Ultra High Frequency Data: Gold Mining Opportunities for Regulation

Giampiero M. Gallo
Dipartimento di Statistica, Informatica, Applicazioni (DiSIA) *G. Parenti*
Università di Firenze

COEURE Seminar, *European University Institute*
June 6, 2015 Firenze

Outline

My Terms of Reference today:

- Present some elements of Ultra High Frequency Data based research
- Share the viewpoint that regulation should not a response to past crises
- If not ahead of practitioners/traders/IT at least on the same wavelength
- Applied research tends to be US centered
- Recommendation: give European applied research more data (preferebly free) and transparency of what is going on within European exchanges

Why Interest in UHFD?

- Exchange of assets in the presence of constraints, of market regulation, of different horizons of interests (e.g. intra–daily dealers vs institutional investors)
- Market activity is translated into a 'tick': the elementary record (quote or transaction price with a timestamp and additional information)
- Generation of thousands of observations per day. Formidable challenge for IT for data feed, storage, cleaning, manipulation, pattern discovery (archetypal concept of *big data*)
- Trading decisions are made on the basis of sequences of tick data (value for practitioners in real time)
- More accurate characterization of volatility and hence risk measurement.

Where Do UHFD Come From?

Need to pay for them almost invariably.

- Academic research counts on them to characterize market activity behavior.
- Real time data provision on screen (e.g. subscription to Bloomberg or Reuters services)
- Purchase/access more or less expensive data provider services (retrospectively)
- Poor man's alternative: data feed and subsequent data storage (possibility of capturing freely available updates like on finance.yahoo; IT consuming and not very reliable)

What are UHFD?

Essentially three types

- Trades: data on actual transactions, time of execution, price and volume exchanged, where
- Quotes: data on potential transaction: time and best bid price and ask prices
- Limit Order Book: data on the *n* best bid and ask prices with quantities associated with the order. Book depth important for liquidity analysis.

What Do UHFD Look Like?

Reference to TAQ data from NYSE (available through WRDS which acts as a data broker; considerable delay brings the price down). Substantial academic discounts given with the aim to

- Publish new theories and strategies to predict pricing trends and investment behavior
- Backtest existing trading strategies
- Research markets for regulatory or audit activity

Recall Olsen and Associates' (a Forex trading company) effort in mid '90s to foster research in the area of UHFD by distributing Forex tick by tick data as a playing ground for researchers. Special issue of JoEF in 1997.

What Do UHFD Look Like?

Example 1: Trades (sec, from WRDS)

SYMBOL	DATE	TIME	PRICE	G127	CORR	COND	EX	SIZE
AMZN	20120621	6:27:13	223.000	0	0	T	P	100
AMZN	20120621	7:49:15	224.000	0	0	T	P	300
...								
AMZN	20120621	9:30:00	223.840	0	0	@O	Q	19953
AMZN	20120621	9:30:00	223.840	0	0	Q	Q	19953
AMZN	20120621	9:30:00	223.840	0	0		Q	100
AMZN	20120621	9:30:00	223.860	0	0		B	100
...								
AMZN	20120621	15:59:59	220.520	0	0		P	100
AMZN	20120621	16:00:00	220.575	0	0	@6	Q	92001
AMZN	20120621	16:00:00	220.575	0	0	M	Q	92001
AMZN	20120621	16:00:00	220.520	0	0	M	P	100
...								
AMZN	20120621	18:48:35	220.700	0	0	T	P	100
AMZN	20120621	18:48:35	220.700	0	0	T	P	100

What Do UHFD Look Like?

Example 2: Trades (millisec, from WRDS)

SYM_ROOT	DATE	TIME_M	EX	TR_SCOND	SIZE	PRICE	TR_CORR	TR_SEQNUM	TR_SOURCE
AMZN	20120621	6:27:13.814	P	T	100	223.000	00	934	N
AMZN	20120621	7:49:15.534	P	T	300	224.000	00	1075	N
...									
AMZN	20120621	9:30:00.182	Q	@O X	19953	223.840	00	4818	N
AMZN	20120621	9:30:00.182	Q	Q	19953	223.840	00	4819	N
AMZN	20120621	9:30:00.365	Q		100	223.840	00	5085	N
AMZN	20120621	9:30:00.365	B		100	223.860	00	5086	N
...									
AMZN	20120621	15:59:59.400	P		100	220.520	00	1451704	N
AMZN	20120621	16:00:00.270	Q	@6 X	92001	220.575	00	1452350	N
AMZN	20120621	16:00:00.270	Q	M	92001	220.575	00	1452351	N
AMZN	20120621	16:00:00.424	P	M	100	220.520	00	1452601	N
...									
AMZN	20120621	18:48:35.159	P	T	100	220.700	00	1467970	N
AMZN	20120621	18:48:35.763	P	T	100	220.700	00	1467971	N

What Do UHFD Look Like?

