the estimated Realized Variance, continuous component and jump component of the tth day.

To disentangle the two risk components we employ a procedure proposed in Corsi et al. [2010] based on the Threshold–Bipower Variation (section 2.3.3). In particular,

$$TJ_t = I_{\{CTz_t > \Phi_h\}} \cdot (RV_t - TBV_t)^+, \qquad TC_t = RV_t - TJ_t,$$

where I denotes the indicator function, CTz_t is a jump detection test (see Corsi et al. [2010] for details) which is normally distributed under the null of absence of jumps, Φ_h is the quantile function of the Gaussian distribution at confidence level h, TBV_t is the TBV computed for the tth day, and $x^+ = \max(x, 0)$.

Recalling that RV_t is a non–negative valued random variable, in our modelling approach we assume:

$$RV_t = \mu_t \epsilon_t,$$

where ϵ_t are iid with distribution

$$p(\epsilon_t; \alpha) = \Gamma(\alpha)^{-1} \epsilon_t^{\alpha-1} \alpha^\alpha \exp(-\alpha\epsilon_t) I_{\{\epsilon_t > 0\}}, \quad \alpha > 0.$$

Regarding the time-varying parameter updating equation, we employ two different specification (in both cases we employ the conditional information matrix to scale the score (see 3.2.1)):

- **MEM(p,q)**: it is a GAS(p,q) specification with $\beta_t = \mu_t$ (this model is denoted by MEM(p,q) because it is equivalent to the Multiplicative Error Model of Engle and Gallo [2006] and Cipollini et al. [2012]):

$$\beta_{t+1} = \theta^{(\omega)} + \sum_{i=1}^{p} \theta_i^{(a)} (RV_{t-i+1} - \beta_{t-i+1}) + \sum_{j=1}^{q} \theta_j^{(c)} \beta_{t-j+1};$$

- **GAS(p,q)**: it is a GAS(p,q) specification with $\beta_t = \ln \mu_t$:

$$\beta_{t+1} = \theta^{(\omega)} + \sum_{i=1}^{p} \theta_i^{(a)} \frac{(RV_{t-i+1} - e^{\beta_{t-i+1}})}{e^{\beta_{t-i+1}}} + \sum_{j=1}^{q} \theta_j^{(c)} \beta_{t-j+1}.$$

Both previous specifications can be easily extended to include TC_t and TJ_t:

- **MEM$_{\text{Jump}}$(p,q)**:

$$\beta_{t+1} = \theta^{(\omega)} + \sum_{i=1}^{p} \theta_i^{(a)} (TC_{t-i+1} - \beta_{t-i+1}) + \sum_{j=1}^{q} \theta_j^{(c)} \beta_{t-j+1} + \theta^{(TJ)} TJ_t;$$

- **GAS$_{\text{Jump}}$(p,q)**:

$$\beta_{t+1} = \theta^{(\omega)} + \sum_{i=1}^{p} \theta_i^{(a)} \frac{(TC_{t-i+1} - e^{\beta_{t-i+1}})}{e^{\beta_{t-i+1}}} + \sum_{j=1}^{q} \theta_j^{(c)} \beta_{t-j+1} + \theta^{(TJ)} TJ_t.$$

From this specifications it follows that

$$E(RV_t|\mathcal{F}_{t-1}) = \mu_t, \qquad \text{var}(RV_t|\mathcal{F}_{t-1}) = \frac{\mu_t^2}{\alpha},$$

with μ_t being a function of the information set $\mathcal{F}_{t-1}$.

3.6.2 Results

We estimate the parameters in the four models by the method of maximum likelihood. The time–varying parameter approach is strongly supported (all the considered test statistics are comfortably within the rejection region) by the results of the $LM_{GAS(p,0)}$ and Müller–Petalas testing procedure reported in table 3.2. The stocks symbols are explained in table 1.4.

	ANF	**BAC**	**C**	**F**	**GE**
$LM_{GAS(1,0)}$	407.80	490.03	275.21	132.73	49.12
$LM_{GAS(5,0)}$	568.81	621.97	472.86	286.07	165.20
Müller–Petalas	-405.62	-769.09	-843.55	-493.85	-640.85

Table 3.2: Parameter instability test statistics.

