MTH 9879 Market Microstructure Models, Spring 2019

Lecture 7: Long memory of order flow and market impact

Jim Gatheral
Department of Mathematics

Outline of Lecture 7

- Long memory of order flow
- Definitions of market impact
- Why market impact is a concave function of volume.
- Two models of market impact:
  - Permanent, state-dependent impact
  - Transient impact
- Equivalence of these two models
- Including limit orders and cancelations

Long memory of order flow

- Prices evolve as a function of order flow and the arrival of new orders in response to that order flow.
- Price dynamics (especially dynamics of the mid-quote) are well-described by Brownian motion.
- As we have seen already, order flow is a highly autocorrelated long-memory process.

- We know that the price process is reasonably efficient, i.e., a martingale
  - That is price changes are almost uncorrelated (martingality of the price process) which implies that variance grows approximately linearly in trading time.
- It follows that the market response to order flow must strongly depend on the past history of order flow.

Long memory of order flow

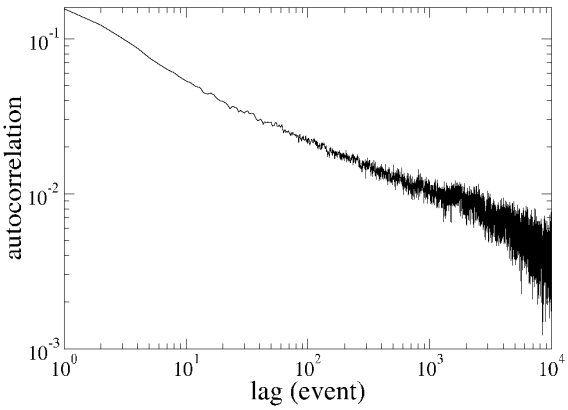


Figure 1. Autocorrelation function of the time series of signs of orders that result in immediate trades (effective market orders) for the stock Vodafone traded on the London Stock Exchange in the period May 2000 - December 2002, a total of 5.8×10^5 events.

Autocorrelation of BAC trade signs

In [1]:
```
download.file(url="https://mfe.baruch.cuny.edu/wp-content/uploads/2018/02/tqDataB
AC_20170919.zip", destfile="tq.zip")
unzip(zipfile="tq.zip")

load("tqDataBAC_20170919.rData")
Sys.setenv(TZ='EST')
```

In [3]:
```
library(highfrequency)
```

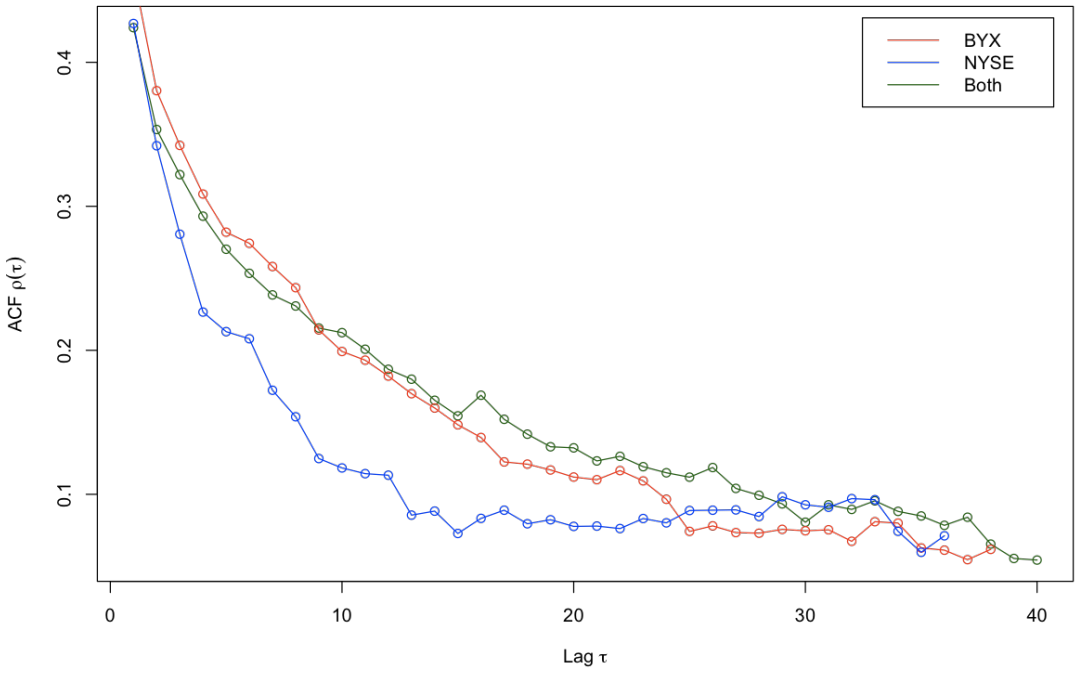
In [4]:
```
tmp <- sapply(getTradeDirection(tqdata), toString)
ts <- cbind(tqdata, tmp)
colnames(ts) <- c(colnames(tqdata), "SIGN")
```

In [5]:
```
library(repr)
options(repr.plot.height=7, repr.plot.width=10)
```

In [6]:
```
# Compute tradesigns
ts.Y <- as.numeric(ts$SIGN[ts$EX=="Y"])
ts.N <- as.numeric(ts$SIGN[ts$EX=="N"])
ts.both <- as.numeric(ts$SIGN[(ts$EX=="Y")|(ts$EX=="N")])
# Compute autocorrelation function of trade signs
ac.Y <- acf(ts.Y, plot=F)
ac.N <- acf(ts.N, plot=F)
ac.both <- acf(ts.both,plot=F)

# AC computations
lag.Y <- ac.Y$lag[-1]
acf.Y <- ac.Y$acf[-1]

lag.N <- ac.N$lag[-1]
acf.N <- ac.N$acf[-1]

lag.both <- ac.both$lag[-1]
acf.both <- ac.both$acf[-1]
```

ACF plot

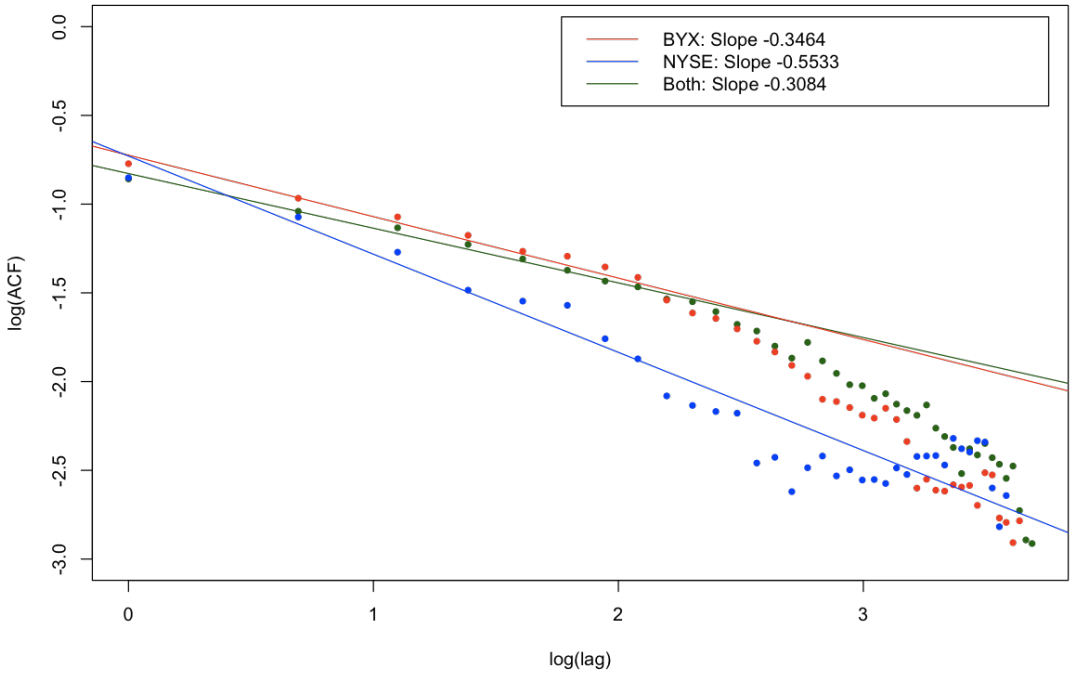
In [7]:
```
plot(lag.both, acf.both, type="o", xlab=expression(paste("Lag ",tau)),
     ylab=expression(paste("ACF ", rho(tau))), col="dark green")
lines(lag.Y, acf.Y, type="o", xlab="log(lag)", ylab="log(ACF)", col="red")
lines(lag.N, acf.N, type="o", xlab="log(lag)", ylab="log(ACF)", col="blue")
legend("topright", c("BYX", "NYSE", "Both"), col=c("red", "blue", "dark green"),
       lty=1, inset=0.02)
```