Example 3: Quotes (millisec, from WRDS)

SYM_ROOT	DATE	TIME_M	EX	BID	BIDSIZ	ASK	ASKSIZ	QU_COND	BIDEX	ASKEX	QU_SEQNUM	NASDBBO_IND	QU_SOURCE
AMZN	20120621	9:30:00.011	B	223.19	1	224.03	1	R	B	B	514524	0	N
AMZN	20120621	9:30:00.011	Y	223.19	1	224.03	1	R	Y	Y	514534	0	N
AMZN	20120621	9:30:00.015	Z	223.03	5	268.74	1	R	Z	Z	514591	0	N
AMZN	20120621	9:30:00.042	Y	223.19	1	224.02	1	R	Y	Y	514814	0	N
AMZN	20120621	9:30:00.046	B	223.19	1	224.02	1	R	B	B	514838	0	N
AMZN	20120621	9:30:00.052	Y	223.19	1	224.01	1	R	Y	Y	514895	0	N
AMZN	20120621	9:30:00.058	B	223.19	1	224.01	1	R	B	B	515010	0	N
AMZN	20120621	9:30:00.091	B	223.19	1	224.02	1	R	B	B	515423	0	N
AMZN	20120621	9:30:00.094	Y	223.19	1	224.02	1	R	Y	Y	515467	0	N
AMZN	20120621	9:30:00.112	B	223.19	1	224.03	1	R	B	B	515658	0	N
...													

What Do UHFD Look Like?

Example 4: Limit Order Book (nanosec, from LOBSTER)

Nikolaus Hautsch's Project (Berlin-Vienna)

Time	Type	Order ID	Size	Price	Direction	APrice1	ASize1	BPrice1	BSize1	APrice2	ASize2	BPrice2	BSize2	APrice3	ASize3	BPrice3	BSize3
34200.004241176	1	16113575	18	58533	1	58594	200	58533	18	58598	200	58530	150	58610	200	58510	5
34200.004260640	1	16113584	18	58532	1	58594	200	58533	18	58598	200	58532	18	58610	200	58530	150
34200.004447484	1	16113594	18	58531	1	58594	200	58533	18	58598	200	58532	18	58610	200	58531	18
34200.025551909	1	16120456	18	58591	-1	58591	18	58533	18	58594	200	58532	18	58598	200	58531	18
34200.025579546	1	16120480	18	58592	-1	58591	18	58533	18	58592	18	58532	18	58594	200	58531	18
34200.025613151	1	16120503	18	58593	-1	58591	18	58533	18	58592	18	58532	18	58593	18	58531	18
34200.201517942	1	16166035	100	58593	-1	58591	18	58533	18	58592	18	58532	18	58593	118	58531	18
...																	

UHFD Quality

Take tick–by–tick stock data from TAQ/WRDS

- Raw data may have some data outside market opening hours and 'wrong' records (outliers). Example: JNJ, 1998-2013.

Records	Off Time Scale	Outliers	Off Time Scale (pct)
107279000	72613	27938	0.0676862

- Irregularly spaced data; may need aggregation at regular intervals (more below). Example: 15–minute aggregated data include 108675 records
- TAQMNGR (R package – free): clean, aggregate, read data, according to Brownlees and Gallo (2006).

Realm of Applicability

- Tick data:
  - Durations (Engle and Russell, 1998) modeled as an autoregressive process;
  - Interaction between trades and quotes (Engle and Lunde, 2003)
  - Time and price impact of a trade (Dufour and Engle, 2000)
  - Order Book dynamics (LOBSTER; Hautsch and Huang, 2012)
  - Market microstructure dynamics: the role of informed and uninformed traders (Easley *et al.*, 2008)

Realm of Applicability

- Intra–daily manipulation.
  - Irregularly spaced: price or volume durations
  - Regularly spaced: n–minute intervals; returns, volatility, volume, number of trades
- End–of–day manipulation: Realized Volatility literature (Andersen *et al.*, 2008)

Conditional Modeling Paradigm

- Observed series are sequences of numbers indexed by time. What we know up to time $t-1$ can be included in I_{t-1}, information set available at time $t-1$.
- Any variable of interest X_t can be seen as decomposable into two components $\mu_t = E(X_t|I_{t-1})$ a known function of I_{t-1}, and ϵ_t, a random variable unpredictable as of $t-1$ but *neutral* relative to I_{t-1}. We can have
  - Additive error models $X_t = \mu_t + \epsilon_t$, $E(\epsilon_t|I_{t-1}) = 0$
  - Multiplicative error models $X_t = \mu_t\,\epsilon_t$, $E(\epsilon_t|I_{t-1}) = 1$

Multiplicative Error Models

Most financial time series cannot take on negative values. New class of models (Engle, 2002; Engle and Gallo, 2006)

$$X_t = \mu_t \epsilon_t$$

with

$$\mu_t = \omega + \alpha_1 X_{t-1} + \beta_1 \mu_{t-1} + \cdots ; \qquad \epsilon \sim \mathit{Gamma}(1, \phi^2)$$

