10

Long-Range Persistence of Order Flow

An unfailing memory is not a very powerful incentive to the study of the phenomena of memory.

(Marcel Proust)

In the previous chapter, we noted that activity in financial markets tends to cluster in time. As we also noted, such clustering is not explained by local correlations of market activity, but instead suggests that market activity is a *long-memory process*. In this chapter, we discuss another type of long memory that is conceptually unrelated to this long memory of activity: the highly persistent nature of the sequence of binary variables ε_t that describe the direction of market orders. As we will see, buy orders tend to follow other buy orders and sell orders tend to follow other sell orders, both for very long periods of time.

More formally, let ε_t denote the sign of the t^{th} market order, with $\varepsilon_t = +1$ for a buy market order and $\varepsilon_t = -1$ for a sell market order, where t is discrete and counts the number of market orders. In this event-time framework, one can characterise the statistical properties of the time series of signs via the market-order **sign autocorrelation function**

$$C(\ell) := \mathrm{Cov}[\varepsilon_t, \varepsilon_{t+\ell}]. \tag{10.1}$$

As we will see in this chapter, the surprising empirical result is that $C(\ell)$ decays extremely slowly with ℓ, and is well approximated by a power-law $\ell^{-\gamma}$ with $\gamma < 1$. Importantly, this effect is different from activity clustering in time, which we considered in Chapter 9. For example, a process with exponentially distributed inter-arrival times – as would occur if arrivals are described by a homogeneous Poisson process – can still have long-range persistence in order signs. Conversely, a process in which order signs are uncorrelated could still have long-range autocorrelations in inter-arrival times. Therefore, the underlying mechanisms explaining these two phenomena could be completely different.[1]

In this chapter, we first review the empirical evidence for long-range persistence in ε_t. We then consider some consequences of this fact, including the apparent **efficiency paradox**, which asks the question of how prices can remain unpredictable when order flow (which impacts the price) is so predictable. We then introduce two models that are capable of producing ε_t series with long-range autocorrelations, and consider how to calibrate such models on empirical data to make predictions of future order signs. Finally, we will discuss the possible origins of the long-range autocorrelations that we observe empirically.

10.1 Empirical Evidence

The order-sign autocorrelation function defined by Equation (10.1) has been studied by many authors and on many different asset classes, including equities, FX and futures (see references in Section 10.7). While it has been known for a long time that market order signs are positively autocorrelated, it came as a surprise that these autocorrelations decay extremely slowly. Figure 10.1 shows $C(\ell)$ for our four stocks on NASDAQ. Consistently with the existing empirical studies of other assets, the figure suggests that $C(\ell)$ decays as a power-law $\ell^{-\gamma}$ with $\gamma < 1$, at least up to very large lags (beyond which statistical precision is lost). Mathematically, the value $\gamma < 1$ means that $C(\ell)$ is non-integrable, which means that the order-sign process is a **long-memory** process.[2] Note that the initial decrease of $C(\ell)$ is often faster (exponential-like), followed by a slow, power-law regime.

The fact that $C(\ell)$ decays so slowly has many important consequences. For a long-memory process, the conditional expectation of $\varepsilon_{t+\ell}$ given that $\varepsilon_t = 1$ is

$$\mathbb{E}[\varepsilon_{t+\ell}|\varepsilon_t = 1] = C(\ell) \sim \frac{c_\infty}{\ell^\gamma}. \tag{10.2}$$

Numerically, with the realistic values $c_\infty = 0.5$ and $\gamma = \frac{1}{2}$, this gives $C(10000) \approx 0.005$. Therefore, if we observe a buy market order now, the probability that a market order 10,000 trades in the future is a buy order exceeds that for a sell order by more than 0.5%. In modern equities markets for liquid assets, a time-lag of 10,000 trades corresponds to a few hours or even a few days of trading. Where does this long-range predictability come from? We will return to this question later in this chapter.

[1] For example, using a Hawkes process to generate the arrival times of orders whose signs are uncorrelated would not create long-range autocorrelations in the order-sign series.

[2] For a detailed introduction to long-memory processes, see, e.g., Beran, J. (1994). *Statistics for long-memory processes*. Chapman & Hall.

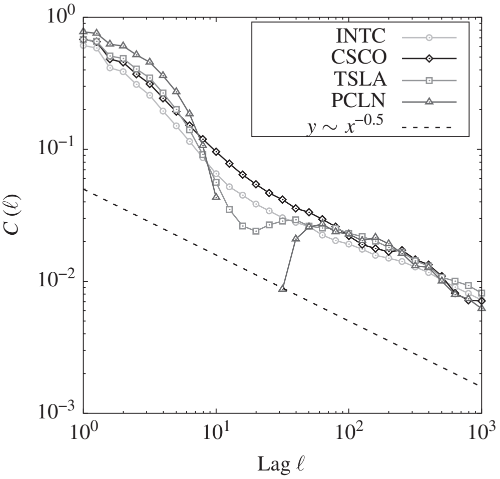

Figure 10.1. Autocorrelation function of the market order sign process for INTC, CSCO, TSLA and PCLN. The dashed line represents a power-law with exponent −0.5. The missing data points correspond to negative values.

10.2 Order Size and Aggressiveness

Until now, we have treated all market orders on an equal footing, independently of their volume. In theory, we could consider the time series of signed volumes $\varepsilon_t \upsilon_t$ (where υ_t denotes the volume of the market order arriving at time t). However, the distribution of order volumes has heavy tails, which complicates the estimation of the corresponding autocorrelation function. Furthermore, simply knowing that a market order is large only tells us half the story, because the impact caused by the arrival of a large buy (respectively, sell) market order depends on the volume available at the ask (respectively, bid) price.

