Quantitative Finance

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rquf20

Liquidity fluctuations and the latent dynamics of price impact

Luca Philippe Mertens, Alberto Ciacci, Fabrizio Lillo & Giulia Livieri

Quantitative Finance, 2021
https://doi.org/10.1080/14697688.2021.1947511

Liquidity fluctuations and the latent dynamics of price impact

LUCA PHILIPPE MERTENS†, ALBERTO CIACCI‡, FABRIZIO LILLO §¶ and GIULIA LIVIERI*§

†Bloomberg L.P., New York, NY, USA
‡Blackett Laboratory and Centre for Complexity Science, Imperial College London, London, UK
§Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy
¶Department of Mathematics, University of Bologna, Piazza di Porta San Donato 5, 40126 Bologna, Italy

(*Received 22 December 2020; accepted 18 June 2021; published online 19 July 2021*)

Market liquidity is a latent and dynamic variable. We propose a dynamical linear price impact model at high frequency in which the price impact coefficient is a product of a daily, a diurnal, and an auto-regressive stochastic intraday component. We estimate the model using a Kalman filter on order book data for stocks traded on the NASDAQ in 2016. We show that our price changes estimates, conditional on order flow imbalance, on average 82% of real price changes variance. Evidence is also provided on the fact that the conditioning on filtered information improves the estimate of the LOB liquidity with respect to the one obtained from a static estimation of the price impact. In addition, an out-of-sample analysis shows that our model provides a superior out-of-sample forecast of price impact with respect to historical estimates.

Keywords: Liquidity; Price impact; Order flow imbalance; Kalman filter

1. Introduction

Statistical modeling of centralized limit order books, or LOBs, is a well-established and active area of research. Central to this topic is the concept of price impact or 'the correlation between an incoming order and the subsequent price change' (Bouchaud 2010). Grounded in the economic intuition that price changes should reflect the aggregate imbalance between supply and demand, Cont *et al.* (2014) introduce the *Order Flow Imbalance* (OFI), the sum of signed volume of all incoming orders at the best quotes over a time interval. The authors prove that, at the minute time-scale, price change is a linear function of the order flow imbalance and identify the price impact of order book events with the regression coefficient of this linear relationship.

This paper makes two contributions to the price impact literature. First, we provide empirical evidence of the autoregressive nature of price impact at short time scales. Second, leveraging this statistical feature, we describe an econometric model for *conditional* high frequency estimates of price impact. While the order flow imbalance is a strong predictor of price change at the minute time scale, at a higher frequency, the discreteness of the price change variable becomes a dominant feature, weakening the linear relationship between price changes and OFI, and thus the explanatory power of a simple regression model. We prove that, over short time scales, price impact has a statistically significant autocorrelation, even after controlling for the presence of the intraday pattern. This statistical feature has important consequences for estimation purposes: at a high frequency, price impact has a weaker dependence on OFI, but a higher dependence on recent LOB activity. By using a filtering approach, where price impact is modeled as a latent auto-regressive variable with noisy observations, we can exploit past information to form a prior on the current level of price impact and lower the estimation error.

Precisely, we introduce a model where price impact at time ℓ of day i is determined by the product of three components: a daily price impact component β_i, a deterministic intraday pattern π_ℓ, and a stochastic auto-regressive component $q_{i,\ell}$. This type of modeling is reminiscent of analogous models in the conditional variance literature (cfr. Engle and Sokalska 2012, among others). The resulting price impact process $\beta_{i,\ell}$ describes the time-dependent relationship between price change—normalized by the tick size δ—$\Delta P_{i,\ell}$ and the

*Corresponding author. Email: giulia.livieri@sns.it

order flow imbalance $\mathrm{OFI}_{i,\ell}$. In particular, our model can be thus written as

$$\Delta P_{i,\ell} = \beta_{i,\ell}\mathrm{OFI}_{i,\ell} + \epsilon_{i,\ell} \qquad \epsilon_{i,\ell} \sim \mathrm{NID}\left(0, \sigma_\epsilon^2\right) \tag{1}$$

$$\beta_{i,\ell} = \beta_i\,\pi_t\,q_{i,\ell} \tag{2}$$

$$q_{i,\ell} = 1 + \rho\left(q_{i,\ell-1} - 1\right) + \eta_{i,\ell} \quad \eta_{i,\ell} \sim \mathrm{NID}\left(0, \sigma_\eta^2\right) \tag{3}$$

We refer to Appendix 2 for the details on how we set the initial values for the parameters and for the process $(q_{i,\ell})$. In the setting in (1), price impact is linear in the order flow and permanent, but its value at a given day/time of the day is determined by the *price impact process* (2)–(3). Some observations are in order. The model postulates that even at the time-scale considered in the present paper, i.e. 1 min, the contemporaneous OFI is the main determinant of price changes and, in particular, it does not describe the joint dynamics of price changes and OFI. Importantly, we empirically verify the reliability of this exogeneity assumption on OFI by performing a study following the lines of Hasbrouck (1991). Also, the model assumes no commonality in the OFI of assets i and j with $j \neq i$ (Capponi and Cont 2020); the accounting for such a factor in our time-varying model represents an interesting direction for future research. Admittedly, the Gaussian assumption in (1) may be suspect because the price change variable may be confined to a discrete price grid. However, we reassure, through some classical diagnostic testing for the Kalman filter, that the forecast errors are (mostly) normal. Finally, the model suggests—and the empirical analysis verifies—that variation in intraday liquidity is not caused only by a deterministic and recurrent diurnal pattern, but also by a stochastic auto-regressive component, which can be associated with transitory phenomena. In other words, the predictable diurnal pattern of market depth (Cont *et al.* 2014) is not sufficient to explain all the auto-correlation of price impact. The model described in (1)–(3) has a linear Gaussian state–space representation that can be easily estimated using a Kalman filter (Durbin and Koopman 2012) and standard maximum-likelihood methods.

Our model has interesting applications for execution. Execution strategies typically slice an order into child orders, which are then executed within a given time window according to a trading schedule. Tactical execution of child orders could benefit from the real-time price impact forecast provided by our model. For example, a trader could defer a trade when the price impact is higher than the expected average level, waiting for market conditions where the likelihood of moving the price against her are lower.

Our model could be used to design an optimal execution schedule by generalizing the approach of Almgren and Chriss (2001). We remind that in this framework there are two kinds of market impact: a temporary impact and a permanent impact. In our model, we account only for the permanent market impact given by $\beta_{i,\ell}\mathrm{OFI}_{i,\ell}$; see (1). The trading volume v_ℓ due to the optimal execution (either with limit or with market orders) will bring an additional contribution to the OFI and the variations of $\beta_{i,\ell}$ influence the impact of both OFI and v_ℓ. In particular, the permanent impact of the latter will be $\beta_{i,\ell}v_\ell$. Thus the main difference with Almgren and Chriss (2001) is that the pre-factor of permanent impact is now time varying and must be filtered during execution. For this reason, differently from Almgren and Chriss (2001), the optimal execution schedule is not statically determined, but requires the use of stochastic optimization. Notice also that one could make the model more complex by adding a dynamic equation for the OFI in a spirit similar to Cartea and Jaimungal (2016). Also, we mention the works of Casgrain and Jaimungal (2019) and (2020) where the authors consider the optimal execution problem and market impact games, respectively, in the presence of order flow models driven by latent elements. These are potentially interesting applications of our modeling framework that we leave for future research.

We select five stocks listed on the National Association of Securities Dealers Automated Quotation System (NASDAQ) 100 index and we filter intraday time-series of price impact estimates. Our analysis leads to the following results: (i) the auto-regressive coefficient ρ is statistically significant with an average value of approximately 0.5, suggesting a half-life of 1 min for the process q; (2) conditioning on real-time information improves the estimate of the LOB liquidity with respect to the one obtained from the deterministic intraday pattern alone. (3) The variance of the price changes explained by $\beta_{i,\ell}\mathrm{OFI}$ as in (4) is significantly larger than the ones explained by both a model in which $\beta_{i,l}$ is statically computed and a model in which $\beta_{i,l}$ is book-reconstructed and set equal to half the inverse of the depth (as in the stylized LOB model, Cont *et al.* 2014). Moreover, it turns out that the price impact is much higher after the opening auction than during the rest of the trading day in that it exhibits a flat behavior followed by a decline before the closing auction.

The paper is organized as follows. Section 2 is a literature review. Section 3 describes the data. Section 4 introduces the model of Cont *et al.* (2014) and provides a descriptive analysis of the price impact coefficient constructed with their methodology. The focus is on the performance of the model for different estimation timescales. Ultimately, the empirical findings presented in this section are used to justify the assumptions of our model. In section 5, we present our model and describe the estimation methodology in detail. We study the dynamics of price impact, we perform an out-of-sample exercise, and we investigate the relation between price impact and market depth in section 6. Section 7 concludes. Technical details and further investigations are confined to the appendix.

2. Literature review

Market liquidity is one of the key characteristics of financial markets. Given its relevance, both from a theoretical (e.g. for price formation (Datar *et al.* 1998, Chung and Chuwonganant 2018)) and a practical perspective (e.g. for the optimal liquidation of institutional orders (Bertsimas and Lo 1998, Almgren and Chriss 2001)), an extensive body of literature has been written in the effort to define, measure, and understand liquidity (see Foucault *et al.* 2013, for an authoritative introduction on the subject and its ramifications). Also, market liquidity is a slippery and elusive concept with various dimensions. One of them, which is the focus of this paper, is price impact (known also as ‘depth’ or ‘resiliency’ in the language of Kyle (1985)), i.e. the reaction of prices to trades. In his foundational paper, Kyle (1985) derives an equilibrium solution in a framework where price fluctuations and traded volumes are linearly related. Moreover, he shows that the reciprocal of the coefficient of this linear relationship (i.e. the notorious Kyle's lambda) is a measure of the depth of the market. Since then, several models of market impact have been proposed (cfr. Glosten and Milgrom 1985, Glosten and Harris 1988, Madhavan *et al.* 1997, Bouchaud *et al.* 2004, Hasbrouck 2007, Bouchaud *et al.* 2009, Cont *et al.* 2014, among many others).

Although price impact can be derived from an equilibrium solution, it can be seen also as the result of the arrival and cancellation of orders in the market. Over the last two decades, the increasing availability of high quality granular data has motivated a stream of studies focused on the modeling and measurement of the price impact of orders. The impact of market orders has been extensively investigated in Lillo *et al.* (2003) and Farmer *et al.* (2004), where the connection between the state of the LOB and price fluctuations has been elucidated. Moreover, it has been shown that limit orders and cancellations also have a significant price impact (Eisler *et al.* 2012, Hautsch and Huang 2012).