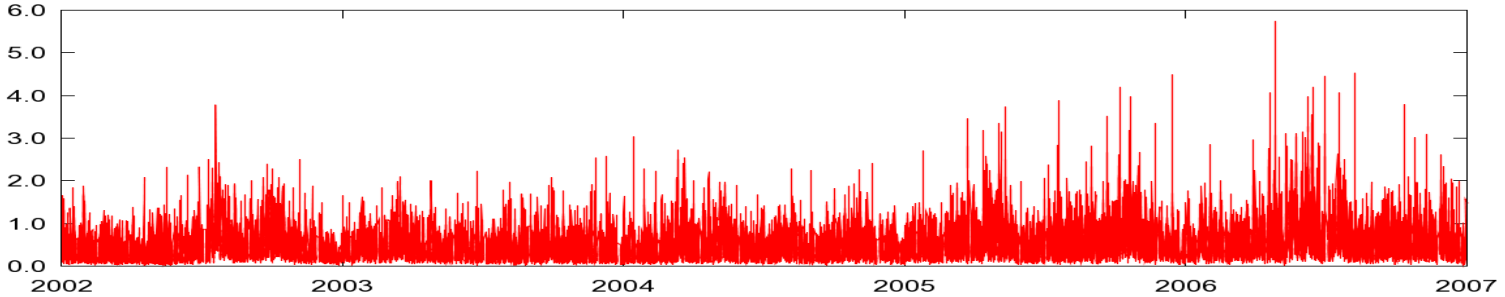
Example: $\text{Volume}_{t,\tau}$ = Expected $\text{Volume}_{t,\tau}$ as of $(t, \tau - 1)$ or $t - 1 \times$ unpredictable error. How to specify the Expected $\text{Volume}_{t,\tau}$ given the past will be suggested by data features. San Diego approach to financial time series analysis.

An Example: 15–minute Volume Modeling

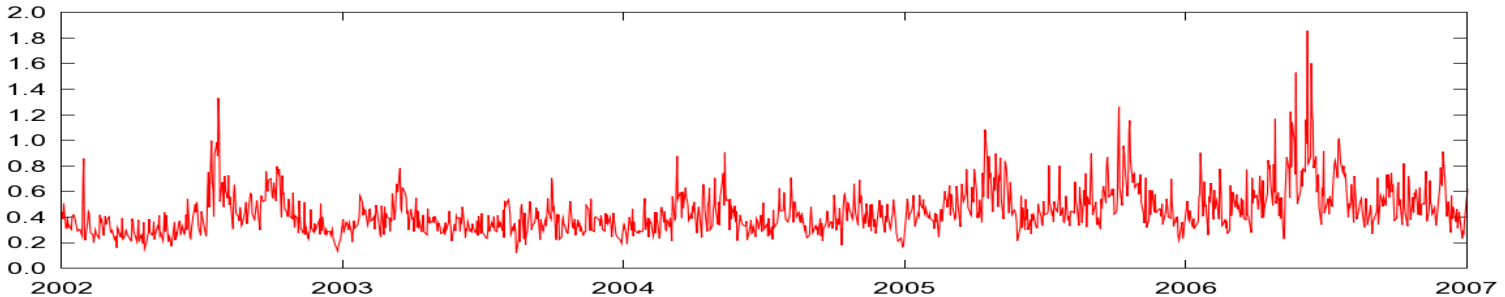
- Algorithmic Trading trying to reproduce the Volume Weighted Average Price (sometimes guaranteed to customers)
- Brownlees *et al.* (2011) show that it is relevant to forecast 15-minute volumes
- Component MEM accommodates various dynamics (intra–daily periodic, intra–daily non periodic, daily).

