Modeling market impact and timing risk in volume time

Slava Mazur

Abstract. Intraday volatility and market impact models in volume time are proposed. We build an intraday volatility profile to capture non-stationarity of intraday price returns and utilize a fractional Brownian motion process to measure deviations from square root scaling rule of volatility.

We propose a generalized, scalable market impact model that encompasses two mainstream approaches: an aggregated impact of a series of trades on a sufficiently long trading horizon and a transient impact of individual trades.

We give an intuitive interpretation of the model parameters and provide a generalized formulation of the optimal trading horizon and efficient trading frontier.

The self-similarity feature of an aggregated model allows for its application to smaller trading horizons and modeling of transient impact of sliced orders. We formulate conditions when the impact of sliced orders can be consistently aggregated to the total impact of the original order and deduce relationships between parameters of macro and micro level models to enforce such consistency.

We demonstrate that the parameters of aggregated and transient impact models are intimately related to the auto-covariance function of trade signs. We give an explicit formulation of such a relationship when the stated auto-covariance function has a power law form.

Keywords: Intraday Volatility, Fractional Brownian Motion, Hurst Index, Market Impact, Efficient Trading Frontier, Transient Impact, Decay Kernel

1. Introduction

There are two major approaches to modeling market impact. In an aggregated market impact model, such as that of MSCI Barra (1997), Kissell and Glantz (2003) and Almgren et al. (2005), price changes caused by trading activity are modeled based on the amount of shares executed during a certain time period. There, a concept of temporal and permanent components of the total market impact is introduced and the power law of market impact is proposed. An alternative, more recent approach is considered for example in Bouchaud et al. (2004), Obizhaeva and Wang (2013), Bouchaud et al. (2008), Gatheral (2010) and Gatheral et al. (2011), where the market impact is modeled on a micro-level of each individual trade followed by an aggregation technique based on the concept of the transient nature of instantaneous impact and decay kernel to evaluate the aggregated impact for a given time period. In this paper we introduce a general model that encompasses both approaches on the basis of scalability of market impact model.

First we build a scalable aggregated market impact model and derive conditions under which a macro-level model can be scaled down to evaluate a micro-level instantaneous market impact and the latter can be consistently aggregated back to a larger scale with the help of a suitably chosen decay kernel. We demonstrate how parameters of the macro-level model relate to dissipation rate of transient market impact to maintain such consistency.

Then we take advantage of universal nature of the model under consideration to evaluate a price reversion time after trade ends and thus to estimate the temporal component of total market impact.
Finally, we analyze the relationship between transient market impact and a decay kernel on the one side and auto-covariance function of trade signs on another, and discuss applicability of the market efficiency hypothesis to modeling market impact.

All the results presented in this paper are presented in volume time. This is similar to Bouchaud et al. (2004), Bouchaud et al. (2008) and other sources and dictated by the nature of the task at hand, which suggests that volume or trading time is more natural and convenient in this case than a clock time. Indeed, we are more inclined to evaluate a price change not because a certain amount of time elapsed, but rather because a certain volume or number of trades executed. In this paper we present a model of price change in volume time based on market data, thus we estimate how a particular stock reacts to perturbations caused by trading activity.

Our analysis is two-fold. First, we propose a model of volatility in volume time. Its purpose is to evaluate the timing risk of a trading strategy. Another purpose, which is not less important, is to provide scaling and normalization rules applicable to modeling market impact. The rest of the paper is devoted to the market impact model and its properties.

2. Intraday volatility profile and timing risk

First we observe that an intraday price change process is far from stationary. Intuitively it is clear that volatility, as a means of measuring the magnitude of price changes, is higher near market open and close times and lower around noon. In non-US markets additional intraday patterns may be observable. It is also natural to assume that the impact of a trade in a stock at a certain time of the trading day depends on the volatility of the stock at that time. A model for intraday volatility in clock time and a method of removing its non-stationary component were presented in Sokalska et al. (2005) and later enhanced and implemented in the Liquidnet intraday volatility model (Mazur, 2009). In volume time a certain pattern of intraday volatility is also observable although slightly different from the one in clock time, see Figure 1. For many stocks intraday volatility pattern in volume time is arguably more informative compared to that in clock time. This is particularly true for illiquid, thin-traded stocks.

To evaluate an intraday volatility profile in volume time we divide a trading day’s volume into 390 equal segments and for each time segment we estimate the standard deviation of the “one volume minute”

![Intraday Volatility Profile](image)

Fig. 1. Intraday volatility profile and the scaling rule on a log-log scale. A dashed line corresponds to an ordinary Brownian motion.
log-returns on the same time segment for 60 days sample of historical data.