Tables 3.3–3.7 report, for each considered stock and each model specification, the parameters estimates and corresponding standard errors (computed by using the inverse of minus the Hessian evaluated at the maximum likelihood estimate), the Akaike and the Bayesian Information Criterion (AIC and BIC respectively) and the heteroskedasticity adjusted root mean square error suggested in Bollerslev and Ghysels [1996]:

$$HRMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\frac{RV_t - \widehat{RV}_t}{RV_t}\right)^2},$$

where $\widehat{RV}_t$ is the forecast. Figures 3.11 and 3.12 depict the actual series (RV_t) and the in-sample one-step ahead forecasts ($\widehat{RV}_t = \widehat{\mu}_t$) for each considered stock and each estimated model. Everywhere we use annualized volatility measures. The number of included lags represents an overall reasonable choice in terms of *cleaning* the residuals autocorrelation.

Two main findings can be derived by interpreting the results. First: the MEM and the GAS models show almost the same performance in terms of fitting. Second: the inclusion in the models of the disentangled volatility components doesn't really improve the models fitting. This is clear by observing both information criterion (the smaller the better), in the sense that the bias induced by adding the

new parameter for the jump component is not justified in terms of log–likelihood increase.

However, this is just the starting point, and, probably, the chosen one is not the proper way to include the two risk components in a model for RV_t. Various specifications can be derived by using the GAS approach, but this is the topic of future research.

ANF

	MEM$(2,1)$	GAS$(2,1)$	MEM$_{\text{Jump}}(2,1)$	GAS$_{\text{Jump}}(2,1)$
α	14.5744	14.4451	14.5519	14.4305
	(0.5008)	(0.4963)	(0.5000)	(0.4958)
$\theta^{(\omega)}$	0.0051	-0.0116	0.0061	-0.0194
	(0.0016)	(0.0042)	(0.0019)	(0.0056)
$\theta_1^{(a)}$	0.3642	0.3283	0.3796	0.3488
	(0.0243)	(0.0218)	(0.0271)	(0.0249)
$\theta_2^{(a)}$	-0.1859	-0.1659	-0.1749	-0.1622
	(0.0246)	(0.0229)	(0.0272)	(0.0261)
$\theta_1^{(c)}$	0.9864	0.9884	0.9869	0.9772
	(0.0048)	(0.0039)	(0.0056)	(0.0050)
$\theta^{(TJ)}$	–	–	0.1586	0.3502
	–	–	(0.0279)	(0.0615)
AIC	-4.7290	-4.7196	-2.7039	-2.7128
BIC	22.2955	22.3049	29.7181	29.7093
HRMSE	0.2678	0.2694	0.2691	0.2678

Table 3.3: Estimation results table of the stock Abercrombie & Fitch Co.

BAC

	MEM(2, 1)	GAS(2, 1)	$\text{MEM}_{\text{Jump}}(2,1)$	$\text{GAS}_{\text{Jump}}(2,1)$
α	11.7115	11.4506	11.9371	11.8145
	(0.4013)	(0.3922)	(0.4091)	(0.4049)
$\theta^{(\omega)}$	0.0052	-0.0173	0.0082	-0.0230
	(0.0014)	(0.0051)	(0.0016)	(0.0066)
$\theta_1^{(a)}$	0.5456	0.4542	0.5637	0.4905
	(0.0273)	(0.0222)	(0.0290)	(0.02393)
$\theta_2^{(a)}$	-0.1762	-0.1706	-0.1425	-0.1508
	(0.0272)	(0.0231)	(0.0290)	(0.0262)
$\theta_1^{(c)}$	0.9864	0.9850	0.9897	0.9707
	(0.0057)	(0.0039)	(0.0063)	(0.0050)
$\theta^{(TJ)}$	–	–	0.1088	0.2948
	–	–	(0.0486)	(0.1053)
AIC	-4.8655	-4.8430	-2.8595	-2.8697
BIC	22.1593	22.1814	29.5625	29.5524
HRMSE	0.3067	0.3129	0.3016	0.2996

Table 3.4: Estimation results table of the stock Bank of America Corporation.