BAC autocorrelation log-log plot

In [8]:
```
fit.both <- lm(log(acf.both)[1:10] ~ log(lag.both)[1:10])
slope.both <- fit.both$coef[2]
fit.Y <- lm(log(acf.Y)[1:10] ~ log(lag.Y)[1:10])
slope.Y <- fit.Y$coef[2]
fit.N <- lm(log(acf.N)[1:10] ~ log(lag.N)[1:10])
slope.N <- fit.N$coef[2]
```

In [9]:
```
plot(log(lag.both), log(acf.both), xlab="log(lag)", ylab="log(ACF)", col="dark gr
een",
     pch=20, ylim=c(-3,0))
abline(fit.both, col="dark green")
points(log(lag.Y), log(acf.Y), xlab="log(lag)", ylab="log(ACF)", col="red", pch=2
0)
abline(fit.Y, col="red")
points(log(lag.N), log(acf.N), xlab="log(lag)", ylab="log(ACF)", col="blue", pch=
20)
abline(fit.N, col="blue")

leg1 <- paste("BYX: Slope", format(slope.Y, digits = 4), "               ")
leg2 <- paste("NYSE: Slope", format(slope.N, digits = 4))
leg3 <- paste("Both: Slope", format(slope.both, digits = 4))

legend("topright", c(leg1, leg2, leg3), col=c("red", "blue", "dark green"), lty=1
,
       inset=0.02)
```



Data from 19-Sep-2017 is from the SIAC feed courtesy of Richard Holowczak.

Long memory processes

- Stochastic processes for which the autocorrelation function decays asymptotically as a power-law with an exponent smaller than one are called *long-memory* processes.
- For such a process,

$$\rho(\tau) \sim \frac{L(\tau)}{\tau^{\alpha}}$$

where $L(\cdot)$ is a slowly-varying function and $\alpha < 1$.
  - $L(x)$ is a slowly varying function if

$$\lim_{x\to\infty} \frac{L(t\,x)}{L(x)} = 1,\ \forall t.$$

- Models of long-memory processes include Fractional Brownian Motion and ARFIMA (sometimes called FARIMA).
  - Autoregressive Fractionally Integrated Moving Average
- Volume and volatility were also both widely believed to be long-memory processes.
  - At least until our work on rough volatility!

ARFIMA

- A series $\{X_t\}$ is $ARFIMA(p, d, q)$ if the series $(1 - B)^d\, X_t$ is a stationary $ARMA(p, q)$ time series where B is the backward shift operator.
- The special case $d = 0$ gives $ARMA(p, q)$ and the case $d = 1$ gives $ARIMA(p, q)$.

- The fractional difference operator $(1 - B)^d$ is defined by the binomial series

$$(1 - B)^d = 1 + \sum_{j=1}^{\infty} \frac{d\,(d-1)\ldots(d-j+1)}{j!}\, B^j$$

- When $d < 1/2$, the autocorrelation function

$$\rho(k) \sim k^{2\,d-1} \qquad \text{as } k \to \infty$$

Unwrapping the definition of ARFIMA

- An $ARMA(p, q)$ model has the form

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i\, X_{t-i} + \eta_t - \sum_{j=1}^{q} \theta_j\, \eta_{t-j}$$

- This can be rewritten using the backshift operator B as

$$\left(1 - \sum_{i=1}^{p} \phi_i\, B^i\right) X_t = \phi_0 + \left(1 - \sum_{j=1}^{q} \theta_j\, B^j\right) \eta_t$$

- If X_t is $ARIMA(p, q)$, then $Y_t = (1 - B)X_t$ is $ARMA(p, q)$. X_t is then "integrated" $ARMA(p, q)$.
- If $(1 - B)^d X_t$ is $ARMA(p, q)$ with $d \in (-1/2, 1/2)$, X_t is then "fractionally integrated" $ARMA(p, q)$.

The simplest case: $ARFIMA(0, d, 0)$

- In this case, $(1 - B)^d X_t = \eta_t$.
  - $d = 0$ gives white noise; $d = 1$ gives a random walk.
- Then X_t has the MA representation

(1)

$$X_t = \eta_t + \sum_{j=1}^{\infty} \psi_j \eta_{t-j}$$

$$\text{with } \psi_j = (-1)^j \binom{-d}{j} = \frac{d\,(d-1)\ldots\,(d-j+1)}{j!}$$

- The ACF of X_t is (2)

$$\rho(k) = \frac{\Gamma(1-2\,d)}{\{\Gamma(1-d)\}^2}\,\frac{d\,(1+d)\cdots\,(k-1+d)}{(1-d)\,(2-d)\,\cdots\,(k-d)}$$
$$\sim \frac{\Gamma(1-2\,d)}{\Gamma(d)\,\Gamma(1-d)}\,k^{2\,d-1} \qquad \text{as } k \to \infty$$

Empirical results

- In an earlier slide, we found $\rho(\tau) \sim \tau^{-\alpha}$ with $\alpha \approx 0.3$ for BAC.
- Lillo and Farmer found $\alpha \approx 0.6$ on the LSE.
- Bouchaud et al. found $\alpha \in (0.2, 0.7)$ for stocks on the Paris Stock Exchange.
- Note that estimating α as we just did by fitting a straight line in a log-log plot of the autocorrelation function is notoriously inaccurate. Much more dependable techniques are available.

Explanations for long-memory

There are two main explanations:

- Herding behavior by traders.
  - We saw previously that herding can be optimal strategic behavior. Buy orders follow buy orders etc.
  - There can be two types of herding:
    - Traders respond in the same way to public information.
    - Some traders copy other traders.
- Splitting of large trades.
  - At any given time, there is insufficient depth in the order book to accommodate a large trade. Such trades need to be split.
  - This is also consistent with the optimal strategy of the informed trader in the Kyle model.
- [Tóth, Lillo et al.][9] argue convincingly that the dominant effect is order-splitting.

Order splitting

- We call a large order that is yet to be revealed to the market a *hidden* order or a *metaorder*.
  - The trader knows his intentions but others in the market do not.
  - The objective of the trader is to minimize market impact by hiding his intentions for as long as possible.
- Large institutional trades such as pension fund manager transitions often take many weeks to complete.

A simple model for order splitting

In this model from [Lillo, Mike and Farmer][6],

- There are N meta orders in the market whose sizes V_n, $n \in \{1, \cdots, N\}$, are drawn from the distribution $p(V)$.
- Meta order signs s_n are randomly chosen, iid, with equal probability.
- At time t, an existing meta order $O_t = n$ is chosen with uniform probability and one unit is traded so that

$$V_n(t+1) = V_n(t) - 1.$$

  - This generates a child order of size 1 and sign s_n.
- A meta order dies if $V_n(t+1) = 0$ in which case a new meta order is generated.