Our model builds on that of Cont *et al.* (2014). The main innovations of our model with respect to theirs are the following: (i) the price impact coefficient is considered a latent and dynamical variable; (ii) we are able to disentangle and quantify the contribution of the different components (daily component, time-of-the-day pattern and intra-day variance) of the price impact coefficient estimates; (iii) we can describe the statistical properties of the stochastic intraday component $q_{i,\ell}$, such as its persistence. We point out that a previous version of this model appeared in a preprint written by one of us. Here we add several changes to that model, the most important one is the disentangling of the periodic intraday component, which is responsible of a significant fraction of the persistence in price impact. Our model is broadly in line with the class of History Dependent Impact Models (HDIM) (Lillo and Farmer 2004, Eisler *et al.* 2012) in which the price impact is permanent but variable. Indeed, although in our methodology the price impact is not explicitly history dependent (i.e. it does not depend explicitly on previous values of the OFI), its dynamics is filtered out through the use of Kalman Filter (see section 5.1). As a consequence of this procedure, it becomes dependent on previous values of the OFI.

Finally, we notice that the assumption of time varying liquidity is consistent with various theoretical and empirical works in the literature (cfr.,for instance Saar 2001, Chiyachantana *et al.* 2004, Pereira and Zhang 2010) for which price impact varies according to the underlying economic environment. Moreover, other authors in the financial literature postulate an auto-regressive dynamics for the price impact coefficient (Amihud and Mendelson 1986, Pástor and Stambaugh 2003, Acharya and Pedersen 2005, Pereira and Zhang 2010). However, while models for high-frequency intra-day conditional variance of financial returns are well-explored (cfr. for instance Andersen and Bollerslev 1997, Andresen and Bollerslev 1998, Bollerslev *et al.* 2000, Giot 2005, Engle and Sokalska 2012, Stroud and Johannes 2014, and references therein), to the best of our knowledge, price impact models in which the price impact has an its own (unobserved) dynamics have not been subject of deep investigations. The main contribution of our study is the introduction of an enhanced methodology for the characterization of security price dynamics within the state-space models framework. Broader finance applications would include liquidity effects on asset pricing, optimal trading strategies, or market design. Still, the latter applications may represent an interesting direction for future research.

3. Data

We use data collected from the NASDAQ Historical TotalView-ITCH, and extracted through LOBSTER (Huang and Polak 2011). The dataset contains the history of all trades, orders and cancellations submitted at the best quotes to the NASDAQ stock exchange during standard trading hours†

On the NASDAQ platform each stock is traded in a separate LOB with price-time priority and a tick size of $\delta = \$\,0.01$. Although this tick size is the same for all stocks, the prices of different stocks vary across several orders of magnitude. As it is customary within the financial literature, we refer to *large-tick stocks* if the ratio between δ and the average stock price (i.e. the *relative tick size*) is large and to *small-tick stocks* if this quantity is small. The framework of Cont *et al.* (2014) is particularly suitable for large-tick assets because small tick stocks are characterized by small queues and sparse LOB (Farmer *et al.* 2004, Lillo and Farmer 2005). Thus all our developments are illustrated on five specific examples of large-tick stocks from the NASDAQ 100 index: Microsoft Corp. (MSFT), Comcast Corp. (CMCSA), Intel Corp. (INTC), Cisco Systems Inc. (CSCO), Apple Inc. (AAPL). The relative tick sizes span between 4.27 and 1.36 basis points. Dates range from January 4th, 2016, to June 30, 2016, for a total of 125 days. These stocks have been chosen as a result of a well-defined selection process. First, we ranked the NASDAQ 100 index by average market capitalization. Then, we ranked the top 20 stocks (i.e. the largest 20 capitalization) by relative tick size. Finally, the top five stocks from the list have been selected for the main empirical investigation performed in this work. Whereas the bottom five have been selected as 'small tick' stock; the appendix discusses the results for small-tick stocks which have a relative tick size ranging between 0.36 and 0.06 basis points. The average mid-price of each stock along with its standard deviation over the sample period is reported in the third column of table 1. The fifth and seventh columns of table 1 outline time-weighted bid-ask spreads in dollars and as a percentage of the prevailing quoted midpoint.‡ Table 2 presents a more detailed summary statistics for the LOB. Quantities characterizing the latter (i.e. limit orders, market orders and cancellations) are reported

† NASDAQ standard trading hours are between 09:30 and 16:00 EST.

‡ We note, however, that spreads calculated on displayed liquidity may overestimate the effective spreads actually paid or received due to non-displayed orders. Remarkably, on NASDAQ non-displayed orders are not visible until executed.

Table 1. Descriptive statistics of investigated stocks over the sample period.

Stock	Ticker	Mid-price Avg.	Mid-price Std	Spread Avg.	Spread Std	Relative spread Avg.	Relative spread Std
Cisco Systems Inc.	CSCO	26.910	1.870	0.011	0.000	4.27	0.35
Intel Corp.	INTC	30.994	1.209	0.012	0.000	3.74	0.21
Microsoft Corp.	MSFT	52.150	1.819	0.012	0.000	2.35	0.10
Comcast Corp.	CMCSA	59.779	2.896	0.013	0.000	2.10	0.14
Apple Inc.	AAPL	99.502	5.463	0.013	0.001	1.36	0.11

Note: The sample period is from January 1st, 2016, to June 30th, 2016. Mid-price and Spread are reported in dollar unit ($). The Relative Spread is reported in basis point unit. Stocks are sorted by average price (or by spread), i.e. inversely by relative tick size.

both in terms of number of events (# Ev.) and of average volume (Avg. Vol.). We observe significantly more activities in limit orders and cancellations than in market orders. Moreover the activity on the ask and the bid side of the LOB is quite symmetric. Notice that all averages in tables 1 and 2 are taken over the period considered in our data and are measured in event-by-event time-scale.

4. The model of Cont et al. (2014)

First, for sake of completeness and for fixing notations, we review the model of Cont *et al.* (2014). In particular, we recall both the notion of order flow imbalance and price impact coefficient.

Cont *et al.* (2014) introduced a stylized model of the LOB leading to a simultaneous relation between order flow and price change. More precisely, consider an equally spaced partition $I_\ell = [t_{\ell-1}, t_\ell)$ of length Δ_ℓ of the time interval $[0, T]$ (i.e. a trading day) and denote respectively by L_ℓ^b, C_ℓ^b and M_ℓ^s the total number of shares of limit buy orders, cancellation of buy orders, and market sell orders that has occurred at the bid price within I_ℓ. Similarly, L_ℓ^s, C_ℓ^s, and M_ℓ^b represent the total number of shares of limit, cancellation and market orders that affected the ask price within I_ℓ. Finally, let P_ℓ^b, P_ℓ^s and let P_ℓ be the bid, the ask and the mid price at time t_ℓ, expressed as multiples of the tick size δ. Cont *et al.* (2014) define a stylized LOB whose dynamics are driven by the following rules: (1) the number of shares at each price level of the order book beyond the best bid and ask is given by D. (2) limit, market, and cancellation orders occur only at the best bid and ask prices. (iii) when the bid size reaches D the bid price moves upwards of one tick, when it reaches 0 the bid price moves downwards of one tick. Specular rules hold for the ask price.

Under these assumptions, authors show that the mid price change is determined by

$$\Delta P_\ell = \beta_\ell \,\mathrm{OFI}_\ell + \epsilon_\ell \qquad \epsilon_\ell \sim \mathrm{NID}\left(0, \sigma_\epsilon^2\right) \tag{4}$$

where $\mathrm{OFI}_\ell \equiv M_\ell^b + L_\ell^b - C_\ell^b - M_\ell^s - L_\ell^s + C_\ell^s$ is the *order flow imbalance* and $\beta_\ell \equiv 1/(2\,D_\ell)$ is the *price impact coefficient* relative to the ℓ-th half-hour time window. In their work, authors assume that β_ℓ is constant over each half-hour interval and estimate a separate $\hat{\beta}_\ell$ via ordinary least squares (OLS) regression in each 30-min window. To obtain a proxy of the intraday pattern of market impact, they average $\hat{\beta}_\ell$ for each half-hour interval across the days.

4.1. The Cont et al. (2014) model on different time scales

The estimation of the model proposed by Cont *et al.* (2014) relies on two distinct time scales. The first one, Δ_K, is the time interval over which a single β is estimated from the linear model, and the second one, Δ_ℓ, with $\Delta_\ell < \Delta_K$, is the frequency at which single realizations of price changes and order flow imbalance are sampled. We use the following notation: days in our sample are indexed by $i\,(i = 1, \ldots, N)$ and the trading day is divided in multiple non-overlapping intervals (bins onwards) indexed by $\ell\,(\ell = 0, \ldots, L)$.

In the original paper, authors divided each day in 13 non-overlapping intervals of $\Delta_K = 30$ min and considered time-series of ΔP and OFI with sampling frequency $\Delta_\ell = 10$ s. Since we are interested in modeling liquidity at a significantly higher frequencies (1 min), we conduct an empirical analysis in order to test the performance of the Cont *et al.* (2014) model on different time scales. We estimate (4) for different values of Δ_K and, for a fixed Δ_K, for different values of Δ_ℓ. Qualitatively, the results obtained from different stocks are highly comparable. This allows us to display MSFT as a representative of the entire pool.

The left panel of figure 1 shows the estimated value of β when Δ_K is fixed and equal to 30 min and Δ_ℓ varies from 300 to 0.5 s. The right panel, instead, displays a contour plot of the coefficients of determination as a function of the length (in seconds) of the regression window Δ_K and of the sampling frequency Δ_ℓ. Reported statistics have been averaged across 100 days, starting from January 4, 2016. The following observations are in order. The left panel shows that the average value of β weakly depends on Δ_ℓ. A mildly divergence from this behavior occurs when Δ_ℓ is equal to 5 min. On the other hand, as discussed in Cont *et al.* (2014), the R^2 strongly depends on Δ_ℓ (see the right panel). For a fixed Δ_K, R^2 is a non monotonic function of Δ_ℓ: it initially increases and reaches a maximum around $\Delta_\ell = 10 - 30$ s, which is the time scale used in Cont *et al.* (2014). Moreover, it is a monotonically decreasing function of Δ_K for a fixed Δ_ℓ.

This said, figure 1 indicates that at short time scales the OFI regression is less accurate, thus suggesting that the estimation of the price impact at a higher frequencies is likely to be quite noisy. To overcome this problem, we propose a dynamical model of price impact (see section 4), where β is a latent

Table 2. Main sample statistics of the limit order book averaged over the sample period.