A more useful idea is to label market orders with a simple binary **aggressiveness** variable $\pi_t \in \{0, 1\}$, which indicates whether or not the size of the market order is sufficiently large to trigger an immediate price change ($\pi_t = 1$) or not ($\pi_t = 0$). Put another way, a market order has $\pi_t = 1$ if and only if its size is greater than or equal to the volume available at the opposite-side best quote at its time of arrival.[3]

Throughout the book, we will write MO^1 to denote market orders for which $\pi_t = 1$ and MO^0 to denote market orders for which $\pi_t = 0$. We must then specify four correlation functions, depending on the type of events that we wish to consider. For clarity, we introduce an indicator variable $I(\pi_t = \pi)$, which is equal to 1 if the event at time t is of type π, and 0 otherwise. By standard properties of an indicator

[3] Note that in some cases, the emptied queue is then immediately refilled, reverting the initial price change. This is the case, for example, when hidden liquidity (e.g. iceberg orders) are present. Conversely, a market order smaller than the size of the queue can be immediately followed by a wave of cancellations, leading to a subsequent price change. Our definition is such that the first case corresponds to $\pi = 1$ and the second case to $\pi = 0$.

function, it follows that

$$\mathbb{P}(\pi) = \mathbb{E}[I(\pi_t = \pi)].$$

For a pair of events π and π' both in $\{\text{MO}^0, \text{MO}^1\}$, we define the **conditional correlation** of order signs as:

$$\begin{aligned} C_{\pi,\pi'}(\ell) &:= \mathbb{E}[\varepsilon_t \varepsilon_{t+\ell} | \pi_t = \pi, \pi_{t+\ell} = \pi'], \\ &= \frac{\mathbb{E}[\varepsilon_t I(\pi_t = \pi) \cdot \varepsilon_{t+\ell} I(\pi_{t+\ell} = \pi')]}{\mathbb{P}(\pi)\mathbb{P}(\pi')}. \end{aligned} \tag{10.3}$$

For $\ell > 0$, the first subscript of $C_{\pi,\pi'}(\ell)$ indicates the type of the event that happened first. The unconditional sign-correlation function is given by

$$C(\ell) = \sum_{\pi=0,1} \sum_{\pi'=0,1} \mathbb{P}(\pi)\mathbb{P}(\pi')C_{\pi,\pi'}(\ell).$$

By symmetry in the definition of C, it follows that

$$C_{\pi,\pi'}(\ell) = C_{\pi',\pi}(-\ell).$$

If $\pi \neq \pi'$, then in general $C_{\pi,\pi'}(\ell) \neq C_{\pi',\pi}(\ell)$. In other words, the correlation of the sign of an aggressive market order at time t with that of a non-aggressive market order at time $t + \ell$ has no reason to be equal to the corresponding correlation when the non-aggressive order arrives first.

Figure 10.2 shows the empirical values of these four conditional correlation functions. These results help to shed light on the origin of the slow decay of $C(\ell)$. For small-tick stocks, all four correlation functions behave similarly to each other. This makes sense, because an aggressive market order for a small-tick stock typically induces only a small relative price change, so there is little difference between market orders in MO^0 and those in MO^1 in this case. For large-tick stocks, by contrast, we observe significant differences between the behaviour for MO^0 and MO^1. Although the long-range power-law decay still occurs for pairs of market orders in MO^0, the sign of market orders in MO^1 is *negatively* correlated with that of the next market order. This also makes sense: after a substantial move up (respectively, down) induced by an aggressive buy (respectively, sell) market order, one expects that the flow inverts as sellers (respectively, buyers) are enticed by the new, more favourable price. Note finally that $C_{0,1}(\ell)$ decays faster than $C_{0,0}(\ell)$, but still approximately as a power-law.