2.1. Scaling rule and another dimension of volatility

Since our goal is to build models on different time scales, it is important to require scalability of intraday volatility model as well. Consider the following stochastic model of price change. Let $S_\tau$ be a log-price at volume time $\tau$. Assuming the time interval $[0, \tau]$ is sufficiently small so that the volatility does not vary much and we can consider a price change on such an interval as being stationary, we model such a price change from time zero to $\tau$ via a stochastic integral with respect to a fractional Brownian motion (fBm) process, see e.g. Embrechts and Maejima (2002):

$$S_\tau - S_0 = \bar{\sigma} \int_0^\tau dB_H(s)$$ (1)

A fractional Brownian motion process possesses two remarkable properties that makes it particularly suitable for our purposes. First, it is a self-similar process and this makes it applicable across different time scales. Second, it explains the long-range dependency in asset returns in a parsimonious way. The phenomenon of long-range dependency in financial markets has been observed since the 1960s and reported in numerous publications e.g. Mandelbrot (1967), Mandelbrot (1971), Greene and Fielitz (1977), Ding et al. (1993), Cont (2005) and Bayraktar et al. (2008). The idea of utilizing fBm processes to account for long-range dependency can be traced from Cutland et al. (1995), Mandelbrot (1997) and Shiryaev (1999).

It is well known that the applicability of fBm processes to financial modeling is limited by the fact that it does not belong to the class of semi-martingales and thus is incompatible with the market efficiency hypothesis. An in-depth discussion concerning this topic can be found in Cont (2005). Rogers (1997) builds a concrete example of arbitrage and even calls fBm as being an “absurd candidate” for modeling a price process. In our model, however, we do not rely on the market efficiency hypothesis. Quite the contrary, our assumption is that systematic, non-random trading activity brings about deviation from market efficiency. Such a deviation particularly manifests in creating a long-range dependency in stock returns. Our goal is to build a model for such a phenomenon and thus within the scope of our model an fBm process seems to be quite relevant for asset price dynamics.

The standard deviation of price change in (1) depends on the length of the time interval $\tau$ and the Hurst index $H$ and is given by

$$\sigma_H(\tau) = \sqrt{\text{Var}[S_\tau]} = \bar{\sigma} \tau^H$$ (2)

from which immediately follows a volatility scaling rule:

$$\sigma_H(\tau) = \sigma_H(s) \left( \frac{\tau}{s} \right)^H$$ (3)

In the following sections we show that the Hurst index $H$ plays an important role in modeling market impact as well. There are several methods of estimation of Hurst index: an $R/S$ analysis proposed by Mandelbrot and Taqqu (1979), a method based on wavelet analysis (Bayraktar et al., 2008), a method based on estimation of probability density of returns (Cont, 2005) – these are just a few examples of variety of methods available. In our model we took a simplified approach due to Mandelbrot (1997).

Assuming the Hurst index is constant throughout a day, we calculate the mean of the following quantities based on a suitable sample of market data:

$$\left\{ \frac{\log(\Delta S_\tau)}{\log(D)}, 0 < \tau < 1 \right\}$$ (4)

where $\Delta$ is a sufficiently small sampling interval in volume time (sufficiently long though so that the price change $\Delta S_\tau$ is not zero) and $D$ represents the log-difference between daily high and low prices.

For many stocks and samples $H$ is close to $1/2$, meaning that the price dynamics in volume time is close to an ordinary Brownian motion. However, there are many cases where we observe significant deviations of $H$ from $1/2$ on both intraday, as demonstrated in Figure 1 and daily scales as shown in the Figure 2, which also justifies our preference of fractional Brownian motion over the standard Brownian motion for modeling intraday volatility.

3. Aggregated market impact model

In an aggregated market impact model we operate on a “macro-level” and build a model that explains dependency of changes in a stock price on buy-sell imbalances for all market participants. Consider a certain trading horizon in volume time $[0, \tau]$, $0 < \tau \leq 1$. We split the total market volume $\tau$ to
two categories: a “buy-initiated” $v_B(\tau)$ or a “sell-initiated” $v_S(\tau)$ volume depending on whether a trade execution price is above or below the bid-offer midpoint respectively. We do not consider trades executed at mid-point as they usually indicate execution in dark venues which produce little or no market impact and are thus irrelevant to our goal. Now we can introduce an absolute buy-sell trading imbalance as being

$$Q(\tau) = v_B(\tau) - v_S(\tau)$$  \hspace{1cm} (5)

A relative buy-sell imbalance can thus be introduced in a straightforward way:

$$\frac{v_B(\tau) - v_S(\tau)}{v_B(\tau) + v_S(\tau)} = \frac{Q(\tau)}{\tau}.$$  \hspace{1cm} (6)

Depending on our goal we will consider buy-sell imbalances as being either random or deterministic quantities. It is convenient to consider them random for example during the model setup and calibration phase as demonstrated below and in the next section. During the model application phase however, we want to evaluate (a hypothetical) impact of our own trading activity and thus equating a buy-sell imbalance to the order size in question, which is known upfront and thus presents a deterministic quantity. The estimated impact is hypothetical because it does not take into account activity of other market participants and other factors affecting price change like news. In reality one cannot separate one’s own trading activity from all these external factors (nor is there an efficient way to take them into account) and thus one cannot effectively verify a market impact model based on market data – market impact is not an observable quantity. However, the value of a market impact model is its utility for comparing different trading strategies – this is its major purpose. Another example of deterministic imbalance is a participation strategy, where an agent executes a certain fixed fraction of the total market volume and thus produces a constant relative imbalance.