C

	MEM(2, 1)	GAS(2, 1)	$\text{MEM}_{\text{Jump}}(2,1)$	$\text{GAS}_{\text{Jump}}(2,1)$
α	12.9711	12.5687	13.0991	12.9621
	(0.4451)	(0.4311)	(0.4495)	(0.4448)
$\theta^{(\omega)}$	0.0050	-0.0169	0.0064	-0.0130
	(0.0015)	(0.0049)	(0.0016)	(0.0064)
$\theta_1^{(a)}$	0.5572	0.4708	0.5571	0.4907
	(0.0268)	(0.0230)	(0.0282)	(0.0244)
$\theta_2^{(a)}$	-0.1541	-0.1672	-0.1150	-0.1170
	(0.0269)	(0.0249)	(0.0280)	(0.0278)
$\theta_1^{(c)}$	0.9893	0.9844	0.9962	0.9738
	(0.0058)	(0.0040)	(0.0063)	(0.0052)
$\theta^{(TJ)}$	–	–	0.1938	0.1254
	–	–	(0.0520)	(0.0735)
AIC	-4.7674	-4.7345	-2.7521	-2.7630
BIC	22.2571	22.2901	29.6699	29.6591
HRMSE	0.3038	0.3472	0.2934	0.2925

Table 3.5: Estimation results table of the stock Citigroup, Inc.

F

	MEM$(2,1)$	GAS$(2,1)$	MEM$_{\text{Jump}}(2,1)$	GAS$_{\text{Jump}}(2,1)$
α	12.6654	12.6099	12.6887	12.5644
	(0.4345)	(0.4325)	(0.4353)	(0.4309)
$\theta^{(\omega)}$	0.0059	-0.01167	0.0059	-0.0213
	(0.0018)	(0.0041)	(0.0018)	(0.0056)
$\theta_1^{(a)}$	0.4413	0.4057	0.4538	0.4127
	(0.0240)	(0.0215)	(0.0247)	(0.0221)
$\theta_2^{(a)}$	-0.2026	-0.1872	-0.2092	-0.2041
	(0.0233)	(0.0218)	(0.0242)	(0.0228)
$\theta_1^{(c)}$	0.9853	0.9882	0.9849	0.9744
	(0.0052)	(0.0038)	(0.0059)	(0.0047)
$\theta^{(TJ)}$	–	–	0.2468	0.3065
	–	–	(0.0379)	(0.0475)
AIC	-4.4813	-4.4761	-2.4571	-2.4689
BIC	22.5431	22.5484	29.9649	29.9531
HRMSE	0.2967	0.2999	0.30261	0.2958

Table 3.6: Estimation results table of the stock Ford Motor Co.

GE

	MEM$(2,1)$	GAS$(2,1)$	MEM$_{\text{Jump}}(2,1)$	GAS$_{\text{Jump}}(2,1)$
α	12.3159	10.8494	12.5470	12.2488
	(0.4223)	(0.3714)	(0.4304)	(0.4200)
$\theta^{(\omega)}$	0.0054	-0.0352	0.0055	-0.0107
	(0.0017)	(0.0085)	(0.0016)	(0.0084)
$\theta_1^{(a)}$	0.4776	0.2791	0.5257	0.4718
	(0.0295)	(0.0200)	(0.0298)	(0.0259)
$\theta_2^{(a)}$	-0.0415	-0.0334	-0.0816	-0.1073
	(0.0303)	(0.0288)	(0.0317)	(0.0297)
$\theta_1^{(c)}$	0.9832	0.9756	0.9931	0.9811
	(0.0083)	(0.0057)	(0.0082)	(0.0057)
$\theta^{(TJ)}$	–	–	0.1885	0.0842
	–	–	(0.0478)	(0.0225)
AIC	-5.3567	-5.2561	-3.3379	-3.3564
BIC	21.6677	21.7684	29.0841	29.0656
HRMSE	0.3026	0.6529	0.2829	0.28153

Table 3.7: Estimation results table of the stock General Electric Company.