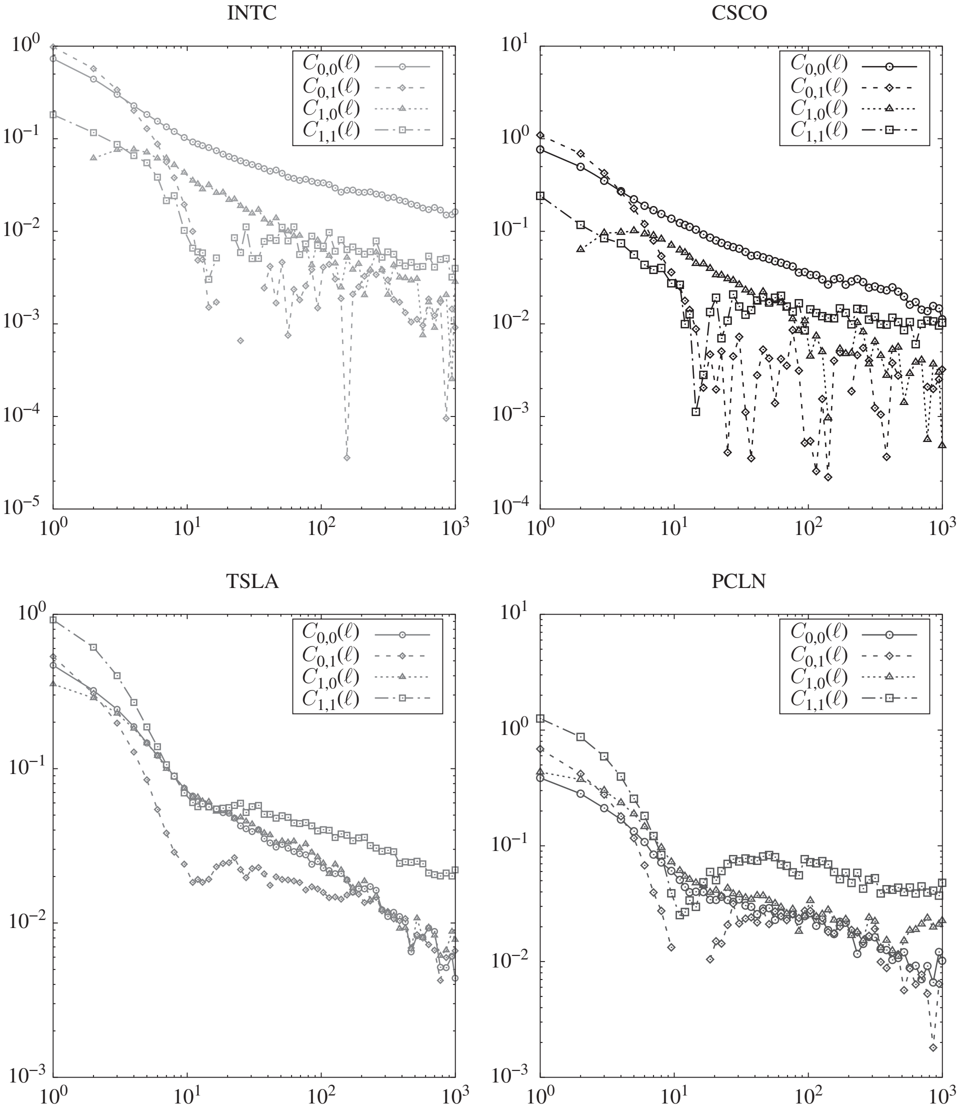

Figure 10.2. Conditional correlation functions (circles) $C_{0,0}$, (diamonds) $C_{0,1}$, (triangles) $C_{1,0}$ and (squares) $C_{1,1}$, of the market order sign process for (top left) INTC, (top right) CSCO, (bottom left) TSLA and (bottom right) PCLN. Missing points correspond to negative values.

10.3 Order-Sign Imbalance

For a finite event-time window of size N, consider the **order-sign imbalance**

$$\frac{1}{N}\sum_{t=1}^{N}\varepsilon_t.$$

The long-memory property of the ε_t series has an important consequence on the behaviour of this quantity. We will assume here and below that buy and sell market orders are such that, unconditionally, $\mathbb{E}[\varepsilon] = 0$. For short-memory processes, the amplitude of the fluctuations of the order imbalance in a window of size N scale as $1/\sqrt{N}$ for large N. More formally,

$$\mathbb{V}\left[\frac{1}{N}\sum_{t=1}^{N}\varepsilon_t\right] := \frac{1}{N} + \frac{2}{N}\sum_{\ell=1}^{N}\left(1-\frac{\ell}{N}\right)C(\ell).$$

When $C(\ell)$ decays faster than $1/\ell$, the second term on the right-hand side also behaves as $1/N$, and we indeed recover that the variance of the order imbalance decays as $1/N$ for large N. If, however, $C(\ell) \sim c_\infty/\ell^\gamma$ with $\gamma < 1$, then the second term instead behaves as $N^{-\gamma}$, and therefore dominates at large N, leading to

$$\mathbb{V}\left[\frac{1}{N}\sum_{t=1}^{N}\varepsilon_t\right] \approx_{N\gg 1} \frac{2c_\infty}{(2-\gamma)(1-\gamma)}N^{-\gamma} \gg N^{-1} \qquad (\gamma < 1).$$

Hence, the fluctuations of order-sign imbalance decay very slowly with increasing N. This is a consequence of the persistence of order flow: although on average there are an equal number of buy and sell market orders, in any finite window there is likely to be an accumulation of one or the other. This is very important for market-makers, who strive to keep a balanced inventory, because these long-range autocorrelations make their job more difficult (see Sections 1.3 and 17.2).

Another way to highlight the problem caused by order-flow correlations is to assume that each market order on average pushes the price by a small amount $G \times \varepsilon_t$. As we will discuss in detail in Chapter 11, this is called **price impact**: buy orders tend to push the price up and sell orders tend to push the price down. If this impact was permanent, the mid-price m_ℓ would evolve according to

$$m_\ell = m_0 + G\sum_{t=0}^{\ell-1}\varepsilon_t + \sum_{t=0}^{\ell-1}\xi_t,$$

where ξ_t is an independent noise term (with variance Σ^2, say) that models all the other possible causes of price changes. For $\gamma < 1$, the variogram of price changes (see Section 2.1.1) would then be

$$\mathcal{V}(\ell) := \mathbb{V}[m_\ell - m_0] \approx_{\ell\gg 1} \frac{2c_\infty G^2}{(2-\gamma)(1-\gamma)}\ell^{2-\gamma} + \Sigma^2\ell.$$

Within our simple framework, the price would therefore be super-diffusive (i.e. the variogram would grow faster than linearly with ℓ for $\gamma < 1$). In other words, the long memory in the order signs would create trends in the price trajectory, because buyers (or sellers) consistently push the price in the same direction for long periods of time. This is at odds with empirical data, which suggests that the above picture is inadequate. As we discuss in Chapter 13, the resolution of this apparent **efficiency paradox** can be found within the framework of the **propagator model**. In other words, the removal of the order-flow correlations is the result of the counter-balancing role of liquidity providers (see also Section 16.2.1).

10.4 Mathematical Models for Persistent Order Flows

10.4.1 Herding versus Splitting

What causes the autocorrelations in market order signs? Intuitively, two possibilities come to mind. The first possibility is **herding**, in which different market participants submit orders with the same sign. For example, a group of market participants could observe the arrival of a buy market order and infer that there is an active buyer interested at the current price. The group of market participants may decide to join the bandwagon, and create many more buy market orders. This behaviour can cascade (similarly to the Hawkes mechanism that we discussed in Chapter 9) and thereby lead to long sequences of trades in the same direction.