The impact of trading is measured by change in the log-price right before the trade commences (time zero) and immediately after it completes (time $\tau$). Our experiments show that for modeling purposes it is more convenient to operate with changes in a log-price normalized by a corresponding period’s volatility. By analogy with z-score we will refer to the corresponding quantity as “z-impact”. If we denote a period log-return by $r(\tau)$,

$$r(\tau) = S_\tau - S_0,$$  \hspace{1cm} (7)

and the same period’s price volatility is represented by $\sigma(\tau)$, then the z-impact is introduced by:

$$R(\tau) = \frac{r(\tau)}{\sigma(\tau)}.$$  \hspace{1cm} (8)

We build a model for aggregated z-impact based on the following two empirical observations supported by historical tick data.

**Empirical observation 1.** For moderate imbalances the z-impact is a linear function of relative imbalance. The following series of charts in Figure 3 supports this statement.
It should be noted that we intentionally consider relatively small imbalances, not more than 25%. The reason behind this is that large imbalances usually involve execution of big blocks. Similar to executions at mid-point discussed earlier, blocks are usually executed in dark venues and are thus irrelevant to our analysis since our goal is to build a model of the impact of trades on lit markets. Consequently, data about blocks executed in dark venues would only distort our model and thus we exclude them by imposing the upper limit for imbalances.

Observe that the slope coefficient $\kappa$ in Figure 3 has a tendency to increase when the trading horizon is increasing. This dependency is consistent across the board and has a recognizable functional form with random fluctuations, which gives us a hint for another empirical observation.

**Empirical observation 2.** The slope coefficient in an aggregated $Z$-impact model itself depends on a trading horizon and can be sufficiently accurately modeled via a power law. Figure 4 demonstrates such a dependency and a power law fit.

Besides goodness-of-fit there is another reason that justifies our choice of power law – its scalability. As will be shown in subsequent sections, this allows us
to scale down the aggregated market impact model and apply it to the instantaneous impact of individual trades, which can be aggregated back consistently with the original macro-level model. This is particularly important from a practical standpoint for optimal order slicing and building optimal trading strategies.

Following our empirical observations we model the aggregated z-impact as:

$$\hat{R}(\tau) = \kappa Q(\tau)\tau^{1-\gamma}, 0 < \gamma < 1$$  \hspace{1cm} (9)

and employing volatility scaling rule (3) with a Hurst index $H$ the model for period $\tau$ price change $\hat{r}(\tau)$ is given by:

$$\hat{r}(\tau) = \kappa \sigma_d Q(\tau)\tau^{H+\gamma-1}$$  \hspace{1cm} (10)

where $\sigma_d$ denotes daily volatility. Observe that for a constant rate participation trading strategy ($Q(\tau) = q_0\tau$) the latter becomes what is known as a “power law of market impact.” Alternately, a constant absolute imbalance $Q = Q_0$ corresponds to the problem of evaluating market impact of a fixed trade size over a variable trading horizon, which is specific, for example, to the Implementation Shortfall (IS) trading strategy. Clearly, in such a case the market impact is inversely proportional to the trading horizon $\tau$ as long as $H + \gamma < 1$. In this case from (10) and volatility scaling rule (3) it is easy to obtain a crude estimator of an optimal trading horizon for IS strategy in the mean-variance approach. We consider this problem in more details later in this paper.

Formula (10) can be interpreted in the following way. Changes in a stock price on a given trading horizon when trading imbalance is known are due to uncontrolled market volatility determined by $\sigma_d$ and Hurst index $H$ and a known trading activity determined by $Q, \kappa$ and $\gamma$. The combined effect of two factors has a multiplicative form and thus facilitates scalability. Observe that (10) confirms what intuitively seems quite logical – a regular trading activity should change the scaling rule of log-price change process making it super-diffusive and thus more predictable. When such a trade sequence ends the stock price should “diffuse” back to its “normal” scale determined by the Hurst index $H$. We address this aspect of our model later in a subsequent section where we evaluate reversion time after the end of trading schedule.

4. Instantaneous impact and decay kernel

In this section we take advantage of model (9)’s scalability and self-similarity to demonstrate how the aggregated model can be scaled down to evaluate an instantaneous market impact and how the latter can be aggregated back to a larger scale so the result of aggregation is consistent with the original model. We also introduce a decay kernel in a power law form and show that its exponent has a simple relationship to the parameter $\gamma$ obtained empirically in the previous section.