The Lillo, Mike, Farmer model of order splitting

We paraphrase the argument of [Lillo, Mike and Farmer][6]:

- Let ϵ_t be the sign of the child order observed at time t.
- Consider the autocorrelation function $\rho(\tau) = \mathbb{E}\left[\epsilon_t \, \epsilon_{t+\tau}\right]$ of order signs. Note that $\mathbb{E}\left[\epsilon_t\right] = 0$ and $\mathbb{E}\left[\epsilon_t^2\right] = 1$ for all t.
- By assumption, if two child orders come from different metaorders, their order signs are uncorrelated, i.e., $\mathbb{E}\left[\epsilon_t \, \epsilon_{t+\tau} | O_t \neq O_{t+\tau}\right] = 0$. On the other hand, we also have $\mathbb{E}\left[\epsilon_t \, \epsilon_{t+\tau} | O_t = O_{t+\tau}, E_{t,t+\tau}\right] = 1$, where $E_{t,t+\tau} := \{V_{O_t}(t) \geq \cdots \geq V_{O_t}(t+\tau)\}$, i.e., the event that the meta order O_t never die out within the time interval $[t, t+\tau]$.
- The probability that a child order drawn at random comes from a metaorder of length L is proportional to $Lp(L)$ where $p(L)$ is the probability that a metaorder has length L.

Note that we can rewrite $\rho(\tau)$ as

$$\begin{aligned}\rho(\tau) &= \mathbb{E}\left[\epsilon_t \epsilon_{t+\tau}\right] \\ &= \mathbb{E}\left[\epsilon_t \epsilon_{t+\tau} | O_t = O_{t+\tau}, E_{t,t+\tau}\right] \mathbb{P}\left[O_t = O_{t+\tau}, E_{t,t+\tau}\right] + \mathbb{E}\left[\epsilon_t \epsilon_{t+\tau} | O_t = O_{t+\tau}, E^c_{t,t+\tau}\right] \mathbb{P}\left[O_t = O_{t+\tau}, E^c_{t,t+\tau}\right] \\ &\quad + \mathbb{E}\left[\epsilon_t \epsilon_{t+\tau} | O_t \neq O_{t+\tau}\right] \mathbb{P}\left[O_t \neq O_{t+\tau}\right] \\ &= \mathbb{P}\left[O_t = O_{t+\tau}, E_{t,t+\tau}\right] \\ &= \sum_L \mathbb{P}\left[O_t = O_{t+\tau}, E_{t,t+\tau} | V_{O_t}(t^*) = L\right] \mathbb{P}\left[V_{O_t}(t^*) = L\right],\end{aligned}$$

where t^* denotes the first time that the meta order O_t was generated.

Note

We shall be following the original paper by using the notations

$$\begin{aligned}Q(L) &:= \mathbb{P}\left[V_{O_t}(t^*) = L\right] \propto Lp(L), \\ q(\tau|L) &:= \mathbb{P}\left[O_t = O_{t+\tau}, E_{t,t+\tau} | V_{O_t}(t^*) = L\right].\end{aligned}$$

In other words, $Q(L)$ is the probability that the executed child order coming from a meta order of length L and $q(\tau|L)$ is the probability that two child orders τ apart come from the same metaorder of length L. Thus,

$$\rho(\tau) = \sum_L Q(L)\, q(\tau|L).$$

We calculate $q(\tau|L)$ as follows.

$$\begin{aligned}q(\tau|L) &= \mathbb{P}\left[O_t = O_{t+\tau}, E_{t,t+\tau} | V_{O_t}(t^*) = L\right] \\ &= \mathbb{P}\left[O_t = O_{t+\tau} | E_{t,t+\tau}, V_{O_t}(t^*) = L\right] \mathbb{P}\left[E_{t,t+\tau} | V_{O_t}(t^*) = L\right] \\ &= \frac{1}{N} \mathbb{P}\left[E_{t,t+\tau} | V_{O_t}(t^*) = L\right]\end{aligned}$$

Denote by $s_{t,t+\tau} := \sum_{s=t+1}^{t+\tau} \mathbf{1}_{\{O_s = O_t\}}$. In other words, $s_{t,t+\tau}$ counts the number to times that the meta order O_t has been chosen to execute within the interval $[t+1, t+\tau]$. Note that we have

$$\begin{aligned}&\mathbb{P}\left[E_{t,t+\tau} | V_{O_t}(t^*) = L\right] \\ &= \mathbb{P}\left[s_{t,t+\tau} < V_{O_t}(t) | V_{O_t}(t^*) = L\right] \\ &= \sum_{\ell=1}^{L} \mathbb{P}\left[s_{t,t+\tau} < V_{O_t}(t) | V_{O_t}(t) = \ell, V_{O_t}(t^*) = L\right] \mathbb{P}\left[V_{O_t}(t) = \ell | V_{O_t}(t^*) = L\right] \\ &= \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{P}\left[s_{t,t+\tau} < \ell | V_{O_t}(t) = \ell, V_{O_t}(t^*) = L\right].\end{aligned}$$

Moreover, we have

$$\mathbb{P}\left[s_{t,t+\tau} < \ell | V_{O_t}(t) = \ell, V_{O_t}(t^*) = L\right] = \sum_{k=0}^{\ell-1} \binom{\tau-1}{k} p^k (1-p)^{\tau-1-k}$$

where $p = \frac{1}{N}$, since $s_{t,t+\tau} \sim B(\tau - 1, p)$.

Computation of $\rho(\tau)$

We have (3)

$$\begin{aligned}
\rho(\tau) &= \sum_L q(\tau|L)Q(L) \\
&= \sum_L \frac{p}{L} Q(L) \sum_{\ell=1}^{L-1} \sum_{k=0}^{\ell} \binom{\tau-1}{k} p^k (1-p)^{\tau-1-k} \\
&\approx \sum_L \frac{p}{L} Q(L) \sum_{\ell=1}^{L-1} N\left(\frac{\ell - (\tau-1)p}{\sqrt{(\tau-1)p(1-p)}} \right) \quad \text{(by CLT)} \\
&\approx \sum_L \frac{p}{L} Q(L) \int_{3/2}^{L-1/2} N\left(\frac{\ell - (\tau-1)p}{\sqrt{(\tau-1)p(1-p)}} \right) d\ell \\
&\approx \int_{3/2}^{\infty} \frac{p}{L} Q(L) \int_{3/2}^{L-1/2} N\left(\frac{\ell - (\tau-1)p}{\sqrt{(\tau-1)p(1-p)}} \right) d\ell dL,
\end{aligned}$$

where $N(\cdot)$ denotes the cdf for standard normal.

The power-law case

In the realistic case where metaorder sizes L are power-law distributed with $p(L) = \gamma L^{-(1+\gamma)}$, which implies $Q(L) \propto L^{-\gamma}$, with $\gamma > 1$, performing the integration (3) explicitly, but technical and tedious, gives as $\tau \to \infty$

$$\rho(\tau) \sim N^{\gamma-2}\, \tau^{1-\gamma}.$$

In particular, if $\gamma = 3/2$ as is more or less the case empirically for many stocks, we have

$$\rho(\tau) \sim \frac{1}{\sqrt{\tau}} \text{ for large } \tau.$$

- The LMF model gives a link between the distribution of order sizes and the autocorrelation function of order signs.

Empirical confirmation

- [Tóth, Lillo et al.][8] perform a careful analysis of order flow data from the London Stock Exchange containing exchange membership identifiers.
- They conclude that order splitting is indeed the dominant cause of the long memory of the order sign process.