The second possibility is **order-splitting**, in which market participants who wish to execute large trades split their intended volume into many smaller orders, which they then submit incrementally. We will discuss why this splitting should occur in real markets in Section 10.5. Clearly, such activity could also lead to long sequences of orders with the same signs, and could therefore cause autocorrelations in the order-sign series.

In the remainder of this section, we will consider two mathematical models: one of herding and one of order-splitting. We will discuss how to calibrate each of these models to empirical data, discuss the autocorrelation structures that each model predicts, then use these results to infer the relative adequacy of these models for explaining the autocorrelations that occur in real markets.

10.4.2 A Model for Herding

In this section, we introduce a family of models called **discrete auto-regressive** (DAR) processes,[4] which can be regarded as discrete analogues of Hawkes processes. A DAR process is constructed as follows. The sign ε_t at time t is thought of as the "child" of a previous sign at time $t-\ell$, where the distance ℓ is a random variable distributed according to a certain discrete distribution $\mathbb{K}(\ell)$, with

$$\sum_{\ell=1}^{\infty} \mathbb{K}(\ell) = 1. \tag{10.4}$$

[4] These processes are sometimes called integer-value auto-regressive (INAR) processes.

For a given value of $k > 0$, if $\mathbb{K}(\ell) \equiv 0$ for all $\ell > k$, the model is called DAR(k), and involves only k lags. Once the mother sign is chosen in the past, one posits that:

$$\begin{aligned} \varepsilon_t &= \varepsilon_{t-\ell} && \text{with probability} \quad p, \\ \varepsilon_t &= -\varepsilon_{t-\ell} && \text{with probability} \quad 1-p, \end{aligned} \tag{10.5}$$

for some constant $p < 1$.

This is a “copy-paste” model for trades: it assumes that a trader trading at time t selects a previous market order that occurred some ℓ steps in the past (where ℓ is chosen according to the probability distribution $\mathbb{K}(\ell)$), and either copies the sign of that market order (with probability p) or goes against the sign of that market order (with probability $1-p$).

In the stationary state, the signs $+1$ and -1 are equiprobable, and the autocorrelation function $C(\ell)$ obeys the following **Yule–Walker equation**:[5]

$$C(\ell \geq 1) = (2p-1)\sum_{n=1}^{\infty} \mathbb{K}(n) C(\ell - n); \qquad C(0) = 1. \tag{10.6}$$

There is therefore a one-to-one relation between $\mathbb{K}(\ell)$ and $C(\ell)$, which allows one to calibrate the model, much as for Hawkes processes (see Section 9.3.5). Note that in the empirical case, where $C(\ell)$ decays as a power-law $\ell^{-\gamma}$ with exponent $\gamma < 1$, one can show (as in the corresponding Hawkes critical case) that $\mathbb{K}(\ell) \sim \ell^{(\gamma-3)/2}$ for large ℓ, together with $p \uparrow 1$ (compare with Equation (9.26)).

By construction, the conditional expectation of ε_t, given the past history of signs, is

$$\widehat{\varepsilon_t} = (2p-1)\sum_{\ell=1}^{\infty} \mathbb{K}(\ell)\varepsilon_{t-\ell}. \tag{10.7}$$

Therefore, the best predictor of the next sign in a DAR processes is a linear combination of the past signs. This result will turn out to be useful in Chapter 13.

10.4.3 A Model for Large Metaorders

An alternative model for the long-range correlation of order signs analyses the consequence of large, incrementally executed trading decisions. This model was originally proposed by Lillo, Mike and Farmer (LMF) in 2005.[6]

The LMF model assumes that the long memory in market order signs comes from very large orders that need to be split and executed slowly, over many individual trades. The underlying parent order is called a **metaorder**. It is well

[5] Note the similarity with the corresponding Equation (9.15) in the context of Hawkes processes.

[6] See: Lillo, F., Mike, S., & Farmer, J. D. (2005). Theory for long memory in supply and demand. *Physical Review E*, 71(6), 066122. Our presentation differs slightly in a few areas to simplify the mathematics, but is otherwise similar to theirs.

known that the amount of assets under management across different financial institutions is very broadly distributed, perhaps itself with a power-law tail. Therefore, if we assume that the size of a metaorder mostly depends on the size of the asset manager or mutual fund from which it originates, then we can reasonably expect the size of these metaorders to be power-law distributed as well.

The LMF model is a stylised model of a market, in which a fixed number M of independent metaorders are being executed at any instant of time. Each active metaorder $i = 1, \ldots, M$ is characterised by a certain sign ε_i and certain **termination rate** $\kappa_i \in (0, 1]$ drawn independently, for each metaorder, from a certain distribution $\rho(\kappa)$. At each discrete time step, one metaorder (say i) is chosen at random. Once chosen, the metaorder is either terminated (with probability κ_i), or has another unit executed (with probability $1 - \kappa_i$).[7] If terminated, the metaorder is replaced by a new metaorder with its own sign and termination rate. If it has a unit executed, the metaorder generates a market order with sign ε_i. Since the signs of metaorders are assumed to be independent, then for each $\ell \geq 1$, the value of the correlation function $C(\ell)$ is equal to the probability that a market order at time t and $t + \ell$ belong to the same metaorder. All other cases average to zero and do not contribute to $C(\ell)$.