Consider a partition of the original trading horizon $[0, \tau]$ to a set of $n$ smaller intervals

$$\Delta = \tau/n$$  \hspace{1cm} (11)

and a trading strategy, representing a slicing of the original order of size $Q = Q(\tau)$, expressed in volume time units, into a sequence of “child” orders $\{Q_j; j = 1, \ldots, n; \sum_{j=1}^n Q_j = Q\}$. The functional form of the aggregated market impact
model allows for scaling the model (9) down to the trading horizon $\Delta$:
\[
\hat{R}_j(\Delta) = \kappa q_j \Delta^\gamma, \quad j = 1 \ldots n
\]  
(12)

where $q_j = \frac{Q_j}{\Delta}$ is a relative imbalance on j-th volume time interval. Obviously, the z-impact calculated on such a partition is sub-additive:
\[
\sum_{j=1}^{n} \hat{R}_j(\Delta) = \kappa Q(\tau) \Delta^{\gamma - 1}
\]
\[
= \kappa Q(\tau) \tau^{\gamma - 1} n^{1 - \gamma} \Rightarrow \hat{R}(\tau)
\]  
(13)

As $n \to \infty$, $Q_j$ represents an instantaneous imbalance and as it follows from (13) the corresponding instantaneous z-impact can not be permanent (otherwise the aggregated impact would tend to infinity). Thus, in order to aggregate the impact of smaller intervals consistently with the original aggregated model, contributions of instantaneous z-impacts should decay with time – the impact is transient, see Bouchaud et al. (2004), Obizhaeva and Wang (2013) and Gatheral (2010). To account for such a decay we introduce weight coefficients $W_j$ to the sum in (13), or a decay kernel, so that the aggregated market impact tends to some finite positive quantity as $n \to \infty$:
\[
0 < \lim_{n \to \infty} \sum_{j=1}^{n} \hat{R}_j(\Delta) W_j < \infty.
\]  
(14)

Following Bouchaud et al. (2004), Bouchaud et al. (2008) and Gatheral (2010) let us consider a decay kernel of the form:
\[
W_j = \frac{\gamma}{(n - j + 1)^\beta}, \quad j = 1, \ldots, n
\]  
(15)

Our next goal is to find a decay rate $d$ in (15) to satisfy (14). Proceeding in a straightforward way:
\[
\sum_{j=1}^{n} \hat{R}_j(\Delta) W_j = \kappa \Delta^{\gamma - 1} \sum_{j=1}^{n} \frac{\gamma q_j \Delta}{(n - j + 1)^\beta}
\]
\[
= \kappa \Delta^{\gamma - 1 + \beta} \sum_{j=1}^{n} \frac{\gamma q_j \Delta}{(n - j + 1) \Delta^\beta}
\]  
(16)

we can easily deduce the required condition:
\[
d = 1 - \gamma.
\]  
(17)

Observing that in this case the sum in (16) tends to a corresponding integral we conclude that
\[
\sum_{j=1}^{n} \hat{R}_j(\Delta) W_j \to \int_{0}^{\tau} \kappa \gamma q(s, \tau) ds
\]
\[
\int_{\tau - s}^{\tau} (s - u)^{\gamma - 1} du,
\]  
(18)

where $q(s, \tau)$ denotes a speed of trading at volume time $s$ given the total trading horizon $\tau$, $0 \leq s \leq \tau$, so that $\int_{0}^{\tau} q(s, \tau) ds = Q(\tau)$ and thus it defines a trading strategy. The aggregated z-impact of such a strategy can be calculated as follows:
\[
\hat{R}_q(\tau) = \int_{0}^{\tau} \kappa \gamma q(s, \tau) ds
\]
\[
\int_{\tau - s}^{\tau} (s - u)^{\gamma - 1} du
\]  
(19)

Note that in terms of (19) the parameter $\gamma$ introduced earlier via pure empirical observations actually determines the rate of decay of transient impact, see (17) – an unobservable quantity specific to each individual stock.

5. Optimal trading horizon and efficient trading frontier

The models for market impact and intraday volatility proposed in the previous sections allow us to build estimators for an optimal trading horizon and efficient trading frontier of Implementation Shortfall (IS) trading strategy similar to those proposed in Almgren and Chriss (2000) and Kissell and Glantz (2003). For simplicity, let us consider the trading strategy $q$ introduced in the previous section having the following parametric form Gatheral et al. (2011):
\[
q(s, \tau) = \frac{Q_0}{\tau B(\beta + 1, \nu + 1)} \left( \frac{s}{\tau} \right)^\beta \left( 1 - \frac{s}{\tau} \right)^\nu, \quad 0 \leq s \leq \tau,
\]  
(20)

where $Q_0$ is the total order size, $\tau$ is a trading horizon, $\beta, \nu$ are the shape parameters of strategy $q$, $-1 \leq \beta \leq 0$, $-1 \leq \nu \leq 0$. Given the market impact model (19) and volatility scaling rule (3) the expected cost of implementation of any strategy $q$ is Gatheral et al. (2011) and Markov et al. (2011):
\[
I(\tau) = \sigma_d \int_{0}^{\tau} q(s, \tau) s^H \int_{0}^{s} \frac{\kappa q(u, \tau)}{(s - u)^{\gamma - 1}} du ds
\]  
(21)
and its timing risk as defined in Kissell and Glantz (2003) is given by
\[ \Sigma^2(\tau) = \sigma_d \int_0^\tau s^{2H} \left( \int_s^\tau q(u, \tau) \, du \right)^2 \, ds. \] (22)

According to the mean-variance approach (Almgren and Chriss, 2000) the optimal trading strategy \( q_{opt} \) is the solution to the following minimization problem:
\[ q_{opt} = \arg \min_{q \in Q^*} \{ I(q) + \lambda \Sigma(q) \}, \] (23)
where \( Q^* \) is a suitably chosen set of feasible trading strategies and \( \lambda \geq 0 \) denotes a risk aversion parameter.