SPY - Empirical Regularities: Volumes

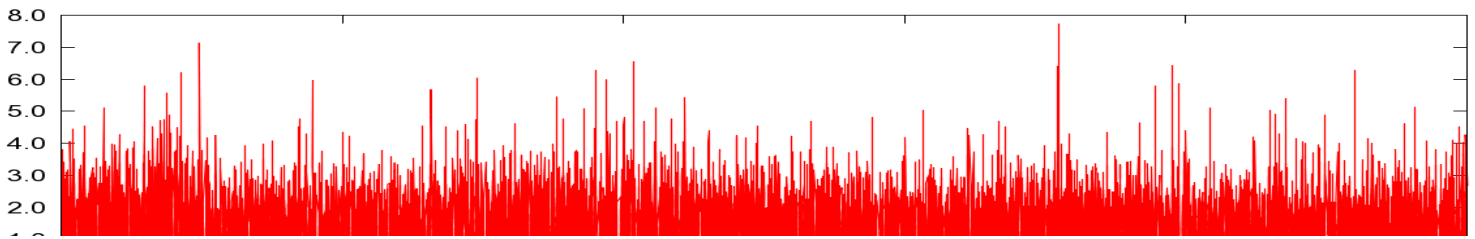
Overall



Daily

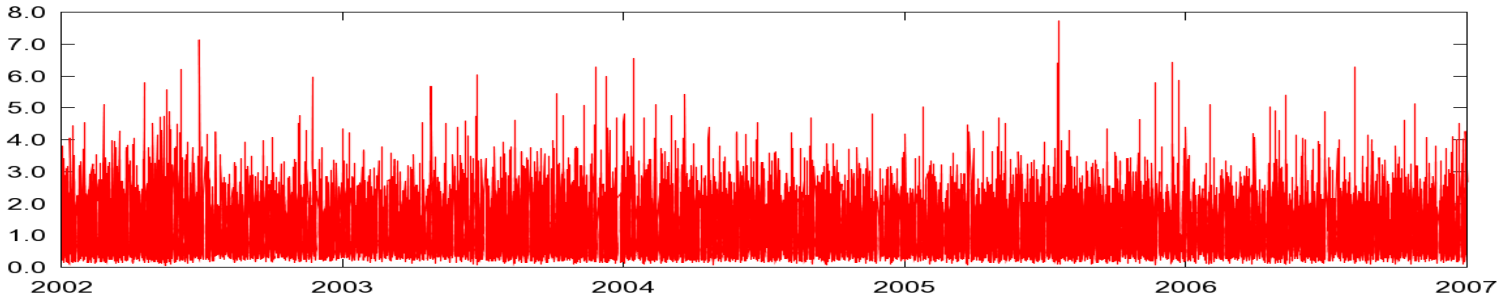


Intra–daily

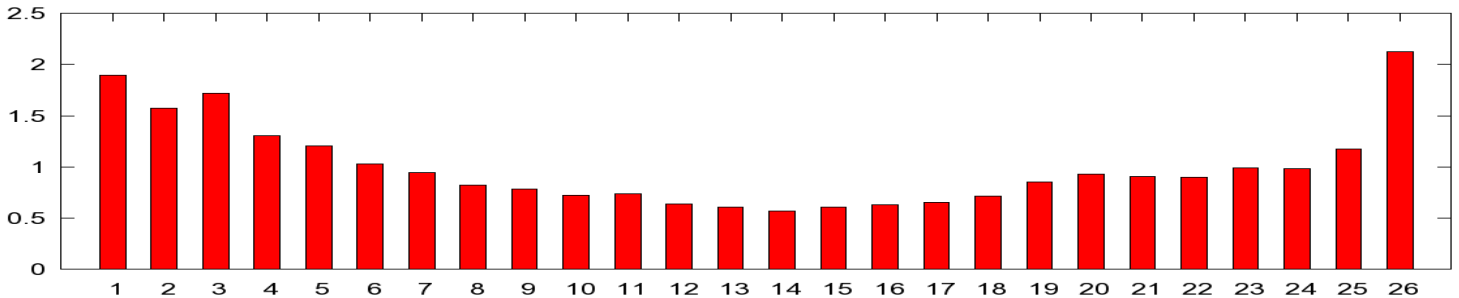


SPY Empirical Regularities: Intra–daily Components

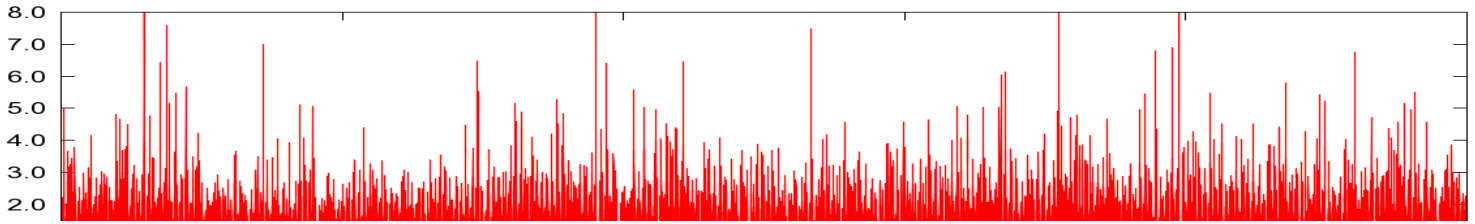
Intra–daily



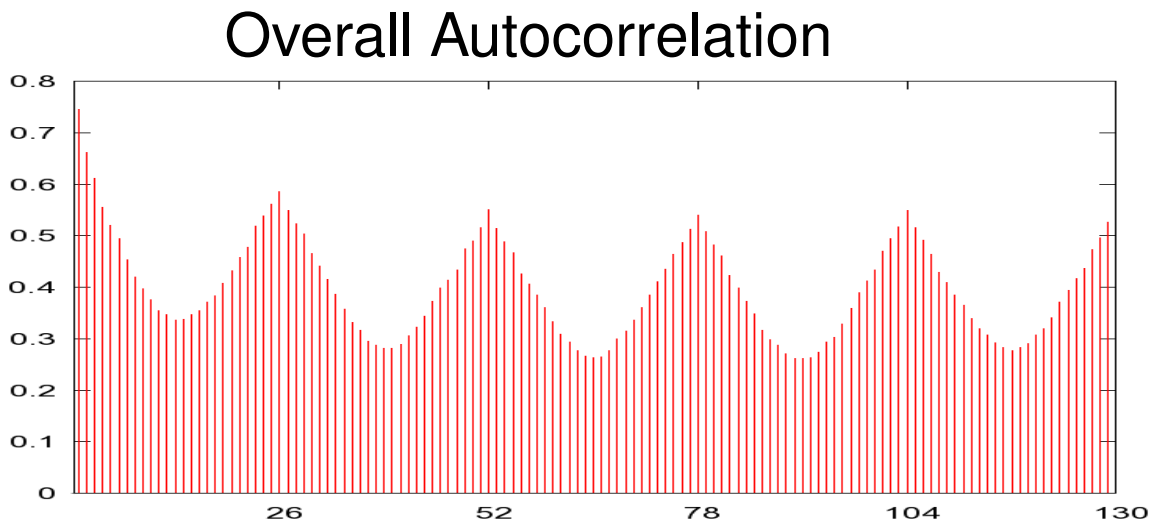