Symbol	Bid $Q_{1,\ell}$		Ask $Q_{1,\ell}$		Bid $Q_{1,c}$		Ask $Q_{1,c}$		Bid $Q_{1,m}$		Ask $Q_{1,m}$	
	#Ev.	Avg. Vol.	#Ev.	Avg. Vol.	# Ev.	Avg. Vol.	#Ev.	Avg. Vol.	#Ev.	Avg. Vol.	#Ev.	Avg. Vol.
CSCO	83746.38	1004.59	85574.85	1018.82	69306.98	809.28	71667.87	823.32	7481.11	54.04	10983.63	70.47
INTC	107735.19	1027.75	109450.99	1052.25	89663.46	824.30	91820.30	849.24	9429.46	60.24	12460.66	76.54
MSFT	190464.74	1776.54	190903.58	1755.47	154535.32	1351.72	156275.93	1355.95	18022.72	141.13	23296.63	196.48
CMCSA	90841.02	715.31	91874.22	724.14	74455.99	561.90	76142.54	575.87	9460.17	62.16	11637.26	80.69
AAPL	206612.51	2460.25	210055.42	2507.16	183007.70	2091.56	185826.46	2137.03	21011.42	247.76	25953.06	320.13

Note: The sample period is from January 1, 2016, to June 30, 2016. The amount of limit orders at the best bid (Bid $Q_{1,\ell}$) and ask (Ask $Q_{1,\ell}$), of cancellations (Bid $Q_{1,c}$ and Ask $Q_{1,c}$) and of market orders (Bid $Q_{1,m}$ and Ask $Q_{1,m}$) for each stock is reported. Quantities are characterized in term of both number of events (#Ev.) and average volume (Avg.Vol.) measured in number of shares.

variable estimated through a Kalman filter approach, over intervals of 1 min. Before introducing our model, we perform a standard regression analysis under the following settings: (i) β is estimated over a window $\Delta_k = 60$ s and (ii) time series of OFI and ΔP are sampled with frequency $\Delta_\ell = 0.5$ s. One of the main aim of this exercise is to obtain further insights on the type of dynamical modeling that is appropriate for empirical data. Results are summarized in figure 2.

The top-left panel plots the histogram of the values $\hat{\beta}_i$, $i = 1, \ldots, N$ obtained by averaging, in each generic day i, $\hat{\beta}_{i,\ell}$ over ℓ. The significant variability range (i.e. between 2 and 4) of $\hat{\beta}_{i,\ell}$ calls for the introduction of a parameter that accounts for the average liquidity on a given day i. Therefore, a proper account of the factors governing price impact must incorporate the (average) daily level of price impact. The intuition behind our assertion is that there are some days in which the market is globally more *liquid* than others.

The top-right panel displays the estimated price impact intraday pattern $\hat{\beta}_\ell$, $\ell = 1, \ldots, L$. In the same spirit of Cont *et al.* (2014), we retrieve the pattern by averaging the estimated price impact across the trading days. In agreement with the empirical findings of Cont *et al.* (2014), the price impact exhibits a pronounced intraday periodic component. This calls for a model specification that explicitly takes it into account. A first visual inspection of the intraday pattern of price impact reveals a characteristic shape. The largest price impact occurs at the open trade. Then, it tapers through the interior period gradually, and fell rapidly at the end of the trading day. We will return to this issue later.

Let $\hat{\beta}_{i,\ell}$ be the estimates in (4) relative to day i and to the ℓth 1-min slot. We compute for each intraday 1-min interval the sample mean and the sample standard deviation of the estimates across days and relate these quantities in the scatter plots presented in the middle-left panel of figure 2. The panel clearly shows a positive relationship between these two statistics. This type of relationship, common for families of random variables with positive support (e.g. the Gamma, the log-normal, and the Weibull are parametric families of random variables with this property), calls for a multiplicative process for market impact. Finally, we rescale each $\hat{\beta}_{i,\ell}$ by the intraday pattern; precisely, we divide each $\hat{\beta}_{i,\ell}$ by π_ℓ, $\ell = 1, \ldots, L$. We perform this further step in order to eliminate the risk of accounting for spurious auto-correlations. We plot the ACF of this normalized quantity (see the middle-right panel of figure 2) for a specific day (May 12, 2016). There is evidence of some short-lived dependences. What type of phenomena can lead to these dependences? We interpret this result as being symptomatic of transient phenomena, which are not captured by the *predictable* intraday pattern (cfr. Cont *et al.* 2014) of the price impact. In particular, this phenomena should exhibit different statistical properties with respect to the time-of-the day effect. We claim that these *fluctuations* around the average level of price impact should be unobserved (latent), similarly to what is assumed for the volatility of price returns in the volatility literature.

In summary, we have identified four main dynamical properties that we will incorporate in the model proposed in section 5: (1) price impact varies across days, thus a global

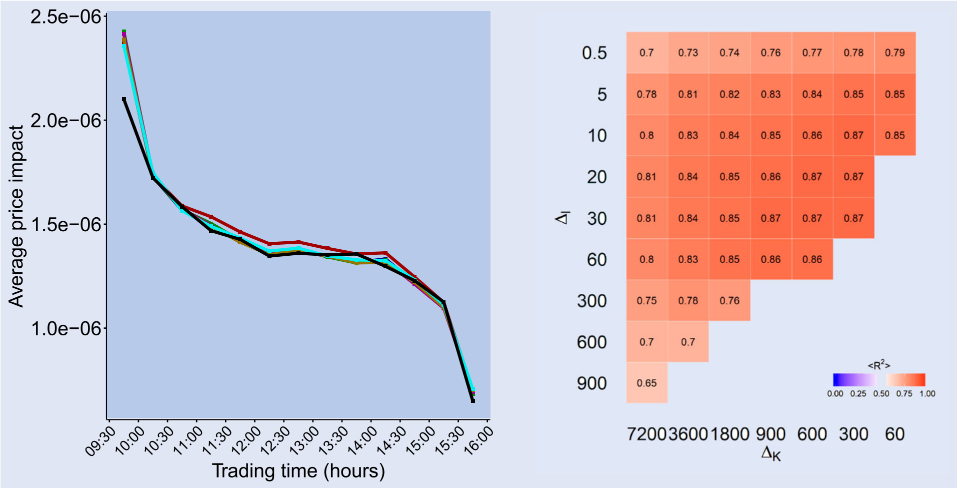

Figure 1. Left panel represents averages across days of the price impact coefficient estimates $\hat{\beta}_k$ as in (4) for MSFT as a function of the time of the day for $\Delta_K = 30$ min and Δ_ℓ equal to: 0.5 s (red), 5 s (blue), 10 s (green), 20 s (magenta), 30 s (gold), 1 min (cyan) and 5 min (black). Right panel represents the contour plot of the averages R^2 across days as a function of the regression window Δ_K and sampling frequency Δ_ℓ.

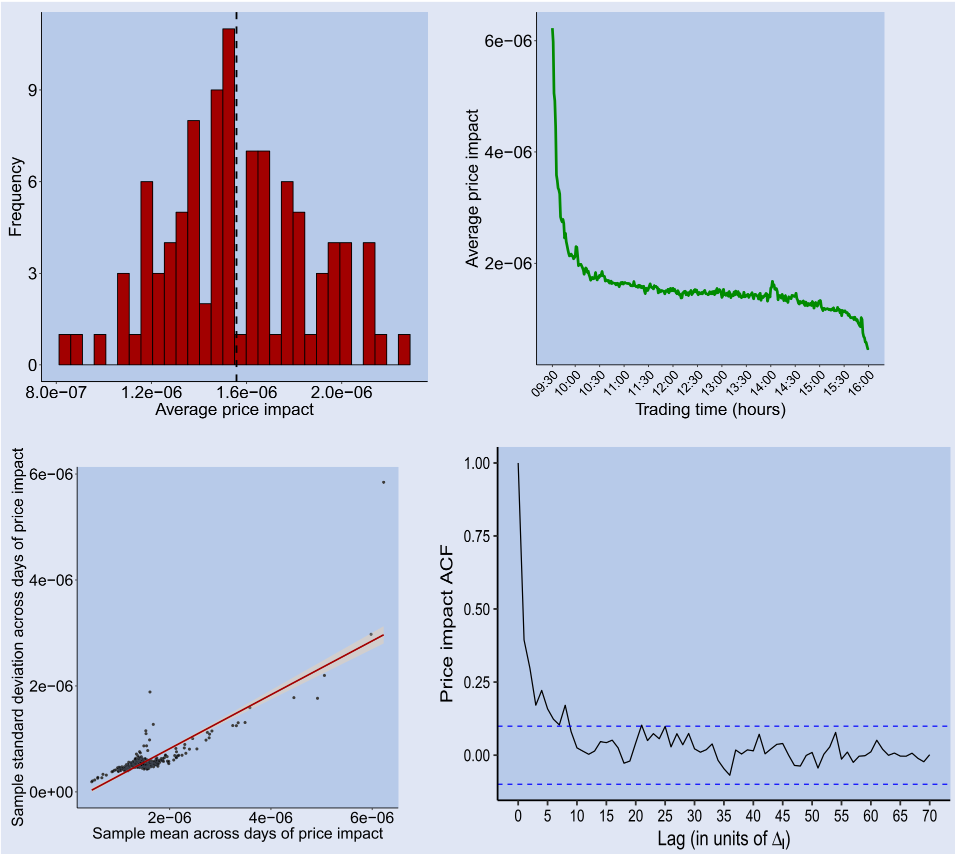

Figure 2. From left to right, from top to bottom: histograms of the average daily level of price impact; estimates of the intraday pattern of price impact coefficient computed according to the procedure of Cont *et al.* (2014) at a time scale of $\Delta_K = 60$ s, $\Delta_\ell = 0.5$ s. Scatter plot of the sample mean versus the sample standard deviation across the days of estimates $\hat{\beta}_{i,\ell}$ computed for each intraday 1-min interval.

level characterizes the average liquidity; (2) there is a pronounced intraday pattern†; (3) there is a linear relation between the mean and standard deviation of price impact, suggesting that the price impact should be modeled by a

† In addition to the previous effects, price impact may exhibit large, sudden shifts around the release of important economic news, such as macroeconomic information (see Andersen and Bollerslev 1998, for a discussion of a similar behavior of *intraday* volatility). Should the price impact also *explicitly* account for economic announcements, one would multiply the price impact coefficient for another factor representing the scheduled announcement effects. This further modeling is beyond the scopes of the present paper, however, we strongly encourage this analysis in future works.

multiplicative process; (4) price impact is temporally (short-range) auto-correlated.

5. A dynamical market impact model

Intraday equity returns (for a day i and time-of-the-day ℓ) are described by the following process:

$$\Delta P_{i,\ell} = \beta_{i,\ell}\, \mathrm{OFI}_{i,\ell} + \epsilon_{i,\ell} \tag{5}$$
$$\beta_{i,\ell} = \beta_i\, \pi_\ell\, q_{i,\ell} \tag{6}$$
$$q_{i,\ell+1} = (1-\rho_i) + \rho_i q_{i,\ell} + \eta_{i,\ell+1}, \tag{7}$$

where

- $\epsilon_{i,\ell}$ and $\eta_{i,\ell}$ are NIID and mutually independent with mean zero and variance respectively equal to $\sigma^2_{i,\epsilon}$ and $\sigma^2_{i,\eta}$.
- The process π_ℓ, which is allowed to change within the day but is otherwise time-invariant across days, captures the diurnal effect.
- β_i is the daily average price impact component on day i, while
- $q_{i,\ell}$ denotes the remaining, potentially correlated, stochastic price impact component which is described through an unobserved (latent) mean-reverting process; notice that the choice of the drift in equation (7) ensures that the unconditional mean of the process $(q_{i,\ell})$ is equal to 1.