For example, the strategies in parametric form (20) depend on three parameters \( (\tau, \beta, \nu) \), their expected cost and timing risk are available in closed-form via special functions and the set of possible strategies is the following region (see the Appendix below):
\[ Q^* = \left\{ 0 < \tau \leq 1; \frac{1 + H + \gamma}{2} < \beta \leq 0; -\gamma < \nu \leq 0 \right\} \] (24)

Performing integration in (21), (22) and taking into account (20) one has:
\[ I(\tau, \beta, \nu) = Q_0^2 \sigma_d \kappa \gamma h(\beta, \nu, \gamma, H) \tau^{\gamma+H-1}, \] (25)
\[ \Sigma^2(\tau, \beta, \nu) = \frac{2Q_0^2 \sigma_d^2 f(\beta, \nu, H)}{2H + 1} \tau^{2H+1} \] (26)

Derivation of (25) and (26) and expression for functions \( f \) and \( h \) are given in the Appendix below.

Practitioners often fix the shape parameters of strategy \( q \) and are interested in the optimal trading horizon \( \tau \) only. In such a case direct minimization in (23)–(26), gives
\[ \tau_{opt} = \left( \frac{Q_0 \kappa (1 - H - \gamma)}{\lambda (H + 0.5) \sqrt{2f(\beta, \nu, H)/(2H + 1)}} \right)^{\frac{1}{\gamma-1}} \] (27)

Further development and application of such an approach can be found in Glukhov (2007) and Markov et al. (2011).

Alternately, one might be interested in the optimal trading schedule for a fixed trading horizon \( \tau \). This leads to a constrained minimization of (23) as a function of the strategy shape parameters \( \beta, \nu \).

If for each aggressiveness level \( \lambda \) we perform a constrained minimization (23)–(26) as a function of three parameters \( (\tau, \beta, \nu) \) and plot corresponding expected impact costs \( I \) against timing risks \( \Sigma \), we obtain what is known as an efficient trading frontier (Almgren and Chriss, 2000), see Figure 5.

6. Reversion time

Consider an arbitrary trading strategy \( q(s, \tau), 0 \leq s \leq \tau \). According to (19), we expect that implementation of such a strategy will result in a change in the stock price by \( R_q(\tau) \) volatility units. It is also natural to expect that after the trading schedule ends the price will revert back to the pre-trade level at some point \( T > \tau \) as demonstrated in Figure 6 below.

Our next goal is to build an estimate for \( T \). Within the scope of model (19) the residual z-impact of strategy \( q \) at the time \( T > \tau \) is given by
\[ \hat{R}_q(T, \tau) = \kappa \gamma \int_0^\tau q(s, \tau) \, ds \] (28)
and obviously represents a decreasing function of \( T \).

We could think of reversion time as a time moment \( T \) when the price process “diffuses back” to its normal “market” level, which in terms of our model means that the decayed perturbation caused by the strategy \( q \)

![AAPL](image-url)  
**Fig. 5.** Efficient trading frontier.
is scaled accordingly to log-returns observed on the market:

\[ r_q(T, \tau) = \hat{R}_q(T, \tau) \sigma(\tau) \propto \sigma(T). \] (29)

The latter is equivalent to

\[ \kappa \gamma \int_0^\tau q(s, \tau) ds \left( \frac{T-s}{T} \right)^{1-\gamma} \leq \delta \frac{\sigma(T)}{\sigma(\tau)} \] (30)

for some constant \( \delta > 0 \). The choice of \( \delta \) is arbitrary. One can consider, for example, a \( \delta \) corresponding to a half of the total impact produced:

\[ \delta = \frac{\hat{R}_q(\tau, \tau)}{2} \] (31)

Observe that \( \delta_0 = 1/2 \) corresponds to the “half-life” of market impact. Inequality (30) can be solved for \( T \) numerically. For example, for strategy \( q \) in the parametric form (20) inequality (30) where \( \delta \) as in (31) becomes

\[ \left( \frac{T}{\tau} \right)^{1-\gamma} _2F_1 \left( \beta + 1, 1 - \gamma; \beta + \nu + 2; \frac{\tau}{T} \right) \times \frac{B(\beta + 1, \nu + 1)}{B(\beta + 1, \gamma + \nu)} \leq \delta_0 \frac{\sigma(T)}{\sigma(\tau)} \] (32)

In the particular case of constant rate participation strategy, \( \beta = \nu = 0 \) and applying scaling rule (3) with the Hurst index \( H \) for volatility \( \sigma \), (32) can be simplified to:

\[ \left( \frac{T}{\tau} \right)^\gamma - \left( \frac{T}{\tau} - 1 \right)^\gamma \leq \frac{1}{2} \left( \frac{T}{\tau} \right)^H \] (33)

7. Decay kernel and auto-covariance of imbalances

In this section we show that the parameter \( \gamma \) introduced empirically in (9) is in fact completely defined by the auto-covariance function of signs of trading imbalances and the Hurst index \( H \). Since the decay kernel depends on \( \gamma \) via (17) we thus present an alternative method of its estimation via said auto-covariances and the Hurst index \( H \).