Evidence based on membership codes

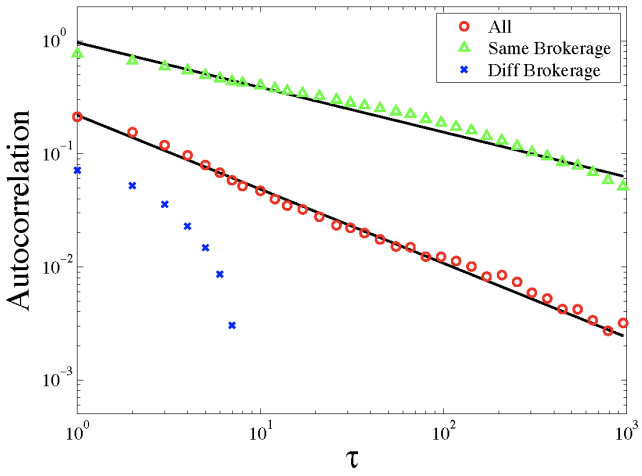


Figure 2. Autocorrelation of signs vs. transaction lag for transactions with same membership code, different membership code, and all transactions irrespective of membership code, plotted on double logarithmic scale. The investigated stock is AstraZeneca (AZN) traded at LSE in the period 2000-2002.

Evidence based on membership codes

- Assuming meta orders are executed using only a few brokers, we would expect to see more autocorrelation for a given broker order flow than for order flow in aggregate.
- Figure 2 from [Bouchaud, Farmer, Lillo][2] shows just that and is a strong evidence for order splitting being the dominant cause of long memory in order flow.

More evidence for order splitting

[Tóth, Lillo et al.][8] present even more convincing evidence for order-splitting as opposed to herding as the principal explanation for the long memory of order flow.

They decompose the sample autocorrelations as follows:

- $\epsilon_t^i = 1$ means a buy order was placed by investor i
- $\epsilon_t^i = 0$ means an order placed by another investor
- $\epsilon_t^i = -1$ means a sell order placed by investor i.

Decomposition of autocorrelation function

Assuming the long-term average order sign is zero, we have

$$C(\tau) = \langle \epsilon_t \, \epsilon_{t+\tau} \rangle = \frac{1}{N} \sum_t \epsilon_t \, \epsilon_{t+\tau} = \frac{1}{N} \sum_t \sum_{i,j} \epsilon_t^i \, \epsilon_{t+\tau}^j$$

Define the autocorrelation between i and j orders as

$$C^{ij}(\tau) = \langle \epsilon_t^i \, \epsilon_{t+\tau}^j \rangle = \frac{1}{N^{ij}(\tau)} \sum_t \epsilon_t^i \, \epsilon_{t+\tau}^j$$

and let $\theta^{ij}(\tau) = N^{ij}(\tau)/N$ be the fraction of times that an order from investor i at time t is followed by an order from investor j at time $t + \tau$.

Then

$$\begin{aligned} C(\tau) &= \sum_{i,j} \theta^{ij}(\tau) \, C^{ij}(\tau) \\ &= \sum_i \theta^{ii}(\tau) \, C^{ii}(\tau) + \sum_{i \neq j} \theta^{ij}(\tau) \, C^{ij}(\tau) \\ &=: C_{split}(\tau) + C_{herd}(\tau). \end{aligned}$$

Decomposition in pictures

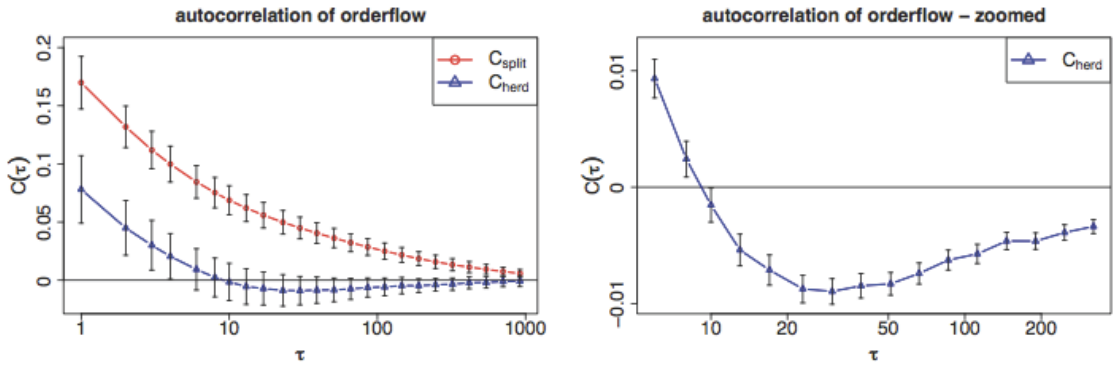


Figure 3: An illustration of the consistency with which splitting dominates herding. *Left panel:* The autocorrelations of market order signs averaged across all 102 samples (spanning nine years and six stocks). The bars are standard deviations. *Right panel:* The herding component in the left panel is magnified to better observe the anti-herding effect. The bars in this panel are standard errors. For both plots we use logarithmic scale on the horizontal axis and linear scale on the vertical axis.

Empirical observation

- Order splitting dominates herding
- There is *anti-herding*!
  - Other market participants tend to trade against a split meta order.

Distribution of volume is fat-tailed

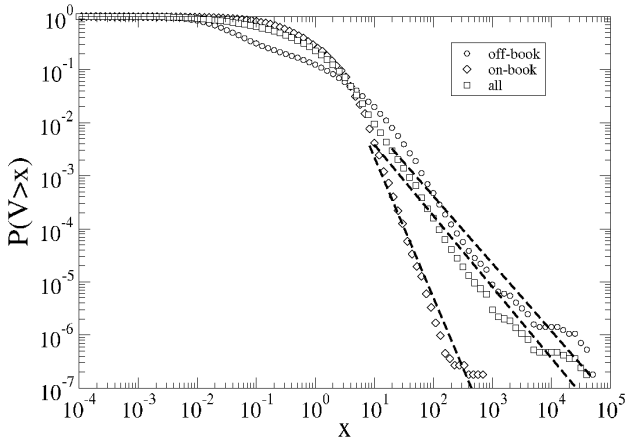


Figure 3. Volume distributions of off-book trades (circles), on-book trades (diamonds), and the aggregate of both (squares). We show this for a collection of 20 different stocks, normalizing the volume of each by the mean volume before combining. The dashed black lines have the slope found by the Hill estimator and are shown for the largest one percent of the data. Adapted from Lillo et al. [2005].

Note

- In the LSE, orders in the *on-book* market are placed publicly but anonymously and execution is completely automated.
- The *off-book* market operators through a bilateral exchange mechanism via phone calls or direct contact of the trading parties.

The Hill estimator

- Denote by $\tilde{X}_i$ the k exceedances in the sample $\{X_1, X_2, \ldots, X_n\}$ over some threshold u.
- The Hill estimator (of $1/\alpha$) is then

$$H_k = \frac{1}{k} \sum_{i=1}^{k} \log \frac{\tilde{X}_i}{u}$$

- To see how this works, consider a density of the form

$$f(x) = \frac{\alpha}{u}\left(\frac{u}{x}\right)^{\alpha+1}.$$

  which corresponds to a distribution with tail-exponent α. Then

$$\mathbb{E}\left[\log\frac{X}{u}\right] = \int_u^\infty \log\frac{x}{u} f(x)\,dx = \frac{1}{\alpha}.$$

- The Hill estimator is also the maximum likelihood estimator (MLE) for the power law with pdf f.
- The Hill estimator works reasonably well when the tail is really Pareto. It can give bad results if the underlying distribution does not have a Pareto tail.

Distribution of volume

- The distribution of block trades appears to have a tail with exponent $\alpha \approx 3/2$.
- Since block trading is in competition with order-splitting as a strategy for executing large trades, we assume that the distribution of meta orders should also have a roughly $3/2$ tail.
  - This has been confirmed by [Bershova and Rakhlin][1] using a database of Alliance Bernstein proprietary meta orders.

Notions of market impact

The term *market impact* can refer to many different phenomena.

Market impact can refer to:

- The impact of individual transactions
- The impact of meta orders
- The impact of aggregate order flow over a given period of time.

Market impact of market orders: Empirical results

- There are many studies in the classic microstructure literature that document that market impact is a concave function of volume. Keim and Madhavan find that block trades also generate concave market impact.
- On the NYSE, Lillo et al. find that

$$\mathbb{E}[\Delta P|V] \sim V^{\psi}$$

  with $\psi \approx 0.5$ for small V and $\psi \approx 0.2$ for large V.