Note that, without additional restrictions, the components of equation (6) are not separately identified. Thus we assume that $\sum_{\ell=1}^{L} \pi_\ell = L$. Setting $\alpha_{i,\ell} \equiv q_{i,\ell} - 1$ (i.e. $\alpha_{i,\ell}$ measures the stochastic component of the price impact deviation from the average level), we define a *price impact fluctuation* to be a non-zero conditional expected value of the process $\beta_i \pi_\ell \alpha_{i,\ell}$. The dynamics in equation (5) can be classified as a latent variable model and its estimation can be performed by using the Kalman filter, as described in section 5.1. We remind that details on how we set the initial values for the parameters and for the process $(q_{i,\ell})$ are given in Appendix 2.

5.1. Estimation

Before showing the estimation procedure, we fix some notations, we introduce the general state–space representation as in Durbin and Koopman (2012) and we derive the corresponding Kalman filter recursions.

We denote the value of a generic variable X (e.g. OFI, price impact coefficient or hyper-parameter) at day i and bin ℓ as $X_{i,\ell}$, $i = 1,\ldots,N$ and $\ell = 1,\ldots,L$. Our methodology is independently applied to each day of the sample, therefore, for the sake of simplicity, the index i will be replaced by $\cdot$. We denote the set of observations of the price adjustment up to time $\ell-1$ with

$$\mathcal{P}_{\cdot,\ell-1} = \{\Delta P_{\cdot,1},\ldots,\Delta P_{\cdot,\ell-1}, \mathrm{OFI}_{\cdot,1},\ldots,\mathrm{OFI}_{\cdot,\ell-1}\}$$

and the hyper-parameters vector with

$$\mathbf{\Theta} = \left(\beta_\cdot, \rho_\cdot, \sigma_{\cdot,\epsilon}, \sigma_{\cdot,\eta}\right).$$

We remind that our proposed model for price impact reads as follows:

$$\Delta P_{\cdot,\ell} = q_{\cdot,\ell}\, \beta_\cdot\, \pi_\ell\, \mathrm{OFI}_{\cdot,\ell} + \epsilon_{\cdot,\ell} \qquad \epsilon_{\cdot,\ell} \sim \mathrm{NID}\left(0, \sigma^2_{\cdot,\epsilon}\right) \tag{8}$$
$$q_{\cdot,\ell+1} = (1-\rho_\cdot) + \rho_\cdot\, q_{\cdot,\ell} + \eta_{\cdot,\ell} \qquad \eta_{\cdot,\ell} \sim \mathrm{NID}\left(0, \sigma^2_{\cdot,\eta}\right), \tag{9}$$

which has a linear Gaussian state–space representation. This has two important implications: (i) the model can be treated with Kalman filter and (ii) the log-likelihood function can be written in a closed form. Denoting by $\mathsf{q}_{\cdot,\ell\,|\,\ell}$ and $\mathsf{P}_{\cdot,\ell\,|\,\ell}$ the conditional mean and variance of $q_{\cdot,\ell}$ given $\mathcal{P}_{\cdot,\ell}$, and by $\mathsf{q}_{\cdot,\ell+1}$ and $\mathsf{P}_{\cdot,\ell+1}$ the conditional mean and variance of $q_{\cdot,\ell+1}$ given $\mathcal{P}_{\cdot,\ell}$, the Kalman filter recursions can be written as follows (see, e.g. Durbin and Koopman 2012):

$$\begin{aligned}
&v_{\cdot,\ell} = \Delta P_{\cdot,\ell} - Z_{\cdot,\ell}\,\mathsf{q}_{\cdot,\ell}, \qquad Z_{\cdot,\ell} = \beta_\cdot\, \pi_\ell\, \mathrm{OFI}_{\cdot,\ell}\\
&F_{\cdot,\ell} = Z^2_{\cdot,\ell}\,\mathsf{P}_{\cdot,\ell} + \sigma^2_{\cdot,\epsilon} \qquad \mathsf{q}_{\cdot,\ell\,|\,\ell} = \mathsf{q}_{\cdot,\ell} + \mathsf{P}_{\cdot,\ell}\, Z_{\cdot,\ell}\, F^{-1}_{\cdot,\ell}\, v_{\cdot,\ell}\\
&\mathsf{P}_{\cdot,\ell\,|\,\ell} = \mathsf{P}_{\cdot,\ell} - \left(\mathsf{P}_{\cdot,\ell}\, Z_{\cdot,\ell}\right)^2 F^{-1}_{\cdot,\ell}\\
&\mathsf{q}_{\cdot,\ell+1} = \rho_\cdot \mathsf{q}_{\cdot,\ell\,|\,\ell} + (1-\rho_\cdot) \qquad \mathsf{P}_{\cdot,\ell+1} = \rho^2_\cdot \mathsf{P}_{\cdot,\ell\,|\,\ell} + \sigma^2_{\cdot,\eta}
\end{aligned} \tag{10}$$

Once $\mathsf{q}_{\cdot,\ell}$, $\mathsf{P}_{\cdot,\ell}$, $\mathsf{q}_{\cdot,\ell\,|\,\ell}$ and $\mathsf{P}_{\cdot,\ell\,|\,\ell}$ are retrieved, one can compute the Kalman smoothing equations by backward recursion. Precisely, we have (Durbin and Koopman 2012)

$$\begin{aligned}
\mathsf{q}_{\cdot,\ell\,|\,L} &= \mathsf{q}_{\cdot,\ell\,|\,\ell} + J_{\cdot,\ell}\left(\mathsf{q}_{\cdot,\ell+1\,|\,L} - \mathsf{q}_{\cdot,\ell+1}\right)\\
\mathsf{P}_{\cdot,\ell\,|\,L} &= \mathsf{P}_{\cdot,\ell\,|\,\ell} + \left(J_{\cdot,\ell}\right)^2\left(\mathsf{P}_{\cdot,\ell+1\,|\,L} - \mathsf{P}_{\cdot,\ell+1}\right)\\
J_{\cdot,\ell} &= \mathsf{P}_{\cdot,\ell\,|\,\ell} F^{-1}_{\cdot,\ell} \mathsf{P}_{\cdot,\ell+1}
\end{aligned} \tag{11}$$

Given the Kalman filter recursions in (10), the log-likelihood function

$$\log L\left(\Delta P_{\cdot,1},\ldots,\Delta P_{\cdot,L}\right) = \sum_{\ell=1}^{L} \log p\left(\Delta P_{\cdot,\ell}\,|\,\mathcal{P}_{\cdot,\ell-1}\right)$$

can be computed in the prediction error decomposition form:

$$\log L\left(\Delta P_{\cdot,1},\ldots,\Delta P_{\cdot,L}\right) = -\frac{L}{2}\log 2\pi - \frac{1}{2}\sum_{\ell=1}^{L}\left(\log F_{\cdot,\ell} + \left(v_{\cdot,\ell}\right)^2 F^{-1}_{\cdot,\ell}\right).$$

As the diurnal pattern is not practically observable, we propose a 'two-stage' approach. In the first stage, the diurnal factor is pre-estimated. In the second stage, the estimator is included in the estimation procedure. In the conditional variance literature, there are several methodologies that capture the intraday unobservable pattern. For instance, Andersen and Bollerslev (1997) and Andresen and Bollerslev (1998) employ Fourier transform techniques, whereas (Müller *et al.* 2011) use functional data analysis methods.

Instead, the approach that we propose in this work can be related to the one introduced by Engle and Sokalska (2012)†.

In our setting, the daily price impact component β_i belongs to the hyper-parameters vector $\mathbf{\Theta}$ (i.e. it is an output of the estimation procedure). To capture π_ℓ, $\ell = 1, \ldots, L$, we proceed as follows. First, we set $\pi_\ell = 1$ for each $\ell = 1, \ldots, L$ and we run the Kalman filter for each day i in our sample to obtain a first estimation of both the hyper-parameters vector, say $\Theta^{(\mathrm{I})}$, and the smoothed state process, say $\mathsf{q}^{(\mathrm{I})}_{i,\ell\,|\,L}$. However, at this stage, the latter incorporates also the diurnal price impact pattern in a multiplicative way. Therefore, the estimated diurnal pattern $\hat{\pi}_\ell$ is obtained by computing the sample mean of $\mathsf{q}^{(\mathrm{I})}_{i,\ell\,|\,L}$ across days i.

In the second step, $\widehat{\pi}_\ell$ is embedded into the Kalman filter recursions in (10), thus leading to the final estimates.

6. Econometric and modeling issues

In this section, we discuss both the modeling and the statistical properties of our model.

6.1. *Exogeneity of the OFI*

The OFI introduced in Cont *et al.* (2014) provides a simple tool to account for both the price impact of trades, i.e. market orders, and orders, limit order placements and deletions. The authors proposed a simple linear regression model of the relationship between OFI and price fluctuations which successfully captures multiple empirical regularities of intraday market liquidity. However, the extreme simplicity of their model limits the study of this relationship to the contemporaneous effects between the two variables. It is worth asking whether this is the case even at 1-min time horizon, i.e. whether the data, at this time-scales, require the OFI to be endogenous. If this were the case, it would be necessary to model the joint dynamics of price changes and OFI. In what follows, guided by the seminal paper of Hasbrouck (1991), we apply a vector auto-regressive (VAR) model as a validation of the exogeneity assumption. In particular, we introduce the following model‡:

$$\begin{bmatrix} 1 & b_0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta P_\ell \\ \mathrm{OFI}_\ell \end{bmatrix} = \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} \begin{bmatrix} \Delta P_{\ell-1} \\ \mathrm{OFI}_{\ell-1} \end{bmatrix} + \begin{bmatrix} v_{1,\ell} \\ v_{2,\ell} \end{bmatrix}, \qquad (12)$$

where the immediate effect of the OFI on price fluctuations is measured by the b_0 coefficient. The goal is to test whether at the time scale of 1 min a joint dynamics for OFI and price changes is necessary. In line with Hasbrouck (1991), v_1 and v_2 are jointly and serially uncorrelated with zero mean. We assert that the settings of the model introduced in (12) capture the fact that the price fluctuations and OFI are, by construction, not determined simultaneously: the change in the mid-price follows the last update of the OFI whereas ΔP_ℓ cannot contemporaneously influence the OFI. In addition, we truncate the potentially infinite order of the VAR representation at one lag. The purpose of the present analysis is to assess whether some lagged effects exist even at a relatively low (in a micro-structure perspective) frequency (1 min). We assert that the eventual presence of serial dependencies is likely to be caused by some micro-structure imperfections, such as price discreteness and delays in the reception of newly released information.