Given that trade t belongs to a metaorder with termination rate κ, what is the probability that trade $t + \ell$ belongs to the same metaorder? Since the process is Markovian, this probability is given by

$$\left(1 - \frac{\kappa}{M}\right)^{\ell} \times \frac{1 - \kappa}{M},$$

because the probability for the metaorder to disappear is κ/M at each step, and the probability that the order at time $t + \ell$ belongs to the same metaorder as the order at time t is $1/M$. Similarly, the probability that a metaorder with termination rate κ remains active for a total number of time steps exactly equal to L is

$$\mathbb{P}(L|\kappa) = \left(1 - \frac{\kappa}{M}\right)^{L-1} \frac{\kappa}{M}.$$

Finally, the probability $\mathbb{P}(\kappa(\mathrm{MO}_t) = \kappa)$ that a market order at a randomly chosen time belongs to a metaorder with termination rate κ and duration L is proportional to $\rho(\kappa)$ and to $L \times \mathbb{P}(L|\kappa)$, since the longer a metaorder is, the more likely it is to be

[7] In contrast to the original LMF model, we assume that when a trader submits a chunk of a metaorder, this action does not diminish the remaining size of their metaorder. Therefore, in our presentation, no metaorders ever diminish to size 0. Instead, we assume that traders simply stop executing their metaorders with a fixed rate. We make this simplification because it greatly simplifies the algebra involved in deriving the model's long-run behaviour.

encountered.[8] Hence

$$\mathbb{P}(\kappa(\mathrm{MO}_t) = \kappa) = \frac{1}{Z}\rho(\kappa)\sum_{L=1}^{\infty} L\mathbb{P}(L|\kappa),$$

where Z is a normalisation constant, such that

$$\int_0^1 \mathrm{d}\kappa \mathbb{P}(\kappa(\mathrm{MO}_t) = \kappa) = 1. \tag{10.8}$$

Putting everything together, one finally arrives at an exact formula for $C(\ell)$, as the probability for the order at time t and the order at time $t+\ell$ belong to the same metaorder:

$$C(\ell) = \frac{1}{MZ}\int_0^1 \mathrm{d}\kappa \rho(\kappa)(1-\kappa)\sum_{L=1}^{\infty} L\mathbb{P}(L|\kappa)\left(1-\frac{\kappa}{M}\right)^{\ell}. \tag{10.9}$$

To make sense of these formulas, we will first make some extra assumptions. First, we will assume that the number of active metaorders is large, implying $\kappa/M \ll 1$, which allows us to obtain the unconditional probability that a metaorder remains active for exactly L steps as:

$$\mathbb{P}(L) = \int_0^1 \mathrm{d}\kappa \rho(\kappa)\mathbb{P}(L|\kappa) \approx_{L\gg 1} \frac{1}{M}\int_0^1 \mathrm{d}\kappa\, \kappa\rho(\kappa)e^{-\kappa L/M}.$$

Second, we will assume that $\rho(\cdot)$ can be written in the form $\rho(\kappa) = \zeta\kappa^{\zeta-1}$ with $\zeta > 1$. This leads to:

$$\mathbb{P}(L) \approx_{L\gg 1} \zeta^2\Gamma[\zeta]\frac{M^{\zeta}}{L^{1+\zeta}}.$$

Assuming that each market order has a size υ_0, this power law for L leads to a power-law decay of the distribution of the size $Q = \upsilon_0 L$ of metaorders, which is the initial motivation of the model and justifies our choice for $\rho(\kappa)$.

Within the same approximation,

$$\mathbb{E}[L] = \sum_{L=1}^{\infty} L\mathbb{P}(L|\kappa) \approx \frac{\kappa}{M}\int_0^{\infty} \mathrm{d}L\, L e^{-\kappa L/M} \approx \frac{M}{\kappa}, \tag{10.10}$$

and hence, using the normalisation condition from Equation (10.8),

$$Z = M\int_0^1 \mathrm{d}\kappa \frac{1}{\kappa}\rho(\kappa) = M\frac{\zeta}{\zeta-1}. \tag{10.11}$$

Finally, from Equation (10.9), for $\ell \gg 1$,

$$C(\ell) \approx \frac{1}{Z}\int_0^1 \mathrm{d}\kappa \frac{1-\kappa}{\kappa}\rho(\kappa)e^{-\kappa\ell/M} \approx_{\ell\gg M} \Gamma[\zeta]\frac{M^{\zeta-2}}{\ell^{\zeta-1}},$$

[8] The sum over all L of $L \times \mathbb{P}(L|\kappa)$ must be convergent for the model to be well defined and stationary.

where the right-hand side is now in the form $c_\infty \ell^{-\gamma}$, as in Equation (10.2). Using this notation, this model produces a power-law decay for the autocorrelation of the signs with an exponent $\gamma = \zeta - 1$. Since the decay exponent γ must be positive, one concludes that $\zeta > 1$. In order to reproduce the empirical value of $\gamma \cong 1/2$, the model requires the underlying metaorder sizes to be distributed roughly as $L^{-1-\zeta}$ with $\zeta = 3/2$.

Clearly, the assumption that the number of active metaorders M is fixed and constant over time is not realistic. However, it is not crucial, because the relation $\gamma = \zeta - 1$ is robust and survives in situations where M fluctuates in time. A hand-waving argument for this is as follows: the probability that two orders separated by ℓ belong to the same metaorder (and contribute to $C(\ell)$) is proportional to the probability that they both belong to a metaorder of size larger than ℓ, which scales like $\int_\ell^\infty \mathrm{d}L\, L\mathbb{P}(L) \sim \ell^{1-\zeta}$.

10.4.4 Herding or Order-Splitting?

As we discussed in Section 10.4.1, herding and order-splitting could both provide plausible explanations for the autocorrelations in market order signs that occur empirically. Which is more likely to be the main reason for the observed phenomena? Although both explanations are likely to play a role, several empirical observations suggest that the influence of order-splitting is much stronger than that of herding.