Consider a partition of trading horizon \([0, \tau]\) into smaller time intervals \( \Delta \) as in (11). In this case, however, we do not assume any particular deterministic trading activity. Instead, we consider the observed imbalances as being random variables with certain auto-covariance structure, which we will discuss momentarily. According to model (10) we estimate a log-price change on the interval \( j, j = 1, \ldots, n \) as follows:

\[ \hat{r}_j(\Delta) = \kappa \sigma q_j(\Delta) \Delta^{H+\gamma}, \ j = 1, \ldots, n. \] (34)

The estimate to the total log-price change is obviously the sum of the changes on smaller intervals:

\[ \hat{r}(\tau) = \sum_{j=1}^n \hat{r}_j(\Delta) = \kappa \sigma q^{H+\gamma} \sum_{j=1}^n q_j(\Delta) \] (35)

We assume that the relative imbalance \( q_j \) can be represented as a product of two independent components: an absolute value \( q \) and a sign \( \epsilon_j \):

\[ q_j(\Delta) = q(\Delta) \epsilon_j, \ E_q(\Delta) = \hat{q}(\Delta) > 0, \ \epsilon_j = \pm 1. \] (36)
Given (36) we calculate the variance of \( \hat{r}(\tau) \) in (35) as follows:

\[
\text{Var}[\hat{r}(\tau)] = (\kappa \sigma_d \tilde{q}(\Delta))^2 \Delta^{2(H+\gamma)} \times \left\{ \sum_{j=1}^{n} \text{Var}[\epsilon_j] + 2 \sum_{i<j} \text{Cov}[\epsilon_i, \epsilon_j] \right\}
\]

(37)

Note that for a sufficiently small partition interval, the imbalance is close to an individual trade size and thus we can consider a trade signs process as a good approximation for imbalance signs. It is reasonable to assume that the trade signs process is stationary with a constant variance.

First we consider a case where the auto-covariance function of trade signs is zero for each non-zero lag. It follows from our model settings that in this case the log-price change process is white noise, \( H = 1/2 \), and from (37):

\[
\text{Var}[\hat{r}(\tau)] \propto (\kappa \sigma_d \tilde{q}(\Delta))^2 \Delta^{2(H+\gamma)} \times 2 \sum_{i<j} \frac{\Delta^2}{(j\Delta - i\Delta)\alpha}
\]

(39)

\[
= (\kappa \sigma_d \tilde{q}(\Delta))^2 \Delta^{2(H+\gamma)} \times \left\{ \frac{\Delta^2}{(j\Delta - i\Delta)\alpha} \right\}
\]

(40)

Thus we conclude that to produce a measurable market impact a random trading sequence must possess a non-trivial auto-covariance structure and this is confirmed by another empirical observation supported by market data.

**Empirical observation 3.** The auto-covariance function of trade signs is a slowly decaying function sufficiently accurately approximated by a power law with exponent \( \alpha < 1 \) (see Figure 7).

Taking into account our assumptions

\[
\text{Var}[\hat{r}(\tau)] = 2(\kappa \sigma_d)^2 \tau^{2-\alpha} \int_0^{\tau} \int_0^x \frac{dy\,dz}{(x-y)^\alpha} = 2(\kappa \sigma_d)^2 \tau^{2-\alpha} \int_0^{1} \int_0^u \frac{dv\,du}{(u-v)^\alpha} = \frac{2(\kappa \sigma_d)^2 \tau^{2-\alpha}}{\alpha^2 - 3\alpha + 2}
\]

(41)

and thus the sought expressions for \( \gamma \) and the decay rate \( d \) are

\[
\gamma = 1 - \frac{\alpha}{2}, \quad d = 1 - \gamma = H + \frac{\alpha}{2}
\]

(42)

Fig. 7. ACF of trade signs from lag 10 and power law fit.
8. Relation to the existing models

Obviously, the model (10) is a generalization of the model proposed in Almgren et al. (2005). In our aggregated model we tried to give a meaning for the parameters obtained there by pure statistical model fit. As it follows from our research, parameters of the power law of market impact are both stock dependent (Figure 2) and time dependent (Figure 1). In addition, we proposed an alternative method of measuring the temporal impact.

The instantaneous market impact model and the aggregation technique are in many ways inspired by research in market micro-structure and a concept of transient market impact developed in Bouchaud et al. (2004), Obizhaeva and Wang (2013), Bouchaud et al. (2008) and Gatheral (2010). Our results in evaluation of the decay rate, however, are different and seem more accurate.