Intra–daily Periodic

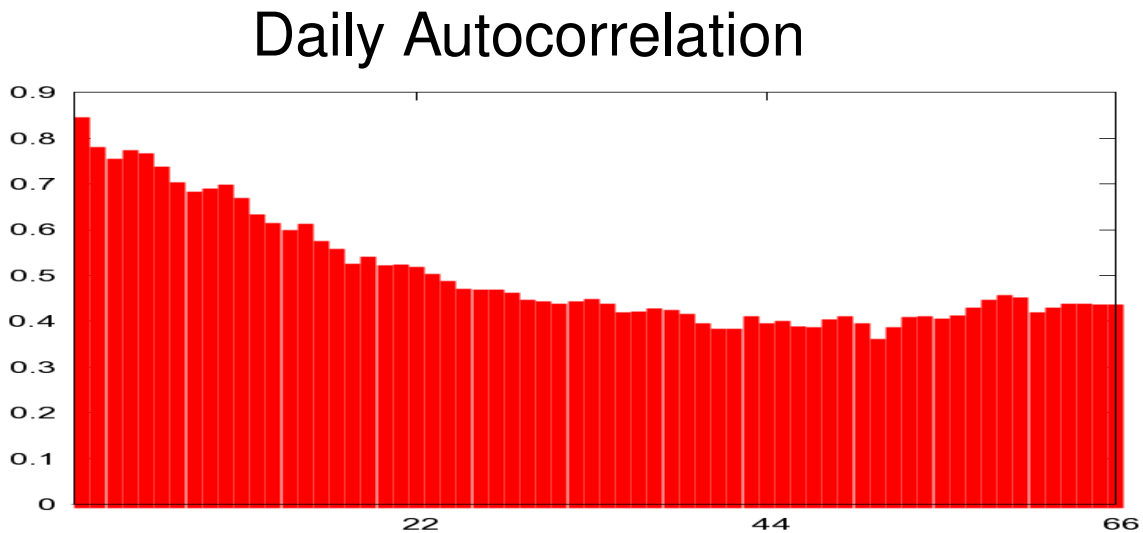


Intra–daily Non–periodic



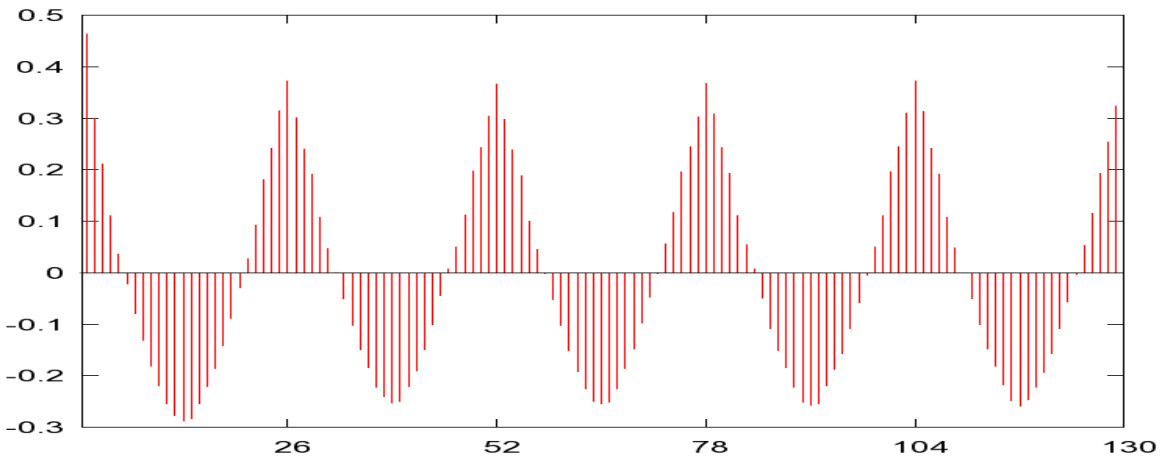
SPY Empirical Regularities: Autocorrelations



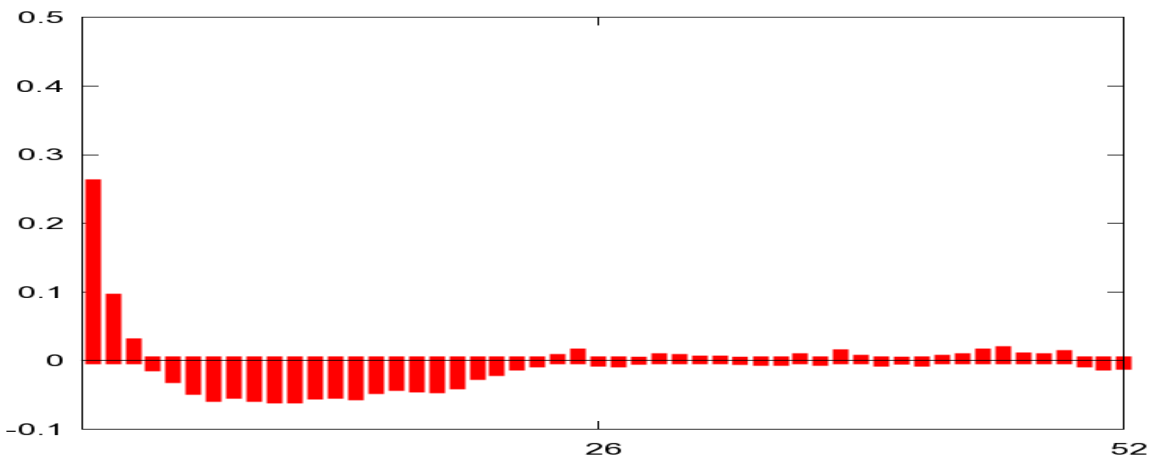


SPY Empirical Regularities: Autocorrelations

Intra–daily Autocorrelation (w/ periodic)