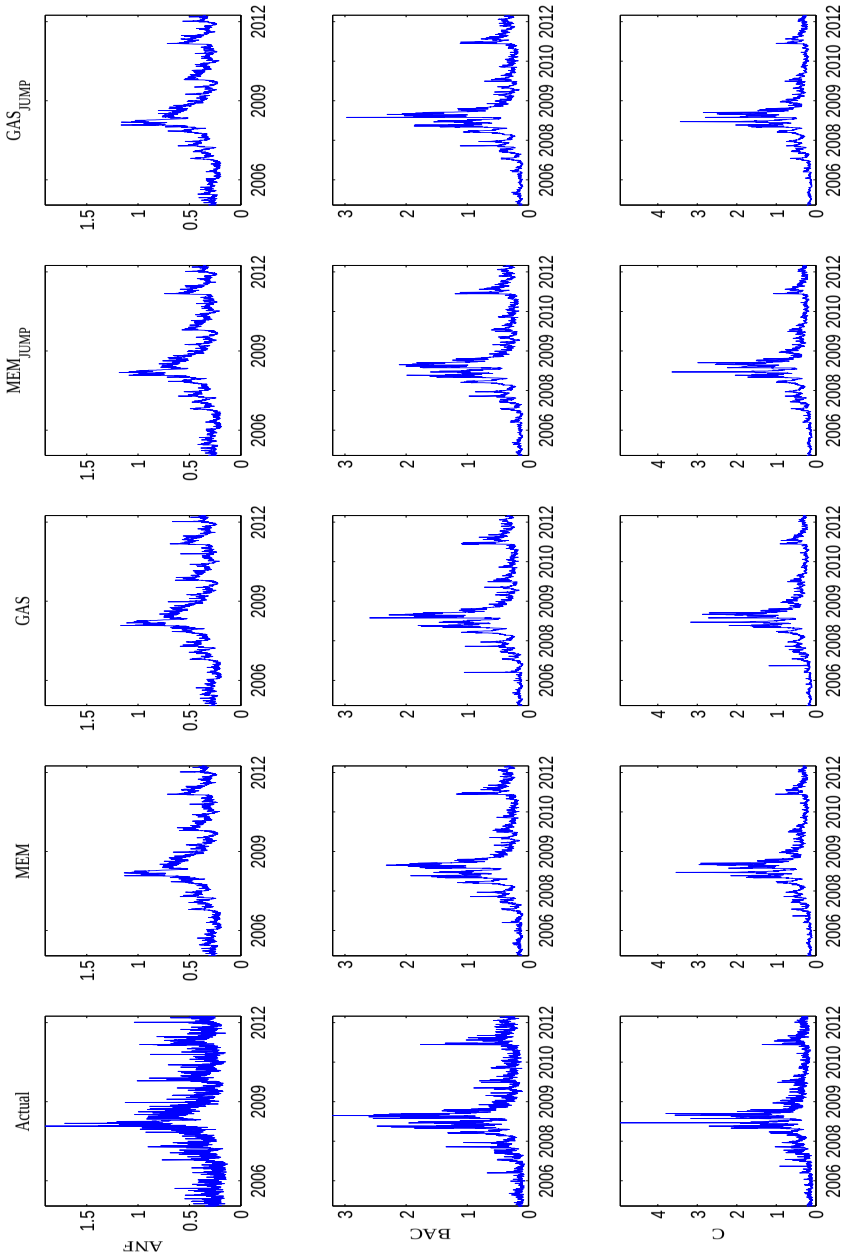

Figure 3.11: Actual and in–sample one–step ahead predicted series.

Figure 3.12: Actual and in–sample one–step ahead predicted series.