Table 3 displays the regression coefficients associated with contemporaneous and lagged effect along with the corresponding t-statistics (between brackets). We report averages of the estimated parameters over the sampling period. Also, we check that the reported results are in line with day-by-day estimates. We observe that the contemporaneous effect of the OFI (i.e. $-b_0$) is positive and significant. On the other hand, the lagged effects are not significant with the notable exception of AAPL. This analysis supports the conjecture that there is a significant instantaneous dependence of price fluctuations on the OFI, but the latter does not display any significant auto-correlation and therefore it is reliable to consider it as exogenous. Ultimately, this finding confirms the validity of the adopted dynamics even at a sampling frequency of 1 min.

6.2. *Is Kalman the right filter?*

From an estimation perspective, the appealing feature of the proposed model is that it can be estimated via Kalman filter, and part of the appeal of Kalman filter comes from its robustness. It is well known (see, for instance Hamilton 1994) that it is the optimal filter (minimum mean-square error) if the innovations are normal. In (8) and (9) the disturbances are *i.i.d.* normal. The normality assumption in (8) is not justified because of the discrete nature of price changes at 1-min time horizon. In particular, the disturbances have to be uncorrelated with the regressors, so, at first sight, normality seems to be implausible. We address this concern explicitly and we quantify the severity of the misspecification. Precisely, we follow the literature (Durbin and Koopman 2012) and we look, for any generic day i, at the time series of the standardized one-step ahead forecast errors: $\hat{\epsilon}_{\cdot,\ell} = \hat{v}_{\cdot,\ell}/\sqrt{\widehat{F}_{\cdot,\ell}}$, $\ell = 1, \ldots, L$ obtained after the estimation of the model. They should be *i.i.d*. normally distributed. In appendix 4, we show sample autocorrelations of $\hat{\epsilon}_{\cdot,\ell}$ and $\hat{\epsilon}^2_{\cdot,\ell}$, as well as quantile–quantile plots for $\hat{\epsilon}_{\cdot,\ell}$. We only note some very weak heteroskedasticity of $\widehat{\epsilon}_{\cdot,\ell}$ and a slight deviation from normality in the tails. We argue that the proposed framework represents an optimal trade-off between a model with a perfect statistical fit of the data and a model that have clear advantages in terms of practical implementability.

6.3. *Statistical issues in the two-step estimation procedure*

In this section, we discuss the statistical properties of the two-step estimator of the model introduced in section 5. The

† Engle and Sokalska (2012) propose a model for high-frequency intraday financial returns, $r_{i,\ell}$. They decompose the conditional variance of the latter as a multiplicative product of daily, h_i, diurnal, s_ℓ, and stochastic, $q_{i,\ell}$ with $\mathbb{E}\left[q_{i,\ell}\right] = 1$, intraday volatility. To proxy the daily variance component h_i, authors use commercially available volatility forecast. Thus once a proxy for h_i is obtained, they simply calculate s_ℓ as the variance of returns in each bin after deflating by the daily variance.

‡ Note that for the sake of consistency the following VAR is conducted in calendar time.

Table 3. VAR parameter estimates. Coefficients which result to be significant at 10%, 5% and 1% confidence levels are marked, respectively, with one, two and three stars.

	Parameters' values (*t*-stats)				
Ticker	b_0	a_1	b_1	c_1	d_1
CSCO	− 2.59e − 06	0.040	− 5.61e − 08	− 27661.86	0.02
	(− 26.85***)	(0.17)	(0.005)	(− 1.19)	(0.34)
INTC	− 5.50e − 07	− 0.005	1.75e − 08	− 124098.13	0.04
	(− 28.81***)	(− 0.025)	(0.14)	(− 0.84)	(0.50)
MSFT	− 1.23e − 06	0.054	− 9.18e − 08	− 65617.70	0.08
	(− 30.83***)	(0.27)	(− 0.23)	(− 1.02)	(1.00)
CMCSA	− 1.03e − 04	− 0.012	2.115e − 08	− 15320.32	0.034
	(− 12.05***)	(− 0.15)	(0.05)	(− 0.86)	(0.48)
AAPL	− 2.94e − 06	0.135	− 8.86e − 07	− 38559.31	0.25
	(− 22.98***)	(0.68)	(− 0.90)	(− 1.48*)	(2.23**)

proposed estimation methodology relies on a two-step procedure. We are aware of the facts that errors emerging in the first phase might propagate through the second stage. We perform a numerical experiment in order to address this statistical concern explicitly. Precisely, appendix 5 reports a numerical exercise based on Monte Carlo simulations confirming that embedding π_ℓ into the Kalman filter recursion has no discernible impact in the estimation.

7. Empirical results

We estimate the model in (1)–(3) by applying the procedure presented in section 5.1 to the dataset described in section 3. We recall that our dataset contains the order book activity of 5 NASDAQ 100 Index stocks over 125 days. We split each trading day in $L = 390$ non overlapping time windows from which we sample the mid-price change and OFI time series as described in Appendix 1. We divide the available data into a training set of 100 days and a test set of the subsequent 25 days.

7.1. In-sample analysis

The analysis on the training set is carried out by estimating the model via MLE through the two-steps procedure discussed in section 5.1. Hyper-parameter estimates of the training set are used to construct the empirical distribution of the hyper-parameters vector $\boldsymbol{\Theta}$. Table 4 reports hyper-parameter estimates after the second stage. Overall, estimates are statistically significant and their inter-quantile ranges indicate that estimates have (on average) a small statistical dispersion.

We now focus on the parameter ρ describing the persistence of price impact. Coherently with the preliminary descriptive investigation presented in section 4.1 and with the numerical exercise displayed in Appendix 5, we note that the proposed two-steps procedure successfully captures and filters out the auto-regressive stochastic dynamics from the intraday pattern. More precisely, the auto-regressive coefficient ρ at stage 2 is significantly smaller than the one estimated at stage 1 (shown in table 5). The difference highlights that ignoring the pervasive diurnal effect of the market impact would turn the process $(q_\ell)_{\ell=1,\dots,L}$ in an almost integrated process (cf. table 5). In particular, this would support the statement that the price impact is *strongly* auto-correlated. When the intraday pattern is properly taken into account, the value of ρ significantly decreases. Nonetheless, estimates remain statistically significant. The reduction of the auto-correlation coefficient has notable implications for the persistence of liquidity fluctuations. For instance, while values for ρ around 0.98 and 0.99 represent a process with a half-life ranging between 30 min and 1 h, values of ρ below 0.5 suggest that the process has a half-life of 1 minute. Thus the liquidity fluctuations predicted by our model are, on average, short lived (but significant).

Figure 3 displays the estimates $\beta_{\cdot,\ell}$, $\ell = 1, \dots, L$, averaged across days. From left to right panels show the price impact process for four stocks (CMCSA, MSFT, INTC, CSCO, AAPL). The dynamics of price impact is interesting and presents a clear intraday pattern. It peaks right after the opening auction, then it quickly decreases to lower levels where it hovers around an average value for most of the trading day and, finally, it decreases again right before the closing auction. The shape of the pattern has an intuitive explanation: the high level of market impact at the beginning of the day is likely to be the consequence of heterogeneity of traders' opinions and adverse selection. The low levels observed at the end of the day can be linked to the market practice of requesting the completion of a parent order within the trading day, a phenomenon that increases the available liquidity right before the closing auction†.

Figure 4 provides a closer look at the dynamics of the price impact process. It combines Kalman smoothed estimates of price impact (black line), Kalman filtered estimates of price impact (green line) and averages across days of the price impact process $\beta_{\cdot,\ell}$, $\ell = 1, \dots, L$, (blue line) in specific days and normalized by the average daily level. Panel A displays CSCO on January 27th 2016, Panel B AAPL on June 15th, 2016, and Panel C INTC on April 27th, 2016. In these days three different meetings of the Federal Open Market Committee (FOMC) have taken place. The behavior of the price impact significantly diverges from the average price impact around 1400 EST, which is the time in which economic information is released, showing how investors promptly react to

† We point out a recent article on the topic, 'The 30 min that have an outsized role in US stock trading', recently appeared (April 24,2018) on the *Financial Times*.

Table 4. Summary of estimation results of model in (1)–(3) on order book data of five NASDAQ 100 Index stock over 100 days in 2016. For each parameter we report: (1) the median (second column), (2) the mean (third column), (3) the average *t*-statistic (fourth column), (4) the lower and the upper quartile.

Ticker	Median	Mean	Average *t*-statistic	Q_1	Q_3
Parameter: β					
CSCO	3.883×10^{-5}	3.835×10^{-5}	41.07	3.414×10^{-5}	4.368×10^{-5}
INTC	6.455×10^{-5}	6.515×10^{-5}	35.06	5.285×10^{-5}	7.979×10^{-5}
MSFT	1.404×10^{-4}	1.443×10^{-4}	40.95	1.213×10^{-4}	1.613×10^{-4}
CMCSA	3.022×10^{-4}	3.124×10^{-4}	37.36	2.576×10^{-4}	3.522×10^{-4}
AAPL	3.392×10^{-4}	3.472×10^{-4}	46.32	2.974×10^{-4}	4.015×10^{-4}
Parameter: ρ					
CSCO	0.3323	0.3505	11.62	0.1167	0.5084
INTC	0.2714	0.3553	12.13	0.1232	0.5271
MSFT	0.4016	0.4603	14.10	0.2419	0.7513
CMCSA	0.3287	0.3855	11.57	0.147	0.6463
AAPL	0.4122	0.4262	14.01	0.1700	0.6203
Parameter: σ_ϵ					
CSCO	0.0035	0.0036	17.52	0.0027	0.0045
INTC	0.0045	0.0050	20.21	0.0039	0.0056
MSFT	0.0096	0.0115	19.34	0.0080	0.0133
CMCSA	0.0103	0.0135	15.05	0.00870	0.0150
AAPL	0.0274	0.0316	20.58	0.0200	0.0400
Parameter: σ_η					
CSCO	0.2547	0.3364	10.29	0.1958	0.3341
INTC	0.2561	0.2941	16.36	0.1942	0.3611
MSFT	0.2752	0.2666	21.06	0.1736	0.3496
CMCSA	0.3091	0.2981	18.79	0.2002	0.3758
AAPL	0.4089	0.3753	24.64	0.2889	0.4639

Table 5. Estimation results for the parameter ρ after the first stage. We report: (1) the median (second column), (2) the mean (third column), (3) the average *t*-statistic (fourth column), (4) the lower and the upper quartile.

Ticker	Median	Mean	Average *t*-statistic	Q_1	Q_3
Parameter: ρ					
CSCO	0.8671	0.7771	32.73	0.6983	0.9470
INTC	0.9218	0.8513	43.51	0.8491	0.9627
MSFT	0.8880	0.8198	40.76	0.7818	0.9548
CMCSA	0.9163	0.8597	48.07	0.8406	0.9598
AAPL	0.6507	0.6109	20.62	0.3744	0.8758

the announcement by updating their positions. This empirical result suggests that our model successfully captures real time variations of the price impact.

7.2. *Out-of-sample performance*

We now discuss the out-of-sample performance of our model. As a benchmark for the true price impact on a given day we select the *ex-post* smoothed estimates of market impact. Note that Kalman smoothing estimates are computed by using the hyper-parameters estimated from real trading data for the considered day; hence they represent the best *ex-post* estimates of market impact. We perform two out-of-sample exercises.