- On the LSE, Lillo et al. find $\psi \approx 0.3$ on average.
- Potters and Bouchaud study trades on NASDAQ and the Paris Stock Exchange (PSE) finding $\mathbb{E}[\Delta P|V] \sim \log V$.

$\psi \approx 0.3$ **on LSE from Lillo et al.**

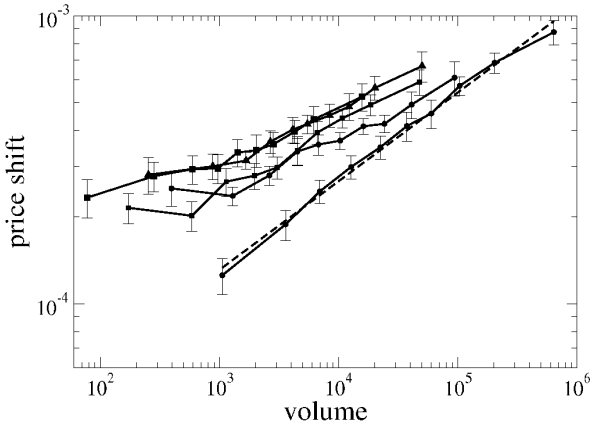


Figure 4. Market impact function of buy market orders for a set of 5 highly capitalized stocks traded in the LSE, specifically AZN (filled squares), DGE (empty squares), LLOY (triangles), SHEL (filled circles), and VOD (empty circles). Trades of different sizes are binned together, and the average size of the logarithmic price change for each bin is shown on the vertical axis. The dashed line is the best fit of the market impact of VOD with a functional form described in Eq.5.1. The value of the fitted exponent for VOD is $\psi = 0.3$.

The square-root formula

- Practitioners typically believe in a square-root relationship between market impact and volume.
- Specifically, the relationship between the price impact ΔP and the size Q of the order should be something like

$$\Delta P \approx \sigma \sqrt{\frac{Q}{V}}$$

where σ is (for example) daily (dollar) volatility and V is daily volume.
  - Note that this formula is dimensionally correct.

- The square root law has recently been verified for Bitcoin and even for vega in options markets.

A heuristic derivation of the square-root market impact formula

- Suppose each trade impacts the mid-log-price of the stock by an amount proportional to $\sqrt{n_i}$ where n_i is the size of the ith trade.

- Then the change in mid-price over one day is given by

$$\Delta P = \sum_i^N \eta \, \epsilon_i \, \sqrt{n_i}$$

where η is the coefficient of market impact, ϵ_i is the sign of the ith trade and N is the (random) number of trades in a day.
  - Note that both the number of trades (N) and the size of each trade (n_i) in a given time interval are random.

- If N, ϵ_i and n_i are all independent, the variance of the one-day price change is given by

\begin{eqnarray} \sigma^2:=\mbox{Var}(\Delta P)=\eta^2\,\sum_i n_i=\eta^2\,V, \end{eqnarray}

where V is the average daily volume.

- It follows that

$$|\Delta P_i| = \eta \sqrt{n_i} = \sigma \sqrt{\frac{n_i}{V}}$$

which is the familiar square-root market impact formula.

Why $\sqrt{n}$?

An inventory risk argument:

- A market maker requires an excess return proportional to the risk of holding inventory.
- Risk is proportional to $\sigma \sqrt{T}$ where T is the holding period.
- The holding period should be proportional to the size of the position.
- So the required excess return must be proportional to $\sqrt{n}$.

Gabaix again

- Volume follows the $3/2$ law:

$$\mathbb{P}\left[\text{Volume} > V\right] \sim \frac{1}{V^{3/2}}$$

- Market impact is proportional to the square-root of volume

$$\Delta P \sim \sqrt{V}$$

- Then

$$\mathbb{P}\left[\Delta P > x\right] \sim \frac{1}{x^3},$$

the so-called cubic law of returns.

Aggregate market impact

- When practitioners are questioned as to whether the square-root model is supposed to estimate the impact of individual orders, meta orders or aggregate order flow, the answer is usually “all of these!”.

- [Hopman][5] measures aggregate order flow on the PSE as $\sum_i \epsilon_i V_i^{\psi}$.
  - He finds that $\psi = 1/2$ best explains daily returns. $\psi = 1$ gives a much lower R^2.
  - He also finds that causality runs from orders to prices and not the other way round, supporting the econophysicists and contradicting the informed trader story.
    - Returns follow orders with a lag.
- Gabaix et al. find $\psi \approx 1/2$ on New York, London and Paris markets.

Aggregation

- For a sequence of N successive trades, let $Q_N := \sum_i^N \epsilon_i V_i$ be the aggregate quantity (net volume) and $R_N := \sum_i^N \log P_i/P_{i-1} = \log P_N - \log P_0$ be the aggregate return.
- Figure 5 shows how the aggregate return $R(Q, N)$ scales with quantity Q_N and N.
- According to [Bouchaud, Farmer, Lillo][2], aggregate impact becomes increasingly linear with increasing N.

Scaling of aggregated returns

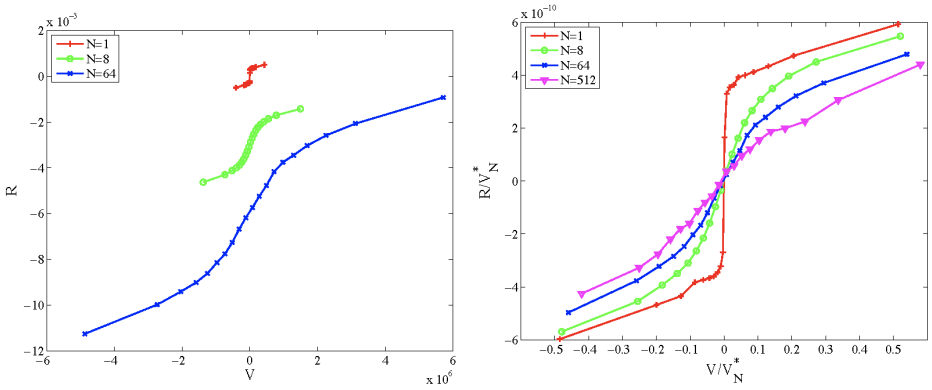


Figure 5. Aggregate market impact $R(Q, N)$ for the LSE stock Astrazeneca for 2000-2002. In (a) we plot the shifted aggregate return $R(Q, N) + R_0$ vs. the aggregate signed volume Q for three values of N. The arbitrary constant R_0 is added to aid visualization; its values are $R_0 = \{0, -3 \times 10^{-3}, -6 \times 10^{-3}\}$ for $N = 1, 8$ and 64 respectively. In (b) for each N we rescale both the horizontal and vertical axes by $Q_N^* = Q_N^{(95)} - Q_N^{(5)}$, where $Q_N^{(5)}$ is the 5% quantile and $Q_N^{(95)}$ is the 95% quantile of Q.

Why is market impact concave?

There are three types of answer in the literature

- The informativeness of a trade depends on size (Easley and O'Hara).
  - Small trades may carry almost as much information as large trades.
- The shape of the order book (Weber and Rosenow).
  - The cumulative depth available at a given price level determines price impact.
- Selective liquidity taking.
  - Traders condition their orders on the quantity available in the order book.