One way to confirm this directly would be to analyse a data set that provides the identity of the initiator of each market order. This would allow detailed analysis of which market participant submitted which market order, and would thereby provide detailed insight into the relative roles of herding and order-splitting. Unfortunately, comprehensive data describing order ownership is very difficult to obtain. However, fragmented data from brokers, consultants and proprietary sources all confirm that order-splitting is pervasive in equity markets, futures markets, FX markets, and even on Bitcoin markets, with a substantial fraction of institutional trades taking several days or even several weeks to complete (see references in Section 9.7).

In some cases (such as the London Stock Exchange (LSE), the Spanish Stock Exchange (SSE), the Australian Stock Exchange (ASE) and the New York Stock Exchange (NYSE)), partial information about the identity of participants can be obtained through the code number associated with the broker who executes the trade. This data suggests that most of the long-range autocorrelations in $C(\ell)$ originate from the same broker submitting several orders with the same sign, rather than from other brokers joining the bandwagon. In fact, although some herding can be detected on short time lags (e.g. for ℓ less than about 10), the behaviour of other brokers at large lags is actually contrarian, and contributes *negatively* to $C(\ell)$.

After a short spree of copy-cat behaviour, other market participants react to a flow of buy market orders by sending sell market orders (and vice-versa), presumably because buy orders move the price up, and therefore create more opportunities for sellers.[9] This indicates that on the time scale over which $C(\ell)$ decays as a power-law, herding is not a relevant factor.

Is there any empirical evidence to suggest that the size of metaorders is power-law distributed, as assumed by the LMF model? Without identification codes, this is again difficult. However, a power-law distribution with an exponent $\zeta \cong 3/2$ has indeed been reported in the analysis of block trades (traded off-book in the upstairs market) on the LSE, in the reconstruction of large metaorders via brokerage codes on the SSE, and in a set of large institutional metaorders executed at Alliance Bernstein's buy-side trading desk in the US equities market. Bitcoin data allows a precise reconstruction of large metaorders, and suggests a smaller exponent $\zeta \cong 1$.[10]

10.5 Liquidity Rationing and Order-Splitting

In Section 10.4.4, we noted that a variety of empirical evidence suggests that order-splitting is the primary cause of the observed long-term autocorrelations in market order signs. This raises an interesting question: why would investors split their orders in the first place, rather than submitting their desired trades as quickly as possible, before the information edge they have (or believe they have) becomes stale?

As we noted at the end of Chapter 1, market participants face a quandary. Many buyers want to buy (and many sellers want to sell) quantities that are, in aggregate, very substantial. For example, the daily traded volume for a typical stock in an equities markets usually amounts to roughly 0.1%–1% of its total market capitalisation. Volumes in some commodity futures or FX markets are breathtaking. However, as we will discuss in this section, the total volume offered for an immediate transaction at any given instant of time is typically very small. Why is this the case? As we will argue, the answer is that most market participants work very hard to hide their real intentions to trade. An important consequence is then that the available liquidity at a given moment in time is rather like an iceberg:

[9] See; Tóth, B., Palit, I., Lillo, F., & Farmer, J. D. (2015). Why is equity order flow so persistent? *Journal of Economic Dynamics and Control*, 51, 218–239 and Tóth, B., Eisler, Z., & Bouchaud, J. P. (2017). https://ssrn.com/abstract=2924029, for the same analysis on CFM proprietary data.

[10] See: Vaglica, G., Lillo, F., Moro, E., & Mantegna, R. (2008). Scaling laws of strategic behavior and size heterogeneity in agent dynamics. *Physical Review E*, 77, 0036110; Bershova, N., & Rakhlin, D. (2013). The non-linear market impact of large trades: Evidence from buy-side order flow. *Quantitative Finance*, 13, 1759–1778; Donier, J., & Bonart, J. (2015). A million metaorder analysis of market impact on the Bitcoin. *Market Microstructure and Liquidity*, 1(02), 1550008.

it reveals only a small fraction of the huge underlying supply and demand. Most of the liquidity remains *latent*.

10.5.1 The Buyer's Conundrum

Imagine that after careful examination of some information, you estimate that the price of some given stock XYZ should increase by 20% in the coming year. After considering your portfolio and performing a suitable risk analysis, you decide to buy 1% of the market capitalisation of XYZ. You also know that the mean daily traded volume of XYZ is 0.5% of its market capitalisation. Even if you decide to participate with $\frac{1}{4}$ of that daily volume with your own trades (which is already quite large), it will take you at least eight days to complete your desired total volume. If we assume that daily volatility is 2%, then over eight days, the value of the stock typically fluctuates by $2\% \times \sqrt{8} \approx 6\%$. Therefore, about one-third (6/20) of your expected profit might evaporate during this execution period. This provides a strong incentive to hurry through your trades as much as possible. How should you proceed?

Frustratingly, there probably also exists a handful of other market participants who are actively trying to offload a position in stock XYZ, and who would be ready to sell you the full 1% at the current market price. However, you do not know who these people are – or where to find them. Moreover, you cannot tell them that you want to buy such a large quantity (for fear they will try to negotiate a much higher price), and they cannot tell you that they want to sell such a large quantity (for fear you will try to negotiate a much lower price).[11] This is precisely the **buyer's conundrum** illustrated by the quote at the beginning of Chapter 1.

If 0.5% of the market cap is traded over the course of a whole day, then roughly 0.005% is traded every five minutes (neglecting intra-day activity patterns and the opening auction and closing auction). This is roughly the volume typically available at the best bid or ask quote for small-tick stocks; the corresponding volume for large-tick stocks is somewhat larger, but only by a factor of about 10. Sending a market order whose volume is much larger than these numbers is clearly a bad idea, because it would mean paying a much worse price than the ask-price, and could possibly wreak havoc in the market (remember that your desired metaorder size is between 20 and 200 times larger than the volume available at the ask!). Similarly, sending a very rapid succession of smaller market orders would presumably send a strong signal that there is a hurried buyer in the market, and would likely cause many sellers to increase their prices.