References

Y. Aït-Sahalia. Telling from discrete data whether the underlying continuous-time model is a diffusion. *The Journal of Finance*, LVII, n. 5:2067–2112, 2002. 40

Y. Aït-Sahalia and J. Jacod. Fisher's information for discretely sampled Lévy processes. *Econometrica*, 76 (4):727–761, 2008. 41

Y. Aït-Sahalia and J. Jacod. Testing for jumps in a discretely observed process. *Annals of Statistics*, 37:184–222, 2009. 41

Y. Aït-Sahalia and J. Jacod. Testing whether jumps have finite or infinite activity. *Annals of Statistics*, in press. 41

T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys. Modeling and forecasting realized volatility. *Econometrica*, 71:579–625, 2003. 1, 41

T. G. Andersen, D. Dobrev, and E. Schaumburg. Jump-robust volatility estimation using nearest neighbor truncation. *NBER Working Paper No. w15533, ssrn.com*, 2009a. 44, 48

T.G. Andersen and T. Bollerslev. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39:885–905, 1998. 1, 31, 32, 38

T.G. Andersen, T. Bollerslev, and F.X. Diebold. Roughing it up: Including jump components in the measurement, modeling and forecasting of return volatility. *The Review of Economics and Statistics*, 89 (4):701–720, 2007. 39, 47

T.G. Andersen, D. Dobrev, and E. Schaumburg. Duration based volatility estimation. In *Global COE Hi-Stat Discussion, paper series 034, Hitotsubashi University*, 2009b. 41

D. W. K. Andrews. Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856, 1993. 60, 64, 67, 70

D.W.K. Andrews and W. Ploberger. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, 62(6):1383–1414, 1994. 65

F.M. Bandi and T.H. Nguyen. On the functional estimation of jump-diffusion models. *J. Econometrics*, 116:293–328, 2003. 41

F.M. Bandi and R. Renó. Non-parametric stochastic volatility. *www.ssrn.com*, 2010. 41

F.M. Bandi and J.R. Russell. Microstructure noise, realized volatility, and optimal sampling. *Review of Economic Studies*, 75 (2):339–369, 2008. 56

O.E. Barndorff-Nielsen and N. Shephard. Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2 (1):1–37, 2004. 42

O.E. Barndorff-Nielsen and N. Shephard. Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics*, 4:1–30, 2006. 39, 42, 43

L. Bauwens and P. Giot. *Econmetric Modelling of Stock Market Intraday Activity*. Kluwer, Dordrecht, 2001. 7

L. Bauwens and D. Veredas. The stochastic conditional duration model: A latent factor model for the analysis of financial durations. *Journal of Econometrics*, 119(2):381–412, 2004. 65

E. Bohemer, J. Gramming, and E. Theissen. Estimating the probability of informed trading - does trade misclassification matter? *Journal of Financial Markets*, 10:26–47, 2007. 14

T. Bollerslev. Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 31:307–327, 1986. 31, 61, 62

T. Bollerslev. Financial econometrics: past developments and future challenges. *Journal of Econometrics*, 100:45–51, 2001. 6

T. Bollerslev and E. Ghysels. Periodic autoregressive conditional heteroskedasticity. *Journal of Business & Economic Statistics*, 14:139–151, 1996. 83

K. Boudt, C. Croux, and S. Laurent. Outlyingness weighted quadratic covariation. *forthcoming in Journal of Financial Econometrics*, 2010. 49

C. T. Brownlees and G. M. Gallo. Comparison of volatility measures: a risk management perspective. *Journal of Financial Econometrics*, 8:29–56, 2010. 39

C. T. Brownless and G. M. Gallo. Financial econometric analysis at ultra-high frequency: data handling concerns. *Computational Statistics & Data Analysis*, 51:2232–2245, 2006a. 2, 7, 13, 14, 18

C. T. Brownless and G. M. Gallo. Financial econometric analysis at ultra-high frequency: Data handling concerns. *Working Paper Dipartimento di Statistica "G. Parenti"*, 2006b. 7

F. Calvori, S. J. Koopman, and A. Lucas. Testing for parameter instability in alternative modeling frameworks. *Working Paper*, 2012. 59

P. Carr, H. Geman, D. Madan, and M. Yor. The fine structure of asset returns: An empirical investigation. *Journal of Business*, 75 (2):305–332, 2002. 37

K. Christensen and M. Podolskij. Realised range-based estimation of integrated variance. *Journal of Econometrics*, 141:323–349, 2007. 50