In the first one, we show that the conditioning on real time information improves the estimate of the LOB liquidity with respect to the one obtained from the deterministic intra-day pattern alone. More precisely, we first estimate our model over an in-sample period of 100 days. Figure 5 presents Kalman filtering performed on out-of-sample data. Precisely, the hyper-parameters values are set equal to the median values of the hyper-parameter distributions computed in the in-sample analysis. Blue lines represent the historical average market impact, green lines the filtered real-time estimation of market impact and, finally, black lines represent the *ex-post* smoothed estimates of market impact. The picture clearly shows that the filter successfully captures the intraday dynamic of market impact. To support our statement, we compute (1) the mean squared error (MSE) of the historical marked impact (resp. of the filtered estimates) with respect to the smoothed one, which is denoted by MSE_1 (resp. by MSE_2) and (2) the *gain* in signal extraction as $(\mathrm{MSE}_1 - \mathrm{MSE}_2)/\mathrm{MSE}_1$. Table 6 reports the average gain (Average Gain 1) across the 25 out-of-sample days†. On average, real-time monitoring reduces the MSE of market impact by a factor ranging between $0.22 - 0.83$, thus confirming a significant reduction in the MSE when the price impact is dynamically filtered.

In the second one, we show the importance of the filtering procedure by comparing our model with the static model of Cont *et al.* (2014) (the yellow line in figure 6). For our model, we use the estimates over the in-sample period of 100 days to obtain (an estimation of) the diurnal pattern. Note that in our model the latter is allowed to change within the day but it is otherwise time-invariant across days. Then, for each of the subsequent 25 days we compute the mean squared

† Daily values for MSE_1 and MSE_2 and gain are available upon request.

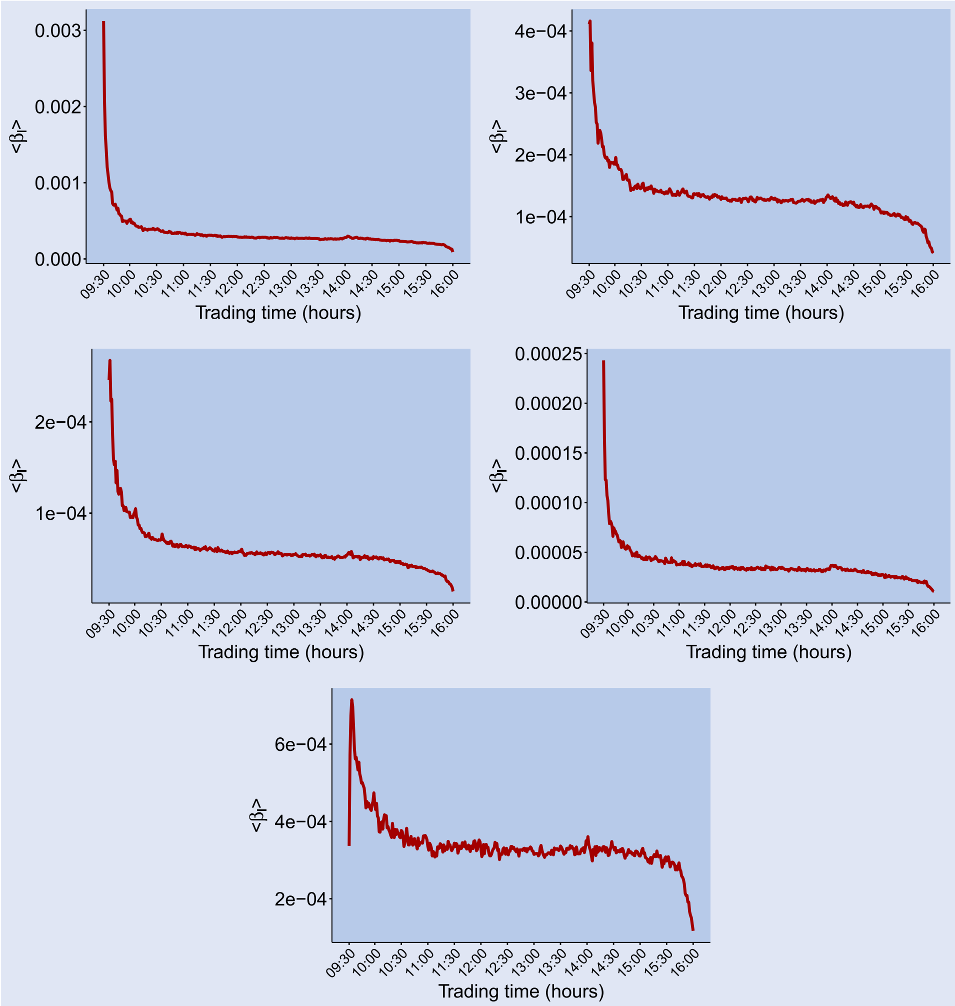

Figure 3. (*From left to right, from top to bottom*) Price impact process $\beta_{\cdot,\ell}$, $\ell = 1, \dots, L$, averaged across days for CMCSA, MSFT, INTC, CSCO, AAPL.

error of the filtered estimates of price impact with respect the smoothed one (MSE_2). The diurnal pattern used in all the out-of-sample period is the one estimated over the in-sample period, i.e. it will not be updated as the experiment proceeds beyond the 100-th day, whereas the values of the hyper-parameters in each day are the one of the day before (e.g. at day 113 we use the ones obtained on day 112). We compare this mean square error with the one of the static model of Cont *et al.* (2014) with respect the Kalman smoothing (MSE_1). More precisely, we estimate in each of the out-of-sample days the model of Cont *et al.* (2014) with a $\Delta_k = 1$ min and $\Delta_\ell = 0.5$ s. It is important to emphasize the difference between the two modeling approaches. The Kalman approach uses the predictive filter, i.e. the expected value of q_ℓ conditional to the past history until $\ell - 1$ and we use for the daily average β_i the daily average price impact of the day before β_{i-1}. On the contrary, for the static model we perform at each interval ℓ the regression (in this sense there is no prediction here) and clearly the level of β will be automatically determined by the estimation. Despite these differences, the second column of table 6, which reports the average gain over the out-of-sample period, shows that the MSE for Kalman filter is smaller than the one for the static approach. By looking at figure 6 it appears that Kalman estimates might have a bias due to the use of the one-day-before level β_i: the average level for impact in a specific day can be quite different from the one of the day before. The static estimates have instead a large variance due to fluctuations, which contributes significantly to the MSE. One could easily improve the performances of the filtering approach, for example by estimating the hyper-parameters (especially β_i) in the first hours of the day and then use it afterwards. This would reduce the above bias, while keeping the variance small. This is beyond the scope of this comparison and we will not be investigated further.

In particular, at this point, one might ask to her/himself what is the correlation between our dynamic measure of the price impact and the static one. To answer this question, we compute the Pearson correlation between the former (i.e. $\beta_{\ell,K} = q_{i,\ell}\beta_i\pi_\ell$, with $q_{i,\ell}$ the filtered estimates) and the static market impact of Cont *et al.* (2014), say $\beta_{\ell,S}$, with a $\Delta_k = 1$ min and $\Delta_\ell = 0.5$ s, for each day i, $i = 1, \dots, 100$, over the in-sample period. Figure 7 displays the box-plots of the measured correlations for each stock in our sample. The two measures of market impact are moderately positively correlated; the correlation coefficient is, in all the cases, around 60%. This

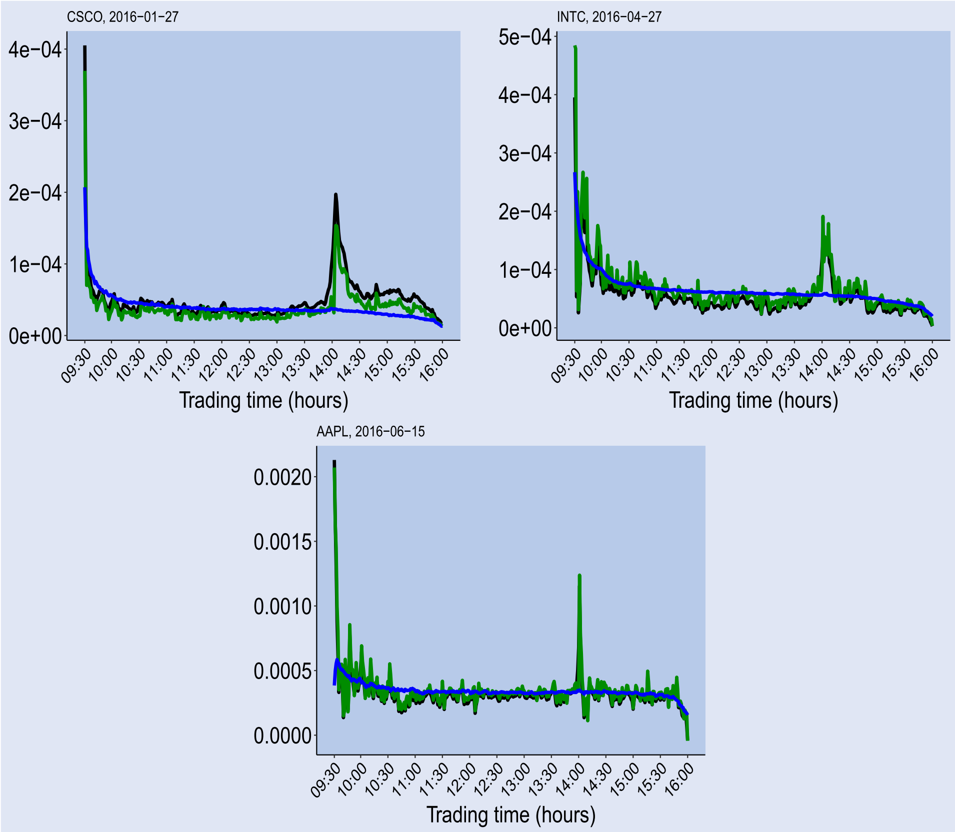

Figure 4. Kalman smoothing estimates of price impact process of CSCO on January 27th, 2016, INTC on April 27th, 2016, and APPL on June 15th, 2016 (black line), together with the correspondent Kalman filtered estimates of price impact (green line) and averages across days over the in-sample period (blue line).