Virtual market impact

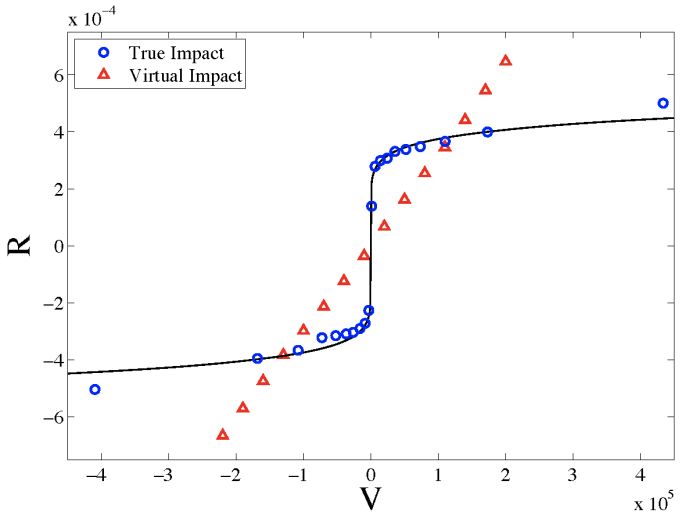


Figure 6. Comparison of virtual to true market impact. True impact is shown in blue circles, virtual impact in red triangles. The fitted curve for true impact (solid black) is of the form $f(v) = Av^{\psi}$, with $\psi = 0.3$.

Selective liquidity taking

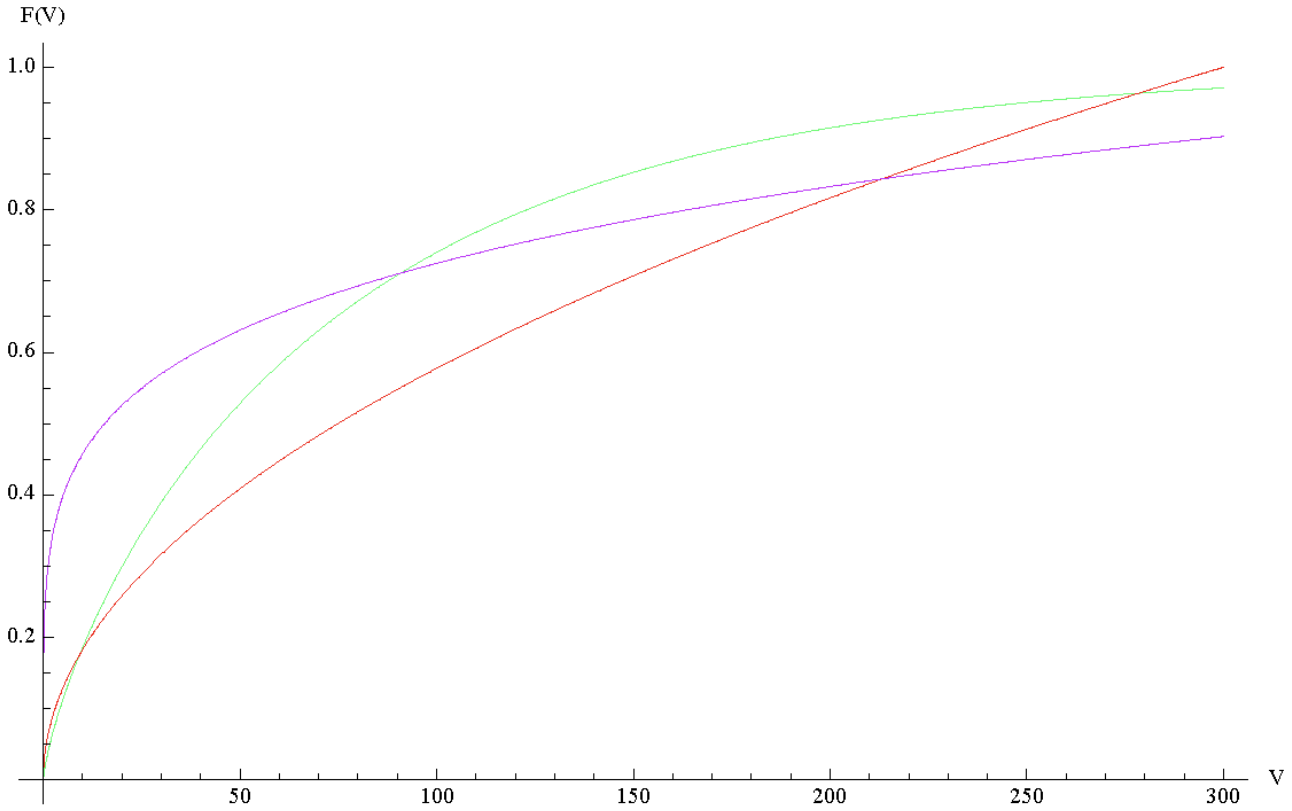
- Empirically, nearly all trades are of sizes less than or equal to the size available at the best quote, and certainly smaller than the total size available at the first two quote levels.
- Suppose then that the price changes only in increments of the spread s. Then
$$\mathbb{E}[\Delta P|V] = \Pr(\Delta P > 0|V)\, s = \Pr(V \geq Q)\, s$$
where Q is the quantity available at the best quote.
- But
$$\Pr(V \geq Q) = \Pr(Q \leq V) =: F_Q(V),$$
the cumulative distribution function of quantity at the best quote, which we expect to be concave wrt V.

Gamma-distributed volume at best quote

For example, according to Bouchaud, Mézard and Potters, the volume at the best quote is gamma-distributed
$$f(V) = \frac{1}{V_0^\gamma \Gamma(\gamma)} V^{\gamma-1}\, e^{-V/V_0}$$
with $\gamma \approx 0.75$.

The cumulative distribution would then be
$$F(V) = 1 - \frac{\Gamma\left(\gamma, \frac{V}{V_0}\right)}{\Gamma(\gamma)},$$
where
$$\Gamma(\gamma, x) := \int_x^\infty t^{\gamma-1} e^{-t} dt$$
denotes the upper incomplete Gamma function

Gamma distributed volume at best quote with $\gamma = 0.75$



The green line is $F(V)$; the red line is $V^{0.5}$; the violet line is $V^{0.2}$.

A fixed permanent impact model

- Suppose that changes in the mid-quote m depend on the sign ϵ of the order, the order size v and possibly on other state variables Ω such as the state of the order book:
$$\Delta m_t = m_t - m_{t-1} = \epsilon_t f(v_t, \Omega_t) + \eta_t$$
with η iid and satisfying $\mathbb{E}[\eta_t] = 0$, $\mathrm{Var}[\eta_t] = \sigma^2$.
- Integrating this gives $$m_{t}=\sum_{k

Variance and autocovariance of price returns

- If $f(\cdot)$ is independent of order-flow, it follows that
$$\mathbb{E}[\Delta m_t \, \Delta m_{t+\tau}] = \mathbb{E}[\epsilon_t \, \epsilon_{t+\tau} f(v_t) f(v_{t+\tau})] \propto \tau^{-\alpha}$$
so that price returns are autocorrelated in time.
- Also, in that case, the variance of lag-τ returns is given by
$$\mathbb{E}[(m_{t+\tau} - m_t)^2] \propto \sum_{i,j=1}^{\tau} \mathbb{E}\left[\epsilon_i \, \epsilon_j\right] f(v_i) f(v_j) \propto \tau^{2-\alpha}.$$
- Both of these properties of the model are inconsistent with efficient prices as empirically observed.

- It follows that either the market impact function $f(\cdot)$ must depend on order flow or that market impact is not permanent.
  - The former point of view was promoted by Farmer, Lillo et al.
  - The latter point of view was promoted by Bouchaud et al.
- We will see that, somewhat surprisingly, the two points of view are compatible with each other.

The Madhavan Richardson and Roomans (MRR) model again

- In the MRR model, as in Glosten-Milgrom, the revision in beliefs is positively correlated with the innovation in the order flow:

  (4)
  $$\Delta V_t = \lambda \, \left(\epsilon_t - \mathbb{E}[\epsilon_t | \mathcal{F}_{t-1}] \right) + e_t$$

  where V_t is the efficient price and e_t represents for example news.

- However, in the MRR model, the revision in beliefs depends only on the unexpected component of order flow.
- MRR modeled the order flow as $AR(1)$: $\epsilon_t = \phi_1 \epsilon_{t-1} + \eta_t$.
- We now model order flow as a long-memory process.