K. Christensen and M. Podolskij. Range-based estimation of quadratic variation. *ssrn.com*, 2010. 50, 51

K. Christensen, R. Oomen, and M. Podolskij. Realised quantile-based estimation of the integrated variance. *Journal of Econometrics*, 159:74–98, 2010. 48, 54

REFERENCES

F. Cipollini, R. F. Engle, and G. M. Gallo. Semiparametric vector MEM. *Journal of Applied Econometrics*, Published online, 2012. 61, 82

R. Cont and P. Tankov. *Financial Modelling with Jump Processes*. CRC Press, 2004. 34, 36, 38

F. Corsi, D. Pirino, and R. Renó. Threshold bipower variation and the impact of jumps on volatility forecasting. *Journal of Econometrics*, 159:276–288, 2010. 2, 39, 46, 47, 81

D. R. Cox. Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics*, 8:93–115, 1981. 60

Drew Creal, Siem Jan Koopman, and André Lucas. Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 27, 2012. 59, 60, 61, 62

M. M. Dacorogna, R. Gencay, U. A. Müller, R. Olsen, and Pictet. *An Introduction to High Frequency Finance*. Academic Press, London, 2001. 17

R. Davidson and J. G. MacKinnon. Specification tests based on artificial regressions. *Journal of the American Statistical Association*, 85(409):220–227, 1990. 63

R. A. Davis, W. T. M. Dunsmuir, and S. Streett. Observation driven models for poisson counts. *Biometrika*, 90(4):777–790, 2003. 61

D. Dobrev. Capturing volatility from large price moves: generalized range theory and applications. *working paper*, 2007. 33, 50, 51, 54

A. Dufour and R. F. Engle. Time and the price impact of a trade. *Journal of Finance*, 555:2467–2498, 2000. 14

J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press, 2nd edition, 2012. 66

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982. 31, 59, 61

R. F. Engle. The economics of ultra high frequency data. *Econometrica*, 68:1–22, 2000. 5

R. F. Engle. New frontiers for arch models. *Journal of Applied Econometrics*, 17: 425–446, 2002. 3

R. F. Engle and G. M. Gallo. A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics*, 131:3–27, 2006. 3, 61, 82

R. F. Engle and J. R. Russell. Autoregressive conditional duration: A new model for volatility using intra-daily data. *Econometrica*, 66(5):1127–1162, 1998. 1, 61

R. F. Engle and J. R. Russell. *Handbook of Financial Econometrics*, chapter Analysis of high-frequency data. Elsevier, 2006. 5

T. N. Falkenberry. High frequency data filtering. *Technical Report, Tick Data*, 2002. 13

C.M. Hafner and H. Manner. Dynamic stochastic copula models: Estimation, inference and applications. *Journal of Applied Econometrics*, (27):269–295, 2012. 65

J. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989. 66

B.E. Hansen. The new econometrics of structural change: Dating breaks in us labor productivity. *Journal of Economic perspectives*, 15(4):117–128, 2001. 65

P.R. Hansen and A. Lunde. Realized variance and market microstructure noise. *J. Bus. Econom. Statist.*, 24 (2):127–218, 2006. 1, 56

A. C. Harvey. *Forecasting, structural time series models and the Kalman Filter*. Cambridge University Press, Cambridge, 1989. 65

J. Hasbrouck. Using the torq database. *NYSE working paper*, 92-05, 1992. 7, 9, 11

REFERENCES

J. Hasbrouck, G. Sofianos, and D. Sosebee. New york stock exchange system and trading procedures. *NYSE working paper*, 93-01, 1993. 9

X. Huang and G. Tauchen. The relative contribution of jumps to total price variance. *Journal of Financial Econometrics*, 3 (4):456–499, 2005. 36

J. Jacod. *Statistics and high-frequency data. Lecture Notes, SEMSTAT in La Manga (Spain) 2007*. in press. 33, 34

J. Jacod and P. Protter. Asymptotic error distributions for the euler method for stochastic differential equations. *The Annals of Probability*, 26 (1):267–307, 1998. 41