Intra–daily Autocorrelations (w/o periodic)



Some Descriptive Statistics

Autocorrelations at selected lags

	overall		daily		intra	w/ per.	intra	w/o per.
	$\hat{\rho}_1$	$\hat{\rho}_{1 \text{ day}}$	$\hat{\rho}_1$	$\hat{\rho}_{1 \text{ week}}$	$\hat{\rho}_1$	$\hat{\rho}_{1 \text{ day}}$	$\hat{\rho}_1$	$\hat{\rho}_{1 \text{ day}}$
DIA	0.60	0.41	0.72	0.60	0.35	0.26	0.18	0.02
QQQQ	0.69	0.53	0.77	0.66	0.53	0.45	0.29	0.01
SPY	0.74	0.59	0.84	0.76	0.46	0.37	0.26	0.00

Is it all here?

- Is the the structure of these data this simple? No :-(
- Complications:
  1. Extreme Observations (Block Trades? System Errors?)
  2. Intra–daily periodic pattern appears to be weekday dependent (For ETFs! only)

Is it all here?

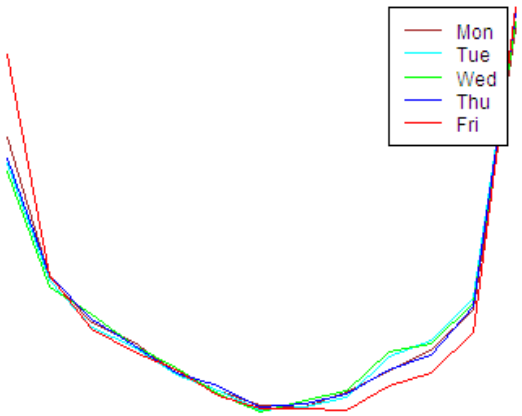
- Is the the structure of these data this simple? No :-(
- Complications:
  1. Extreme Observations (Block Trades? System Errors?)
  2. Intra–daily periodic pattern appears to be weekday dependent (For ETFs! only)

Is it all here?

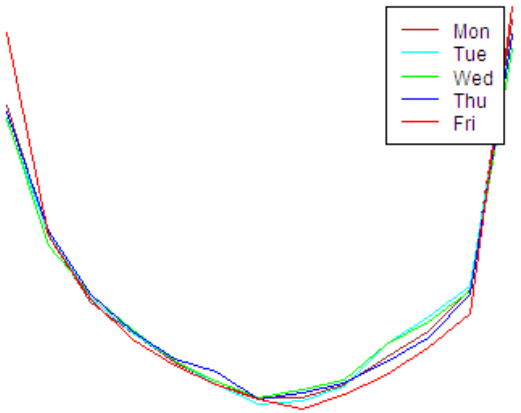
- Is the the structure of these data this simple? No :-(
- Complications:
  1. Extreme Observations (Block Trades? System Errors?)
  2. Intra–daily periodic pattern appears to be weekday dependent (For ETFs! only)

More on the periodic patterns: Stocks Vs. ETFs

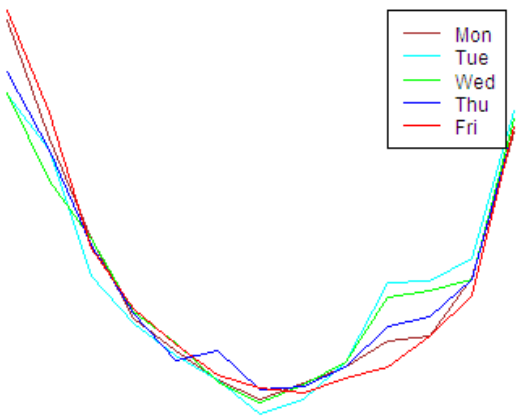
GE



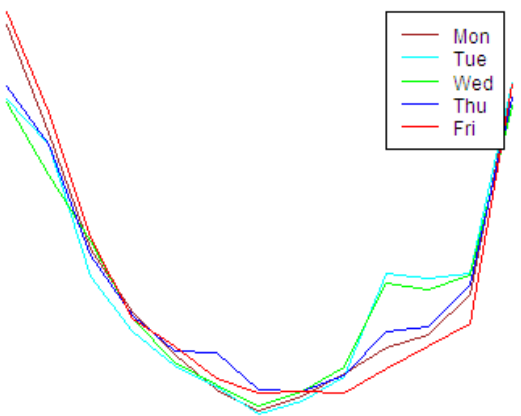
IBM



DIA



SPY



Realized Volatility

- Main field of academic research boosted by UHFD data
- Tick data sampled at 1– or 5–minute intervals: intra–daily returns aggregated into a daily measure of volatility (cf. Andersen *et al.*, 2008)
- Large Realized Volatility literature on different flavors available for more accurate measure of end–of–day return variance (ex post).
- Growing literature on *ex ante* volatility forecasting (expand the information set beyond past realvol).
- Analysis of volatility spillovers (dynamic interdepencence across markets)
- Multivariate extensions Realized Covariance