Expected trade sign

- In the MRR model,
  $$\mathbb{E}[\epsilon_t | \epsilon_{t-1}] = \phi_1 \, \epsilon_{t-1}.$$
  where ϕ_1 is the first order autocorrelation coefficient.
- In our extension of this model, we write
  $$\hat{\epsilon}_t = \mathbb{E}[\epsilon_t | \mathcal{F}_{t-1}] = \sum_{k=1}^{\infty} \phi_k \, \epsilon_{t-k}$$
  so the filtration $\mathcal{F}_t$ is now not just the last order sign ϵ_{t-1} but the entire order flow history.

- In practice, we can fit an $AR(p)$ model with some large order p:
  $$\hat{\epsilon}_t = \mathbb{E}[\epsilon_t | \mathcal{F}_t] = \sum_{k=1}^{p} \phi_k \, \epsilon_{t-k}.$$
- We expect the coefficients to decay as a power-law:
  $$\phi_k \sim k^{-\beta} \text{ for some } \beta.$$

Empirical verification of the model

- Approximating ΔV by Δm, we may rewrite (4) as

  (5)

$$\Delta m_t = \lambda \left(\epsilon_t - \hat{\epsilon}_t\right) + \eta_t$$

- In this model, the impact of a transaction depends on the entire history of the order flow and how predictable the order sign of the transaction is.
  - The most likely outcome has the smallest impact.
- In Figure 10, we see stunning empirical verification of this model.

Market impact relates to unexpected order flow

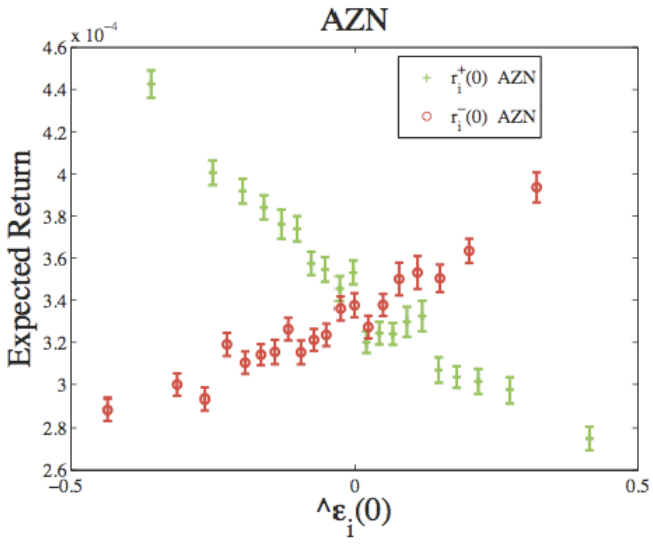


Figure 10. The expected return as a function of the sign predictor $\hat{\epsilon}$. The quantity r^+ (r^-) refer to trades with a sign that is equal (opposite) to the one of the predictor. The data are binned in such a way that each point contains an equal number of observations. Error bars are standard errors. Adapted from Gerig [2007].

Bouchaud's power-law decay argument

- As before, assume that over one day

$$\Delta P = \sum_i^N \eta\, \epsilon_i\, \sqrt{n_i}$$

- The previous heuristic proof of the square-root model assumed that $\mathrm{Cov}[\epsilon_i, \epsilon_j] = 0$ if $i \neq j$ and that all market impact is permanent.
- Empirically, we found that autocorrelation of trade signs shows power-law decay with a small exponent α (very slow decay).

Bouchaud's power-law decay argument continued

$$\begin{aligned}
\mathrm{Var}[\Delta P] &= \eta^2\, \mathrm{Var}\left[\sum_i^N \epsilon_i\, \sqrt{n_i}\right] \\
&= \eta^2\, \left\{ N\, \mathrm{Var}[\sqrt{n_i}] + \sum_{i \neq j} \mathrm{Cov}[\epsilon_i \sqrt{n_i}, \epsilon_j \sqrt{n_j}] \right\} \\
&\approx \eta^2\, \left\{ N\, \mathrm{Var}[\sqrt{n_i}] + \frac{2\, C_1}{(2-\alpha)\,(1-\alpha)}\, \mathbb{E}[\sqrt{n}]^2\, N^{2-\alpha} \right\} \\
&\sim N^{2-\alpha} \text{ as } N \to \infty
\end{aligned}$$

- Empirically, we find that, to a very good approximation, $\mathrm{Var}[\Delta P] \propto N$.
  - Otherwise returns would be serially correlated.
- These observations may be reconciled if market impact decays as a power law.

Computation of daily variance with power-law decay

- Assuming market impact decays as $1/\tau^\gamma$, i.e.,

$$\Delta P = \eta \sum_i^N \frac{\epsilon_i \sqrt{n_i}}{(N-i)^\gamma},$$

then we have

$$\begin{aligned}
\mathrm{Var}[\Delta P] &= \eta^2\, \mathrm{Var}\left[\sum_i^N \frac{\epsilon_i\, \sqrt{n_i}}{(N-i)^\gamma}\right] \\
&= \eta^2\, \left\{ \sum_i^{N-1} \frac{\mathbb{E}[n]}{(N-i)^{2\,\gamma}} \right. \\
&\qquad \left. + 2\, C_1 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N-1} \frac{\mathbb{E}[\sqrt{n}]^2}{(N-i)^\gamma\, (N-j)^\gamma\, (j-i)^\alpha} \right\} \\
&\sim N^{2-\alpha-2\,\gamma} \text{ as } N \to \infty \\
&\sim N \text{ only if } \gamma \approx (1-\alpha)/2.
\end{aligned}$$

Equivalence of the two formulations

We have in the Lillo picture that

(6)

$$\Delta m_t = m_t - m_{t-1} = \lambda \left(\epsilon_t - \hat{\epsilon}_t\right) + \eta_t$$

with

$$\hat{\epsilon}_t = \sum_{k=1}^{\infty} \phi_k \, \epsilon_{t-k}.$$

Write (Bouchaud picture)

$$m_t = \lambda \sum_{i=0}^{\infty} G(i) \, \epsilon_{t-i} + \sum_{j \le t} \eta_j$$

Then

(7) \begin{eqnarray} \Delta m_t = mt - m{t-1}&=&\lambda\,\left(\sum{k=0}^\infty\,G(k)\,\epsilon{t-k}-\sum{k=0}^\infty\,G(k)\,\epsilon{t-1-k}\right)+\eta_t\nonumber\ &=&\lambda\,\epsilont+\lambda\,\sum{k=1}^\infty\,\left{G(k)-G(k-1)\right)}\,\epsilon_{t-k}+\eta_t \end{eqnarray}

To match the expressions (6) and (7), we need

$$G(k) - G(k-1) = -\phi_k.$$

One choice that would work is:

$$G(k) = \sum_{j=k+1}^{\infty} \phi_j$$

If the autocorrelation function $\rho(\tau) \sim \tau^{-\alpha}$ for large τ, and if the underlying process for order signs is ARFIMA, the best linear predictor of order sign is given by

$$\hat{\epsilon}_t = \sum_{k=1}^{\infty} \phi_k \, \epsilon_{t-k}$$

with $\phi_k \sim k^{-(1+\beta)}$ for large k and

$$\beta = \frac{1-\alpha}{2}.$$

Then, for large k,

$$G(k) = \sum_{j=k+1}^{\infty} \phi_j \sim \frac{1}{k^{\beta}}$$

with

$$\beta = \frac{1-\alpha}{2}.$$

That is, the propagator $G(\tau)$ decays as $\tau^{-\gamma}$ with

$$\gamma = \frac{1-\alpha}{2}.$$

as required to show equivalence between the Bouchaud and Lillo pictures.

Explicit computation (Tai-Ho Wang)

Suppose the order sign process is $ARFIMA(0, d, 0)$. Then, by definition,

$$(1 - B)^d \epsilon_t = \eta_t$$

where $\eta_t \sim N(0, 1)$ say.