J. Jacod and A.N. Shiryaev. *Limit Theorems for Stochastic Processes, 2nd ed.* Springer-Verlag, Berlin, 2003. 33, 38

J. Jacod and V. Todorov. Testing for common arrival of jumps for discretely observed multidimensional processes. *Annals of Statistics*, 37:1792–1838, 2009. 46

I. Karatzas and S.E. Shreve. *Brownian motion and stochastic calculus*. Springer, 1999. 44

J. Lahaye, S. Laurent, and Neely C.J. Jumps, co-jumps and macro announcements. *forthcoming in Journal of Applied Econometrics*. 33

C. M. C. Lee and M. J. Ready. Inferring trade direction from intraday data. *Journal of Finance*, 46:733–746, 1991. 7

S. Lee and J. Hanning. Detecting jumps from Lévy jump diffusion processes. *Journal of Financial Economics*, 96 (2):271–290, 2010. 40

Y. Li and P. Mykland. Are volatility estimators robust with respect to modeling assumptions? *Bernoulli*, 13 (3):601–622, 2007. 56

D.B. Madan. *Option Pricing, Interest Rates and Risk Management*, chapter Purely discontinuous asset price processes, pages 105–153. Cambridge University Press, 2001. 33

A. Madhavan and G. Sofianos. An empirical analysis of the nyse specialist trading. *Journal of Financial Economics*, 48:189–210, 1998. 9

C. Mancini. Disentangling the jumps of the diffusion in a geometric jumping brownian motion. *Giornale dell'Istituto Italiano degli Attuari*, LXIV:19–47, 2001. 44

C. Mancini. Estimation of the parameters of jump of a general poisson-diffusion model. *Scandinavian Actuarial Journal*, 1:42–52, 2004. 45

C. Mancini. Optimal threshold for the estimator of integrated variance. *working paper*, 2008. 46

C. Mancini. Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. *Scandinavian Journal of Statistics*, 36:270–296, 2009. 43, 44, 45

C. Mancini. Test for the relevance of noise in observed data. Technical report, 2011, working paper. 56

C. Mancini and R. Renó. Threshold estimation of markov models with jumps and interest rate modeling. *Journal of Econometrics, available on www.ssrn.com*, in press. 41, 45, 46

V. Mattiussi and R. Renó. Spot volatility estimation using delta sequences. *www.ssrn.com*, 2010. 41

U. K. Müller and P. E. Petalas. Efficient estimation of the parameter path in unstable time series models. *The Review of Economic Studies*, 77:1508–1539, 2010. 60, 66, 67, 73

H. L. Ngo and S. Ogawa. A central limit theorem for the functional estimation of the spot volatility. *Monte Carlo Methods and Applications*, 15 (4):353–380, 2009. 41

M. O' Hara. *Market Microstructure Theory*. 1997. 9

W. Ploberger, W. Krämer, and K. Kontrus. A new test for structural stability in the linear regression model. *Journal of Econometrics*, 40(2):307–318, 1989. 65

P.E. Protter. *Stochastic integration and differential equations*. Springer, 2005. 32

J. R. Russell. Econometric modeling of multivariate irregularly-spaced high-frequency data. Unpublished manuscript, University of Chicago, Graduate School of Business, 2001. 61

N. Shephard. *Stochastic Volatility: Selected Readings*. Oxford: Oxford University Press, 2005. 31, 65

G. Sofianos and I. M. Werner. The trades of nyse floor brokers. *Journal of Financial Markets*, 3:139–176, 2000. 11

O. Vergote. How to match trades and quotes for nyse stocks? *K. U. working paper, Katholieke Universiteit Leuven*, 2005. 7, 14

M. Vetter. Limit theorems for bipower variation of semimartingales. *Stochastic Processes and Their Applications*, 120:22–38, 2010. 43

J.H.C. Woerner. *Stochastic Finance*, chapter Power and Multipower variation: inference for high-frequency data, pages 343–364. Springer, 2006. 43

J. Wright and H. Zhou. Bond risk premia and realized jump volatility. *working Paper. Federal Reserve Board.*, 2007. 39