[11] In the old days, traders seeking to perform a large buy (respectively, sell) trade often commissioned a broker to find sellers (respectively, buyers). Such traders hoped that the broker was smart enough to get a good price while hiding their true intentions to trade. This was often a long, drawn-out and possibly costly process, with little transparency. However, even when using a broker, information leakage remained a costly consideration.

Rather than removing liquidity, how about instead contributing to liquidity by posting a very large limit order at or close to the bid-price? Unfortunately, this does not work either, because observing an unusually large limit order also signals a large buying interest. This influences both buyers (who are now tempted to buy at the ask-price rather than hoping to achieve a better price by placing their own limit orders) and sellers (who think that it might be a bad idea to sell now if the price is likely to go up, as suggested by the new limit order arrival). In fact, limit orders – which are often described as "passive" because they provide liquidity – can impact prices considerably.

The so-called **flash crash** of 6 May 2010 provides an example of the possible dangers of submitting unusually large limit orders. On this day, an asset manager decided to sell 75,000 S&P E-mini contracts (representing about 8% of the typical daily volume) using several huge sell limit orders.[12] The appearance of such an enormous volume in an already agitated market created a sudden drop of liquidity on the buy-side of the LOB and a rapid decline of prices that reached −9% after a few minutes, while also spilling over to equity markets. This is an extreme but vivid example of how signalling a desire to buy or sell a large quantity can impact and possibly destabilise prices, even with a purely passive behaviour. One clearly sees how self-reinforcing liquidity droughts can appear in financial markets. In other words, even limit orders that do not directly remove any liquidity from the market can indirectly generate a **liquidity crisis**.

In summary, as a buyer who seeks to purchase a very large quantity of an asset, you have no other realistic choice than to split your desired trade into many small pieces and execute them incrementally, over a period which might span several days or even months. Intuitively, the probability of revealing information increases with the size of a (limit or market) order, because smaller orders are more likely to go unnoticed while larger orders are more likely to attract attention. Therefore, the **information leakage cost** of an order is expected to increase with its volume (relative to the available liquidity). Of course, the same arguments hold for sellers as well. In both cases, these actions are consistent with the idea that market participants' execution of large metaorders can cause long-range autocorrelations in the observed order flow.

10.5.2 Metaorder Execution

In practice, it is difficult to manually execute a large metaorder over a long time period in an efficient way. Therefore, execution algorithms are now routinely used to perform this task. These algorithms are either built in-house by asset managers,

[12] See: Kirilenko, A. A., Kyle, A. S., Samadi, M., & Tuzun, T. (2017). The flash crash: High frequency trading in an electronic market. *The Journal of Finance*. doi:10.1111/jofi.12498.

or proposed by brokers who sell execution as a service (with a fee!). We now provide a brief description of what these algorithms or brokers attempt to achieve. We also return to this question, in the context of optimal execution, in Chapter 21.

Some common execution schedules are:

- **The time-weighted average price (TWAP) benchmark:** TWAP execution aims to achieve an average execution price that is as close as possible to the time-weighted average price available in the market during a specified period (typically one trading day).
- **The volume-weighted average price (VWAP) benchmark:** VWAP execution aims to achieve an average execution price that is as close as possible to the volume-weighted average price available in the market during a specified period (typically one trading day).
- **The Almgren–Chriss optimal schedule:** This algorithm aims to find an execution strategy that minimises a combination of trading costs and the variance of the difference between the execution price and a given reference price (such as the price at the open). We discuss this algorithm in more detail in Chapter 21.

Intuitively, VWAP reflects that large volumes should be more representative of the fair price paid during the day, whereas TWAP is insensitive to the size of the trades. Note that the TWAP and VWAP benchmarks are somewhat misleading since they hide the impact of the executed metaorder: in the limit of a very large metaorder dominating the market, its average execution price is very close to the VWAP, since the VWAP is computed using the trades of the metaorder itself. This misleadingly suggests a high-quality execution, when these benchmark prices themselves are adversely impacted by the metaorder and actually quite far from the decision price (such as the price at the beginning of the day). The Almgren–Chriss algorithm attempts to correct this drawback.

One can also devise more sophisticated algorithms that take into account the local liquidity fluctuations or short-term predictability in the price, or even the simultaneous execution of different metaorders on different instruments. In any case, the important conclusion is that all such execution algorithms slice metaorders into small pieces that are executed incrementally, either as market orders or as limit orders, and thereby result in a long-range autocorrelated order flow.

10.6 Conclusion

The long memory of market order signs is a striking stylised fact in market microstructure. At first sight, the effect is extremely puzzling, because it appears to contradict the near-absence of predictability in price series. How can it be that one can make good predictions of the sign of a market order far in the future without being able to predict that the price will increase or decrease over the same time horizon? As we have discussed in this chapter, it must be the case that the market somehow reacts to the correlated order flow in such a way that the price becomes (approximately) *statistically efficient*. We will return to our discussion of this *efficiency paradox* several times in the coming chapters.

In this chapter, we have also summarised evidence to support that the long memory of order flow is a consequence of metaorder-splitting. Even in so-called "liquid" markets, such as US large-cap stocks, investors are not faced with plentiful liquidity. As we have argued, the volume available in the LOB is only a very small fraction of the total volume desired for trade in the market. Therefore, the only sensible possibility for market participants who wish to execute large trades is to slice and dice their desired metaorders into small quantities, which they execute incrementally over long periods of time.