Inverting this expression gives

(8)

$$\epsilon_t = \eta_t + \sum_{j=1}^{\infty} \psi_j \, \eta_{t-j} = \sum_{j=0}^{\infty} \psi_j \, \eta_{t-j}$$

where we adopt the convention that $\psi_0 = 1$ and for $j > 0$,

$$\psi_j = (-1)^j \begin{pmatrix} -d \\ j \end{pmatrix} = \frac{d\,(d+1)\ldots(d+j-1)}{j!} = \begin{pmatrix} j+d-1 \\ j \end{pmatrix}$$

The autocovariance function of ϵ_t is then

(9) \begin{eqnarray} \rho(k)&=&\sum_{j=0}^\infty\,\psi_j\,\psi_{j+k}\nonumber\ &=&\frac{\Gamma(1-2\,d)}{{\Gamma(1-d)}^2} \,\frac{d\,(1+d)...\,(k-1+d)}{(1-d)\,(2-d)\,...\,(k-d)}\nonumber\ &\sim&\frac{\Gamma(1-2\,d)}{\Gamma(d)\,\Gamma(1-d)}\,\frac{1} {k^{1-2\,d}} ~ \text{ as } k \to \infty \end{eqnarray}

In our notation, $\rho(\tau) \sim \tau^{-\alpha}$ so $d = \frac{1-\alpha}{2}$.

Now invert (8) to get

$$\epsilon_t = \eta_t + \sum_{j=1}^{\infty} (-1)^{j-1} \begin{pmatrix} d \\ j \end{pmatrix} \epsilon_{t-j}.$$

Then

$$\hat{\epsilon}_t = \mathbb{E}[\epsilon_t | \mathcal{F}_{t-1}] = \sum_{k=1}^{p} \phi_k \, \epsilon_{t-k}. \text{ with } \phi_k = (-1)^{j-1} \begin{pmatrix} d \\ j \end{pmatrix}.$$

Finally

$$G(k) = \sum_{j=k+1}^{\infty} \phi_j = \begin{pmatrix} k-d \\ k \end{pmatrix} \sim \frac{1}{\Gamma(1-d)\, k^d} \text{ as } k \to \infty.$$

Thus, the exponent γ of decay of market impact is given by

$$\gamma = d = \frac{1-\alpha}{2}.$$

The Bouchaud and Lillo formulations are exactly equivalent in this case.

When is TIM equivalent to HDIM?

- [Taranto et al.][8] show that the Bouchaud transient impact (TIM) picture is equivalent to the Lillo-Farmer history dependent impact (HDIM) picture whenever the time series of order signs is generated by a so-called Discrete Autoregressive (DAR) process.

- A DAR process may be constructed as follows:
  - For each t, choose a distance $\ell > 0$ with probability λ_ℓ from a distribution with $\sum_\ell \lambda_\ell = 1$.
  - Then

$$\epsilon_t = \begin{cases} \epsilon_{t-\ell} & \text{with probability } \rho \\ -\epsilon_{t-\ell} & \text{with probability } 1-\rho \end{cases}$$

Including other events

Start with the Bouchaud transient impact (TIM) picture again:
$$p_t=\lambda\,\sum_{sThe *propagator* or *decay kernel* $G(\cdot)$ may be estimated from the correlation function

$$\begin{aligned} \mathcal{R}(\ell) &= \langle (p_{t+\ell} - p_t)\, \epsilon_t \rangle \\ &= \sum_{0<n\le\ell} G(n)\, C(\ell - n) + \sum_{n>0} [G(n+\ell) - G(n)]\; C(n). \end{aligned}$$

Then, the second moment of the price difference (variance) may be computed as

$$\mathcal{D}(\ell) = \langle (p_{t+\ell} - p_t)^2 \rangle.$$

Bouchaud transient impact picture vs observation

[Eisler, Bouchaud and Kockelkoren][4] show that the Bouchaud transient impact (TIM) picture is incomplete.

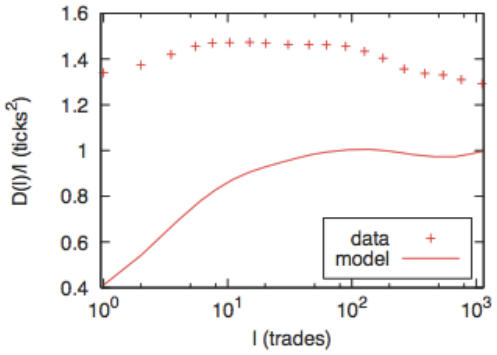


Figure 1: $D(\ell)/\ell$ and its approximation with the transient impact model (TIM) with only trades as events, with $\eta_t = 0$ and for small tick stocks. Results are shown when assuming that all trades have the same, non fluctuating impact $G(\ell)$, calibrated to reproduce $\mathcal{R}(\ell)$. This simple model accounts for $\sim 2/3$ of the long term volatility. Other events and/or the fluctuations of $G(\ell)$ must therefore contribute to the market volatility as well.

Interpretation

The Bouchaud picture is too simplistic

- Market impact varies wildly over time according to the state of the market (for example the shape of the order book)
  - The history of order flow is not the only determinant of market impact
- Not only market orders impact the market price; limit orders and cancelations also impact the price.

Summary

- Order flow is a long memory process.
  - The dominant effect is order-splitting.
- Market impact is concave due to selective liquidity taking.
- Market impact of market orders can be modeled as:
  - Permanent but state-dependent (Lillo)
  - Transient (Bouchaud)
- Both of these formulations are equivalent.

- To get quantitative (as opposed to qualitative) agreement with observation, in principle we need to take into account
  - Time-varying liquidity
  - Limit orders and cancelations
- In practice, it seems (see [Taranto et al.][8]) that distinguishing between market orders that change the price and orders that result in no price change is enough for a surprisingly accurate description of market impact.

References

1. ^Nataliya Bershova and Dmitry Rakhlin, The Non-Linear Market Impact of Large Trades: Evidence from Buy-Side Order Flow, *Quantitative Finance* **13**(11) 1759–1778 (2013).
2. ^Jean-Philippe Bouchaud, J. Doyne Farmer, and Fabrizio Lillo, How Markets Slowly Digest Changes in Supply and Demand, in *Handbook of Financial Markets: Dynamics and Evolution* 57–156. (2009) available at http://tuvalu.santafe.edu/%7Ejdf/papers/MarketsSlowlyDigest.pdf: Sections 4, 5 and 6.
3. ^Jean-Philippe Bouchaud, Yuval Gefen, Marc Potters and Matthieu Wyart, Fluctuations and response in financial markets: the subtle nature of random price changes, *Quantitative Finance* **4**(2) 176--190 (2004).
4. ^Zoltán Eisler, Jean-Philippe Bouchaud, and Julien Kockelkoren, The price impact of order book events: market orders, limit orders and cancellations, *Quantitative Finance* **12**(9) 1395-1419 (2012).
5. ^Carl Hopman, Do supply and demand drive stock prices?, *Quantitative Finance* **7**(1) 37–53 (2007).
6. ^Fabrizio Lillo, Szabolcs Mike, and J Doyne Farmer, Theory for long memory in supply and demand, *Phys. Rev. E* **71**(6) 66122 (2005).
7. ^Gennady Samorodnitsky, Murad S. Taqqu, *Stable non-Gaussian random processes: stochastic models with infinite variance*, Chapman and Hall (1994).
8. ^ Damian Eduardo Taranto, Giacomo Bormetti, Jean-Philippe Bouchaud, Fabrizio Lillo, and Bence Toth, Linear models for the impact of order flow on prices I. Propagators: Transient vs. History Dependent Impact, *Quantitative Finance* forthcoming (2018).
9. ^Bence Tóth, Imon Palit, Fabrizio Lillo, and J Doyne Farmer, Why is order flow so persistent?, *Journal of Economic Dynamics and Control* (51) 218-239 (2015).