First of all, a quick look at the final formula for decay rate \( d = \frac{1-\alpha}{2} \) in our terms, obtained in Bouchaud et al. (2004, eq. 23) and Gatheral (2010, eq. 12) immediately raises the following concern: the faster the auto-correlation function of trade signs vanishes, the slower the decay rate. However, the faster decaying auto-correlations of trade signs is a characteristic of more liquid stocks. For example, for highly liquid US stocks like AAPL, MSFT etc. the auto-correlation function of trade signs vanishes, the slower the decay rate. This also presents another problem from a practical standpoint. When trying to aggregate the impact of a relatively long trading schedule with such a slowly decaying kernel, the resulting aggregated impact quickly becomes unrealistically high.

We believe that the cause of this inconsistency is in attempts to reduce the sum of individual “impacted” returns to a white noise process on the basis of “market efficiency” argument, which in the cited sources means the square root scaling rule for volatility. To demonstrate our point let us consider a quotation from Gatheral (2010, p. 17):

“Empirically, we find that, to a very good approximation, \( \text{Var}[\Delta P] \propto N \). If the variance of price changes grew super-linearly as \( N^{2-\alpha} \), returns would be serially correlated and market efficiency would be glaringly violated: simple trend-following strategies would be consistently profitable. We conclude that market impact cannot be permanent.”

The matter, however, is that an intended, not completely random market activity brings about market inefficiency in a sense that it violates the square root scaling rule for volatility and indeed introduces serial correlation of returns. Such an inefficiency, however, is not quite easy to exploit since educated traders utilize different techniques to conceal their intentions. The task at hand is exactly the opposite to the market efficiency assumption – how to quantify the deviation from a square root law caused by systematic trading and thus the model for it cannot be based on market efficiency argument. On the other side, traders indeed try to detect and exploit their counterparties’ intentions and all these activities end up in historical market data. Consequently, a model that relies on market data cannot at the same time rely on market efficiency hypothesis as the market data is effectively a record of attempts of exploitation of market inefficiencies. Given that, it seems an adequate model should not disregard violations of market efficiency, on the contrary, it should try to quantify them.

What methods can be used for detection and measurement of market inefficiency? Some of them are presented in this paper and utilized in construction of the market impact and volatility models, namely:

1. The Hurst index \( H \). If the market were efficient one would have \( H \approx 1/2 \) consistently.
2. The parameter \( \gamma \). For an efficient market one would have \( \gamma \approx 0 \) (see equation 38).
3. The model for auto-covariance function of trade signs. For an efficient market it would be trivial.

Also, instantaneous impact is indeed transient and decays; however, its decay is completely defined by the auto-correlation structure of trade signs, not the market efficiency argument, and thus no additional adjustments to the sum of instantaneous impacts needed.

The market efficiency argument could be applied not to the trading horizon under consideration, but rather to the time period after the trade is completed. It is quite reasonable to expect that after certain trading activity the scaling rule reverts to its normal form according to the market efficiency hypothesis. We demonstrated how to apply such an idea to calculate reversion time; however, it is based purely on empirical assumptions and was not confirmed by market data (and hardly could be).
9. Conclusion

The proposed approach generalizes the existing market impact models and provides synergy between macro- and micro-level approaches. Key points of such a generalization are as follows:

- A fractional Brownian motion can be considered as an adequate model for intraday volatility and the Hurst index plays an important role in modeling market impact, thus generalizing a well-known “square root law” of market impact.
- The aggregated, macro-level market impact model can be built with scalability in mind and its parameters can be relatively easily estimated from the market data.
- Because of its scalability the aggregated market impact model is applicable to infinitesimal volume time intervals. Such an application yields a micro-level estimator for instantaneous market impact. For consistent aggregation to the original time scale the instantaneous market impact should be transient, that is it should decay in time. To model the transient nature of instantaneous market impact we introduce a decay kernel and formulate conditions of consistency between micro- and macro-level impact models.
- The model framework based on the transient market impact and decay kernel concepts allows for relatively easy evaluation of price reversion time after the end of trading and thus provides an estimate of temporal impact in terms of Almgren et al. (2005).
- Treatment of relative imbalances as random variables and calculation of variance of sum of model returns reveals an important relation between the decay rate of instantaneous market impact and auto-covariance function of trade signs and thus gives an alternative way of estimating model parameters. We provide correct relationships between these quantities and explain the cause of discrepancy in the existing results.

Acknowledgements

The author wishes to express his gratitude to David Saltz, Vladimir Markov, Vacslov Glukhov, David Fellah, Akis Georgiou, Visweswaran Nageswaran, Scott Harrison, Tito Ingargiola and other present and former colleagues at Liquidnet who were abundantly helpful and offered invaluable assistance, support and guidance.