An important conclusion is that at any instant of time, there are huge chunks of metaorders that still await execution. At odds with Walras' picture of price formation, where the price instantaneously clears supply and demand (see Chapter 18), markets in fact only slowly resolve the imbalance between buyers and sellers. Therefore, the revealed liquidity in an LOB far from illuminates the true buying and selling intentions in the market. Instead, most of the liquidity remains latent, as we will explore in detail in Chapter 18.

Take-Home Messages

(i) The signs of arriving market orders have long-range autocorrelations. This makes the signs of future market orders predictable, which seems to be at odds with the (nearly) uncorrelated nature of price returns.

(ii) Empirical studies suggest that the main cause of these autocorrelations is single investors splitting large metaorders, rather than different investors herding.

(iii) Traders seek to minimise the information leakage that occurs when they make their trading intentions public. Therefore, the liquidity in an LOB is typically much smaller than the sum of all latent trading intentions.

(iv) Due to this shortage of liquidity, investors cannot execute large orders quickly without destabilising the market, and therefore need to split them using execution strategies such as VWAP, TWAP and Almgren–Chriss execution.

10.7 Further Reading

Long-Memory Processes, DAR(p) and Other Models

Jacobs, P. A., & Lewis, P. A. (1978). Discrete time series generated by mixtures. I: Correlational and runs properties. *Journal of the Royal Statistical Society*. Series B (Methodological), 40, 94–105.

Jacobs, P. A., & Lewis, P. A. (1978). Discrete time series generated by mixtures. III. Autoregressive processes (DAR (p)). *Naval Postgraduate School Technical Report* (Monterey, CA).

Jacobs, P. A., & Lewis, P. A. (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *Journal of Time Series Analysis*, 4(1), 19–36.

Raftery, A. E. (1985). A model for high-order Markov chains. *Journal of the Royal Statistical Society*. Series B (Methodological), 47, 528–539.

Beran, J. (1994). *Statistics for long-memory processes*. Chapman & Hall.

Berchtold, A., & Raftery, A. E. (2002). The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science*, 17, 328–356.

Lillo, F., Mike, S., & Farmer, J. D. (2005). Theory for long memory in supply and demand. *Physical Review E*, 71(6), 066122.

Taranto, D. E., Bormetti, G., Bouchaud, J. P., Lillo, F., & Tóth, B. (2016). Linear models for the impact of order flow on prices II. The Mixture Transition Distribution model. https://ssrn.com/abstract=2770363.

Long-Range Correlation of Order Flow

Bouchaud, J. P., Gefen, Y., Potters, M., & Wyart, M. (2004). Fluctuations and response in financial markets: The subtle nature of random price changes. *Quantitative Finance*, 4(2), 176–190.

Lillo, F., & Farmer, J. D. (2004). The long memory of the efficient market. *Studies in Nonlinear Dynamics & Econometrics*, 8(3), 1.

Bouchaud, J. P., Kockelkoren, J., & Potters, M. (2006). Random walks, liquidity molasses and critical response in financial markets. *Quantitative Finance*, 6(02), 115–123.

Bouchaud, J. P., Farmer, J. D., & Lillo, F. (2009). How markets slowly digest changes in supply and demand. In Hens, T. & Schenk-Hoppe, K. R. (Eds.), *Handbook of financial markets: Dynamics and evolution*. North-Holland, Elsevier.

Yamamoto, R., & Lebaron, B. (2010). Order-splitting and long-memory in an order-driven market. *The European Physical Journal B-Condensed Matter and Complex Systems*, 73(1), 51–57.

Tóth, B., Palit, I., Lillo, F., & Farmer, J. D. (2015). Why is equity order flow so persistent? *Journal of Economic Dynamics and Control*, 51, 218–239.

Taranto, D. E., Bormetti, G., Bouchaud, J. P., Lillo, F., & Tóth, B. (2016). Linear models for the impact of order flow on prices I. Propagators: Transient vs. history dependent impact. https://ssrn.com/abstract=2770352.

Large Institutional Trades

Chan, L. K., & Lakonishok, J. (1993). Institutional trades and intra-day stock price behavior. *Journal of Financial Economics*, 33(2), 173–199.

Chan, L. K., & Lakonishok, J. (1995). The behavior of stock prices around institutional trades. *The Journal of Finance*, 50(4), 1147–1174.

Gabaix, X., Ramalho, R., & Reuter, J. (2003). *Power laws and mutual fund dynamics*. MIT mimeo.

Gabaix, X., Gopikrishnan, P., Plerou, V., & Stanley, H. E. (2006). Institutional investors and stock market volatility. *The Quarterly Journal of Economics*, 121(2), 461–504.

Vaglica, G., Lillo, F., Moro, E., & Mantegna, R. (2008). Scaling laws of strategic behavior and size heterogeneity in agent dynamics. *Physical Review E*, 77, 0036110.

Schwarzkopf, Y., & Farmer, J. D. (2010). Empirical study of the tails of mutual fund size. *Physical Review E*, 81(6), 066113.

Bershova, N., & Rakhlin, D. (2013). The non-linear market impact of large trades: Evidence from buy-side order flow. *Quantitative Finance*, 13(11), 1759–1778.

Donier, J., & Bonart, J. (2015). A million metaorder analysis of market impact on the Bitcoin. *Market Microstructure and Liquidity*, 1(02), 1550008.

Kyle, A. S., & Obizhaeva, A. A. (2016). Large bets and stock market crashes. https://ssrn.com/abstract=2023776.

Kirilenko, A., Kyle, A. S., Samadi, M., & Tuzun, T. (2017). The flash crash: High frequency trading in an electronic market. *The Journal of Finance*, 72, 967–998.

https://doi.org/10.1017/9781316659335.015