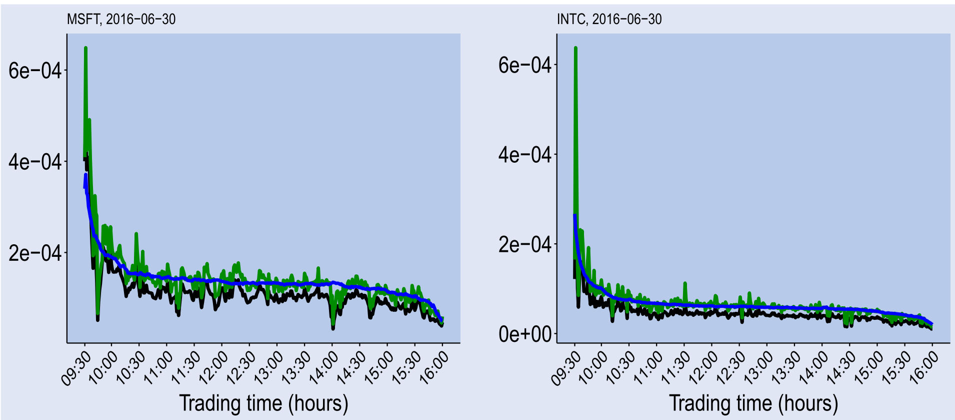

Figure 5. Out-of-sample Kalman filtering of market impact. Panels display three distinct estimates: (1) average market impact (blue), (2) real-time filtered market impact (green) and (3) the ex-post smoothed market impact (black). Left panel: INTC. Right panel: MSFT.

moderate level of correlation might be symptomatic of the fact that in our model price impact becomes history dependent because of the use of the Kalman filter and, as a consequence, is capable of grasping different dimensions of the price impact with respect to those captured by $\beta_{\ell,S}$. We will return on this point when discussing the results in section 6.3, table 7.

We conclude this section by observing that in a recent paper (Xu *et al.* 2019) authors show a reduction of the out-of-sample MSE when comparing their model with the one of Cont *et al.* (2014). Their model is a generalization of the static model of Cont *et al.* (2014) which employs a multi-level OFI (i.e. a vector quantity that measures the net flow of buy and sell orders at different price levels in a (LOB)) instead of the OFI†. The authors of this work performed a comparison also

† More precisely, in order to calculate the out-of-sample RMSE, they use a methodology similar to fivefold cross-validation. For each stock, they first split their data set into five separate folds. For a given fold, they use all the data in the other four folds to fit the parameters of the multi-level OFI via OLS or Ridge regression. Then they calculate the RMSE of the fitted model on these same four folds and they call this the in-sample RMSE. Then they use the same fitted parameters to estimate the RMSE for the other fold (which was not used in the regression fit) and call this the out-of-sample RMSE. They repeat this process for each of the five folds separately, and record the mean out-of-sample RMSE across these five repetitions.

Table 6. Average gain 1 : Average gain across the out-of-sample period (25 days), defined as $(\mathrm{MSE}_1 - \mathrm{MSE}_2)/\mathrm{MSE}_1$, where MSE_1 (resp. MSE_2) is the mean squared error of the historical market impact (resp. of the filtered estimates) with respect to the smoothed market impact. Average gain 2 : Average gain across the out-of-sample period (25 days), defined as $(\mathrm{MSE}_1 - \mathrm{MSE}_2)/\mathrm{MSE}_1$, where MSE_1 (resp. MSE_2) is the mean squared error of the static model (Cont *et al.* 2014) (resp. of the filtered estimates) with respect to the smoothed market impact.

Ticker	Average gain 1	Average gain 2
CSCO	0.40	0.51
INTC	0.28	0.10
MSFT	0.22	0.02
CMCSA	0.83	0.28
AAPL	0.37	0.35

with our model and the performance of their model is in line with ours.

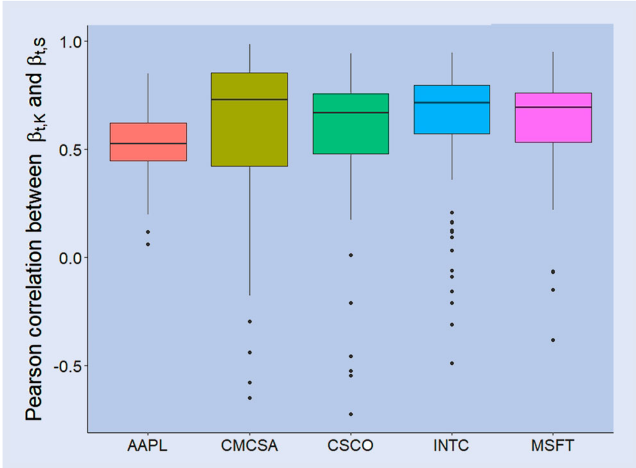

Figure 7. Box-plots of the Pearson correlation between our dynamic measure of the price impact ($\beta_{\ell,K} = q_{i,\ell}\beta_i\pi_\ell$, with $q_{i,\ell}$ the filtered estimates) and the static one ($\beta_{\ell,S}$, i.e. the one of Cont *et al.* (2014) measured with a $\Delta_k = 1$ min and $\Delta_\ell = 1$ s). Pearson correlation is measured for each day i, $i = 1, \ldots, 100$, over the in-sample period.

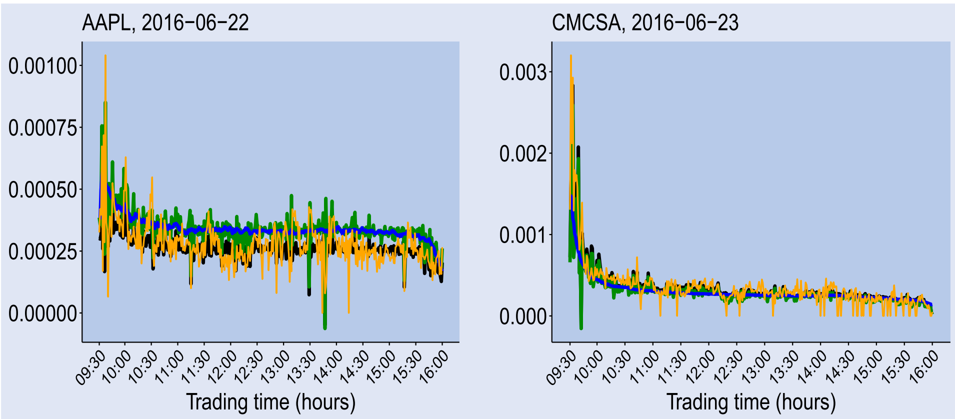

Figure 6. Out-of-sample Kalman filtering of market impact. Panels display four distinct estimates: (1) average market impact (blue), (2) real-time filtered market impact as described in the second exercise (green), (3) the ex-post smoothed market impact (black), (4) the static market impact of Cont *et al.* (2014). Left panel: AAPL. Right panel: CMCSA.

Table 7. Estimates of parameters c and λ in (12) when $\beta_{i,\ell}$ is: (1) estimated employing the methodology of Cont *et al.* (2014) with $(\Delta_K, \Delta_\ell) = (30\,\mathrm{min}, 10\,\mathrm{s})$ (first column) and with $(\Delta_K, \Delta_\ell) = (1\,\mathrm{min}, 0.5\,\mathrm{s})$ (second column); (2) estimated through the Kalman filter (third column).

	Cont *et al.* (2014) (30 min, 10 s)		Cont *et al.* (2014) (1 min , 0.5 s)		Kalman filter	
Parameter: $\hat{c}$	Mean	Median	Mean	Median	Mean	Median
AAPL	0.251	0.216	0.283	0.158	0.008	0.006
CMCSA	0.64	0.71	0.92	0.93	0.07	0.06
CSCO	0.18	0.17	0.35	0.40	0.048	0.045
INTC	0.19	0.21	0.29	0.32	0.04	0.04
MSFT	0.126	0.125	0.157	0.170	0.018	0.016
Grand mean	0.456		0.289		0.02	
Parameter: $\hat{\lambda}$						
AAPL	0.985	0.990	0.987	0.938	0.47	0.46
CSCO	0.97	0.96	1.06	1.04	0.82	0.81
CMCSA	1.14	1.17	1.209	1.213	0.818	0.810
INTC	0.97	0.99	1.03	1.05	0.799	0.806
MSFT	0.91	0.91	0.94	0.96	0.66	0.64
Grand mean	0.91		0.83		0.50	

Table 8. Linear regression estimates as in equation (13) along with the adjusted R^2 of equation (13) for three different proxies of price impact $\beta_{i,\ell}$. First column: the book-reconstructed price impact $\beta_{i,\ell} = 1/(2\,D_{i,\ell})$. Second column: the 'static' price impact obtained by setting $(\Delta_K, \Delta_\ell) = (1\,\text{min}, 0.5\,\text{s})$. Third column: the (Kalman) filtered estimates of price impact.

	Book-reconstructed			'Static' $\beta_{i,\ell}$			Kalman filtered $\beta_{i,\ell}$		
Ticker	$\hat{\alpha}$	$\hat{\gamma}$	R^2	$\hat{\alpha}$	$\hat{\gamma}$	R^2	$\hat{\alpha}$	$\hat{\gamma}$	R^2
AAPL	$4.72\,10^{-3}$	0.29	0.48	$5.84\,10^{-3}$	0.800	0.64	$4.14\,10^{-3}$	1.063	0.82
	(22.09)	(212.26)		(32.52)	(289.31)		(32.9)	(467.4)	
CSCO	6.7210^{-4}	0.44	0.83	4.7410^{-4}	0.94	0.86	3.7610^{-4}	1.01	0.96
	(21.81)	(488.43)		(16.95)	(544.62)		(24.59)	(1055.29)	
CMCSA	2.0510^{-3}	0.46	0.64	1.8910^{-3}	0.92	0.69	1.4310^{-3}	1.03	0.89
	(18.71)	(294.07)		(18.68)	(331.71)		(23.9)	(635.9)	
INTC	8.3410^{-4}	0.41	0.80	6.0610^{-4}	0.95	0.85	4.6110^{-4}	1.02	0.96
	(20.13)	(437.41)		(17.06)	(526.22)		(23.6)	(1013.1)	
MSFT	2.1610^{-3}	0.39	0.74	1.8010^{-3}	0.91	0.81	1.4110^{-3}	1.02	0.91
	(25.86)	(366.00)		(25.14)	(445.67)		(29.07)	(700.08)	

7.3. Price impact and market depth

We now turn to an analysis of the relation existing between the proposed price impact measure and market depth. Moreover, we look at the goodness of fit of (4) when three different proxies of price impact are employed.

We remind that the baseline model for (4) is the model dubbed *stylized order book*, which mirrors the following relation between price impact and market depth:

$$\beta_\ell = \frac{c}{D_\ell^\lambda} + \nu_\ell, \tag{13}$$

where $c = 1/2$, $\lambda = 1$ and ν_ℓ is a noise term. Thus one can straightforwardly set β_ℓ equals to the following identity: $\beta_\ell = 1/(2D_\ell)$ (i.e. to reconstruct the price impact directly from the market depth. Technical details on the construction of market depth are discussed in appendix 1). Cont *et al.* (2014) went a step further by computing price impact estimates through linear regressions of price fluctuations on OFI (see section 4 of the present paper). Relying on their vast universe of stocks, they report the following results with $\Delta_K = 30$ min and $\Delta_\ell = 10$ s (see (17) and (18) in the original work): $\hat{c} = 0.56$ and $\hat{\lambda} = 1.08$. Thus at these time scales estimates of $\hat{c}$ and $\hat{\lambda}$ are very close to values predicted by the stylized order book model. We now ask what would happen to $\hat{c}$ and $\hat{\lambda}$ should one use either different Δ_K and Δ_ℓ or the price impact time-series filtered out through the use of Kalman filter. Is the price impact dynamics (almost) fully explained by quote size alone also in these in the aforementioned cases?