References

10. Appendix

10.1. Residual number of shares

Assuming \( q(u, \tau) \) is in the parametric form (20) for the residual number of shares one has:

\[
X(s, \tau) = q(u, \tau) du
\]

\[
= Q_0 - \frac{Q_0}{\tau B(1 + \beta, 1 + \nu)} \int_0^s \left( \frac{u}{\tau} \right)^\beta \left( 1 - \frac{u}{\tau} \right)^\nu du
\]

\[
= Q_0 - \frac{Q_0 s^{1+\beta}}{\tau^{1+\beta} B(1 + \beta, 1 + \nu)} \int_0^1 y^\beta \left( 1 - \frac{ys}{\tau} \right)^\nu dy
\]

\[
f(\beta, \nu, H) = \frac{\Gamma(\beta + \nu + 2)}{\Gamma(\beta + 1)} \left[ \frac{\Gamma(2H + \beta + 2)}{\Gamma(2H + \beta + \nu + 3)} - \frac{\Gamma(\beta + \nu + 2) \Gamma(2H + 2\beta + 3) \tilde{F}_2(\beta + 1, 2H + 2\beta + 3, -\nu; \beta + 2, 2(H + \beta + 2) + \nu; 1)}{\Gamma(\nu + 1)} \right].
\]

10.2. Derivation of (26)

Integrating by parts in (22) one has:

\[
\Sigma(q)^2 = \sigma_q^2 \int_0^\tau s^{2H} X(s, \tau)^2 ds
\]

\[
= \frac{\sigma_q^2}{2H + 1} \int_0^\tau X(s, \tau)^2 ds^{2H+1}
\]

\[
= -\frac{2\sigma_q^2}{2H + 1} \int_0^\tau s^{2H+1} X(s, \tau) dX(s, \tau)
\]

\[
= \frac{2\sigma_q^2}{2H + 1} \int_0^\tau s^{2H+1} q(s, \tau) X(s, \tau) ds.
\]

(44)

In the formulae above we employed the following obvious properties of \( X(s, \tau) \):

\[
X(\tau, \tau) = 0; \quad \frac{\partial X(s, \tau)}{\partial s} = -q(s, \tau).
\]

(45)

Taking into account (20):

\[
\Sigma(\tau, \beta, \nu)^2 = \frac{2\sigma_q^2 Q_0^{2H-1}}{(2H + 1)B(1 + \beta, 1 + \nu)^2} \times \int_0^\tau s^{1+\beta+2H} \left( 1 - \frac{s}{\tau} \right)^\nu ds
\]

\[
\times \int_0^\tau y^{1+\beta+2H} (1 - y)^\nu dy
\]

\[
\times \int_0^1 x^\beta (1 - x)^\nu dx dy.
\]

(46)

Thus \( \Sigma(\tau, \beta, \nu)^2 \) is (26) where
10.3. Aggregated z-impact

From (19)–(20):

\[
\hat{R}_q(\tau) = \frac{k\gamma Q_0}{\tau^{2-\gamma} B(1+\beta, 1+\nu)} \\
\times \int_0^{\tau} \left( \frac{s}{\tau} \right)^{\beta} \left( 1 - \frac{s}{\tau} \right)^{\nu+\gamma-1} ds \\
= \frac{k\gamma Q_0}{\tau^{1-\gamma} B(1+\beta, 1+\nu)} \\
\times \int_0^{1} x^{\beta} (1-x)^{\nu+\gamma-1} dx \\
= \frac{k\gamma Q_0 B(1+\beta, \gamma+\nu)}{B(1+\beta, 1+\nu)} \tau^{\gamma-1} \\
\]

as long as \( \nu > -\gamma \).

10.4. Derivation of (25)

Taking into account (19) integration in (21) yields:

\[
I(\tau, \beta, \nu) = \frac{\sigma_d k\gamma Q_0^2}{\tau^2 B(1+\beta, 1+\nu)^2} \\
\times \int_0^{\tau} \left( \frac{s}{\tau} \right)^{\beta} \left( 1 - \frac{s}{\tau} \right)^{\nu} s^H ds \\
\times \int_0^{s} \left( \frac{u}{\tau} \right)^{\beta} \left( 1 - \frac{u}{\tau} \right)^{\nu} \left( s - u \right)^{\gamma-1} du ds \\
= \frac{\sigma_d k\gamma Q_0^2 \tau^{H+\gamma+1}}{B(1+\beta, 1+\nu)^2} \\
\times \int_0^{1} y^{2\beta+H+\gamma} (1-y)^{\nu} \\
\times \int_0^{1} x^{\beta} (1-x)^{\gamma-1} (1-xy)^{\nu} dx dy.
\]

Thus \( I(\tau, \beta, \nu) \) is (25) where

\[
h(\beta, \nu, \gamma, H) = \frac{\Gamma(\gamma)\Gamma(\beta+\nu+2)\Gamma(H+2\beta+\gamma+1)}{B(1+\beta, 1+\nu)} \tilde{F}_2(1+\beta, H+2\beta+\gamma+1, -\nu; \beta+\gamma+1, H+2\beta+\gamma+\nu+2; 1).
\]

as long as \( H+2\beta+\gamma > -1 \).