Risk Management/Stress Tests

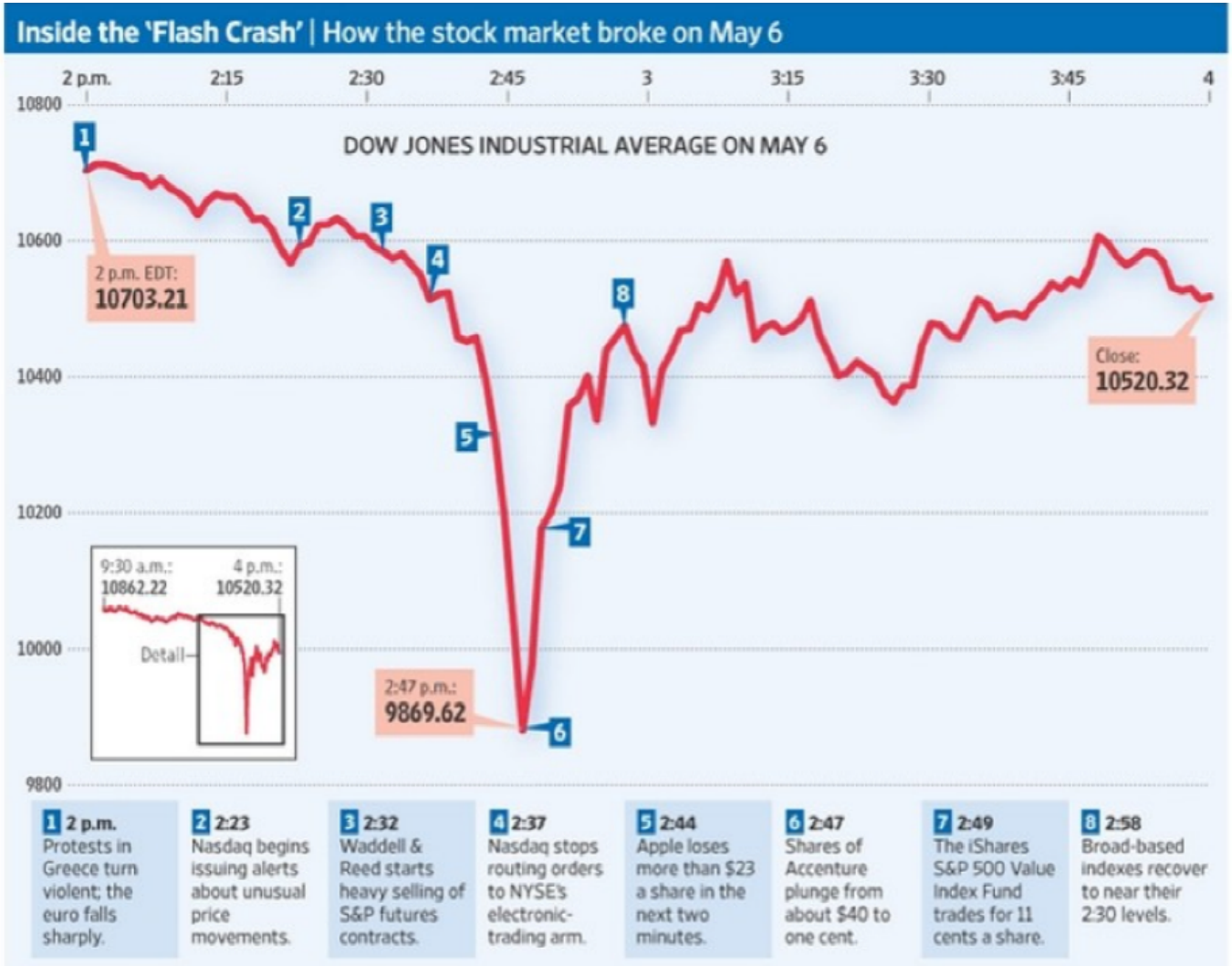
- Stress testing framework built as a synthesis of market risk characterization (conditional distributions)
- Selection of levels of inputs for internal models, but which?
- Hierarchy of risk factors; start from an originator at a given α
- Reproduce correlation structure to derive the consistent levels in other risk factors
- What level of joint probability of occurrence are we considering?

Where Speed Matters

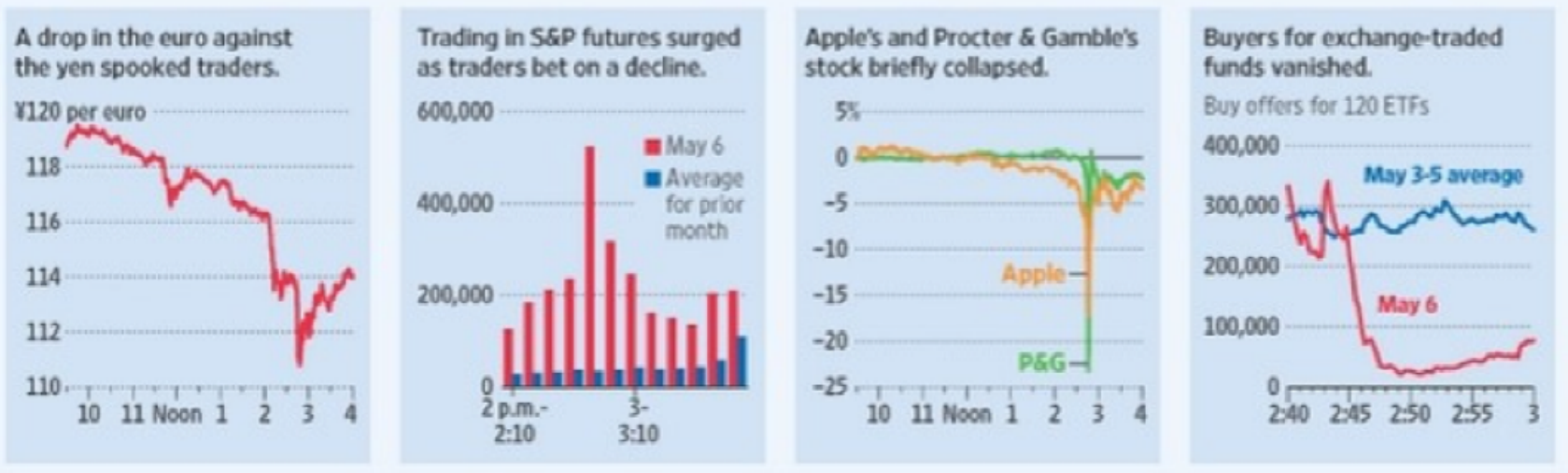
Analysis on how markets work: intra-daily trading considerations

- Algo trading: software driven trading where patterns are detected on an asset realized dynamics and trigger buy/sell decisions
- Play the order book to send signals that may be quickly reverted in order to attract "predictable" orders (being ready on the other side)
- Scan different exchanges in order to detect tiny arbitrage opportunities (e.g. NYSE/BATS). Speed is of the essence: latency arbitrage run on the basis of a few milliseconds

Why Should Regulators Care



Look for Deviant Behavior



Wrapping Up

UHF data useful for

- More accurate measurement of market activity
- Risk Management tools (VaR, Expected Shortfall, CoVaR, etc.)
- Analysis of price formation mechanisms (market liquidity, liquidity risk). Limit order book dynamics for reduced transaction costs
- Detection/Characterization of inefficiencies (bid-ask spread dynamics)
- Impact of high frequency trading on transaction costs, volatility, liquidity
- Detection of insider trading

In a Dream World

Increase the availability of freely available data for all realms of empirical analysis

- Empirical analysis is attracted by the availability of data
- Huge potential of analysis to be redirected toward European markets (difference with US?)
- Institutional knowledge increased, fill the gap *vis á vis* practitioners
- Paradoxical that some free data on European stocks are available only through finance.yahoo
- Some arrogance on the part of exchange managers: need for more transparency
- FRED or QuandL as leading examples

A Concluding Thought

To manage

- to store enough information out of what is being produced,
- to turn that information into knowledge (understanding), and
- to raise that knowledge to a wisdom level (I guess that is where regulation is born of)

A Concluding Thought

To manage

- to store enough information out of what is being produced,
- to turn that information into knowledge (understanding), and
- to raise that knowledge to a wisdom level (I guess that is where regulation is born of)

References

- Andersen, T.G., Bollerslev, T., Christoffersen, P.F., Diebold, F.X. (2006). Volatility and Correlation Forecasting. In Elliott, G., Granger, C.W.J. and Timmermann, A. (eds.), *Handbook of Economic Forecasting*, North Holland.
- Brownlees, C.T. and Gallo, G.M. (2006). Financial Econometric Analysis at Ultra–High Frequency: Data Handling Concerns. *Computational Statistics and Data Analysis*, **51**, 2232-2245.
- Dufour, A. and Engle, R.F. (2000). Time and the Price Impact of a Trade. *The Journal of Finance*, **55**, 2467-2498.
- Easley, D., Engle, R.F. O'Hara M., and Wu L. (2008) Time-Varying Arrival Rates of Informed and Uninformed Trades *Journal of Financial Econometrics*, **6**, 171-207.
- Engle, R. F. (2000). The Econometrics of Ultra-High-Frequency Data. *Econometrica*, **68**, 1-22.
- Engle, R.F. (2002). New Frontiers for ARCH Models. *Journal of Applied Econometrics*, **17**, 425-446.
- Engle, R.F. and Lunde, A. (2003). Trades and Quotes: A Bivariate Point Process. *Journal of Financial Econometrics*, **1**, 159-188.
- Engle, R.F. and Russell, J.R. (1998). Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. *Econometrica*, **66**, 1127-62.
- Hautsch, N. and Huang, R. (2012). The Market Impact of a Limit Order. *Journal of Economic Dynamics and Control*, **36**, 501-522.