To answer this question, we estimate c and λ in (12) when price impact is constructed by employing the methodology of Cont *et al.* (2014) with $(\Delta_K, \Delta_\ell) = (30\,\text{min}, 10\,\text{s})$ and $(\Delta_K, \Delta_\ell) = (1\,\text{min}, 0.5\,\text{s})$, and when price impact is replaced with Kalman estimates. Incidentally, we notice that a percentage (although very low, of the order of 0.10%) of price impact estimates are negative. We thus fit, every month, (12) with a non-linear regression model by robust methods (i.e. by using iterated re-weighted least squares)†. Table 7 reports average values of $\hat{c}$ and $\hat{\lambda}$. The grand mean and median value of these parameters reported in the second and third columns are in line with that found in Cont *et al.* (2014)‡, thus confirming that at these time scales there exists a tight relation between price impact and the reciprocal of the market depth, as suggested by the stylized order book model. Estimates in the fourth and fifth column indicate that the latter is still a good approximation when $(\Delta_K, \Delta_\ell) = (1\,\text{min}, 0.5\,\text{s})$. Finally, the sixth and seventh column indicates that (12) is inadequate if $c = 1/2$ and $\lambda = 1$ when $\beta_{i,\ell}$ are Kalman estimates. This implies that, in this case, the price impact dynamic cannot be fully explained by (a deterministic function of) quote size alone. Again, we postulate that this might be linked to the auto-regressive nature of our price impact measure, which is capable to grasp different dimensions of the price impact in addition to the market depth only.

To support this assertion, we look at the goodness of fit of (5) at a time scale of 1 minute for three different proxies of price impact $\beta_{i,\ell}$: (1) the book-reconstructed price impact $\beta_{i,\ell} = 1/(2\,D_{i,\ell})$; (2) the 'static' price impact obtained by setting $(\Delta_K, \Delta_\ell) = (1\,\text{min}, 0.5\,\text{sec})$; (3) the (Kalman) filtered estimates of price impact. We run regressions of returns $\Delta P_{i,\ell}$ on $\beta_{i,\ell}$, $i = 1, \ldots, N$ and $\ell = 1, \ldots, L$:§

$$\Delta P_{i,\ell} = \alpha + \gamma\, \beta_{i,\ell} \text{OFI}_{i,\ell} + \nu_{i,\ell}, \tag{14}$$

where $\nu_{i,\ell}$ is a random error. Table 8 reports the results and confirms what we have noticed so far. Precisely, in the stylized model introduced by Cont *et al.* (2014), price changes are a mechanical consequence of the provision and depletion of volume at the best quotes. In particular, this allows to express price impact (see (12)) as a function of the size of the best quotes. If this model for price impact was accurate, the utility of alternative estimation methods would then be restricted to the case where order book data are not available. The clearly superior results obtained by price impact estimates based on price change and OFI suggest that this is not the case. Price impact dynamics cannot be fully explained by quote size alone. The method proposed by Cont *et al.* (2014)

† We also implement the two-step linear regressions employed in Cont *et al.* (2014), by ignoring negative estimates. However, no sensitive differences on estimates emerge.

‡ Notice that the data set examined by the authors refers to the month April 2010.

§ Notice that in this regression is present also the intercept; this explain why the $\hat{\gamma}$ coefficient in the case of the static model is not equal to 1.

already captures some of the variation not explained by quote size, as proven by the higher coefficient of determination, but additional improvements can be achieved by introducing a dynamic for price impact as in the model we have proposed in this paper.

8. Concluding remarks

In this paper, we introduced a model in which price impact is linear and permanent, but the price impact coefficient is a latent state of the limit order book subject to its own dynamics. In particular, we define the price impact coefficient to be the product of three components: (1) a daily price impact level, (2) a deterministic intraday pattern, and (3) a stochastic autoregressive component.

An important characteristic of our model is that we do not study price impact components in isolation. This is critical for real-time analysis and decision making. In particular, traders should recognize and quantify the interactions and contributions of the various price impact components as they simultaneously affect the dynamics of the price. Besides, such a decomposition is necessary, since factors display widely different statistical properties.

We calibrate the model on five stocks traded at NASDAQ using 1 min data in 2016. We show that (1) market impact dynamics should be taken into the account on out-of-sample estimation performance, (2) the history dependent nature of our price impact measure, due to the use of the Kalman filter, captures different dimensions of the price impact in addition to the market depth only. In particular, we show that our model does not imply a square-root relation between price changes and trading volume. This statement is true also when using the static estimates of Cont *et al.* (2014) on smaller timescales than in their original work. Finally, (3) our approach to market impact estimation explains a greater dispersion of price fluctuations when compared to other proxies. In an out-of-sample exercise, we show that the MSE for the Kalman filter is smaller than the one obtained when using the static approach of Cont *et al.* (2014). Note that in the static model the average daily level of market impact is automatically determined by the estimation. On the contrary, the Kalman approach uses the predictive filter and, more importantly, it uses it for the daily average of market impact the level of the day before. In spite of this disadvantage, we obtain a reduction in the MSE which is primarily due to a reduction in the variance of the dynamic estimates.

As a by-product, our analysis allows for the estimation of the market impact intraday pattern. Price impact is at its peak at the beginning of the day, right after the opening auction. The level decreases in the next hour to an average daily level, and then drops again steeply in the last thirty minutes of the trading day. We interpret this result as being symptomatic of higher opinion dispersion and adverse selection at the opening of the market, which translates in to a lower order book liquidity, or equivalently, in to a higher market impact. As trading begins, the price discovery process takes place and liquidity gradually reverts from this initial shock to equilibrium level, lowering market impact.

In summary, the main contribution of this paper is the introduction of an improved methodology for characterizing security price dynamics within the cited type of models. Broader finance applications would include liquidity effects on asset pricing, optimal trading strategies, or market design. Still, the latter applications may represent an interesting direction for future research.

Acknowledgments

L.M would like to thank Robert Almgren, Adam Clark-Joseph, Joel Hasbrouck, Arseniy Kukanov, Lira Rocha da Mota, Adrien Nguyen Huu and Jorge Zubelli for precious comments on this work. G.L. would like to thank Angelo Ranaldo, Marcello Rambaldi, Daniele Tantari for comments on this work. A.C. acknowledges a Ph.D. studentship from the Engineering and Physical Sciences Research Council through Grant No. EP/L015129/1. The research activity of G.L. was supported by Unicredit S.p.a. under the project: 'Dynamics and Information Research Institute - Quantum Information (Teoria dell' Informazione), Quantum Technologies'.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Fabrizio Lillo http://orcid.org/0000-0002-4931-4057

References

Acharya, V.V. and Pedersen, L.H., Asset pricing with liquidity risk. *J. Financ. Econ.*, 2005, **77**, 375–410.

Almgren, R. and Chriss, N., Optimal execution of portfolio transactions. *J. Risk*, 2001, **3**, 5–39.

Amihud, Y. and Mendelson, H., Asset pricing and the bid-ask spread. *J. Financ. Econ.*, 1986, **17**, 223–249.

Andersen, T.G. and Bollerslev, T., Intraday periodicity and volatility persistence in financial markets. *J. Empir. Financ.*, 1997, **4**, 115–158.

Andersen, T.G. and Bollerslev, T., Deutsche mark–dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer run dependencies. *J. Finance*, 1998, **53**, 219–265.

Andresen, T. and Bollerslev, T., DM-Dollar volatility: Intraday activity patterns, macroeconomic announcements and longer run dependence. *J. Finance*, 1998, **53**, 219–265.

Bertsimas, D. and Lo, A.W., Optimal control of execution costs. *J. Financ. Markets*, 1998, **1**, 1–50.

Bollerslev, T., Cai, J. and Song, F.M., Intraday periodicity, long memory volatility, and macroeconomic announcement effects in the US treasury bond market. *J. Empir. Financ.*, 2000, **7**, 37–55.

Bouchaud, J.P., Farmer, J.D. and Lillo, F., How markets slowly digest changes in supply and demand. In *Handbook of Financial Markets: Dynamics and Evolution*, edited by T. Hens and K.R. Schenk-Hoppe, pp. 57–160, 2009 (North-Holland: San Diego).

Bouchaud, J.P., Price impact. In *Encyclopedia of Quantitative Finance*, edited by R. Cont, 2010 (John Wiley & Sons: Hoboken, NJ).

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

Capponi, F. and Cont, R., Multi-asset market impact and order flow commonality. Available at SSRN, 2020.

Cartea, A. and Jaimungal, S., Incorporating order-flow into optimal execution. *Math. Financ. Econ.*, 2016, **10**, 339–364.

Casgrain, P. and Jaimungal, S., Mean-field games with differing beliefs for algorithmic trading. *Math. Financ.*, 2020, **30**(3), 995–1034.

Casgrain, P. and Jaimungal, S., Trading algorithms with learning in latent alpha models. *Math. Finance*, 2019, **29**, 735–772.

Chiyachantana, C.N., Jain, P.K., Jiang, C. and Wood, R.A., International evidence on institutional trading behavior and price impact. *J. Finance*, 2004, **59**, 869–898.

Chung, K.H. and Chuwonganant, C., Market volatility and stock returns: The role of liquidity providers. *J. Financ. Markets*, 2018, **37**, 17–34.

Cont, R., Kukanov, A. and Stoikov, S., The price impact of order book events. *J. Financ. Econ.*, 2014, **12**, 47–88.

Datar, V.T., Naik, N.Y. and Radcliffe, R., Liquidity and stock returns: An alternative test. *J. Financ. Markets*, 1998, **1**, 203–219.

Durbin, J. and Koopman, S.J., *Time Series Analysis by State Space Methods*, Vol. 2012 (OUP Oxford: Oxford).

Eisler, Z., Bouchaud, J.P. and Kockelkoren, J., The price impact of order book events: Market orders, limit orders and cancellations. *Quant. Finance*, 2012, **12**, 1395–1419.

Engle, R.F. and Sokalska, M.E., Forecasting intraday volatility in the us equity market. Multiplicative component garch. *J. Financ. Econ.*, 2012, **10**, 54–83.

Farmer, J.D., Gillemot, L., Lillo, F., Mike, S. and Sen, A., What really causes large price changes? *Quant. Finance*, 2004, **4**, 383–397.

Foucault, T., Pagano, M. and Röell, A., *Market Liquidity: Theory, Evidence, and Policy*, 2013 (Oxford University Press: New York).

Giot, P., Market risk models for intraday data. *Eur. J. Finance*, 2005, **11**, 309–324.

Glosten, L.R. and Harris, L.E., Estimating the components of the bid/ask spread. *J. Financ. Econ.*, 1988, **21**, 123–142.

Glosten, L.R. and Milgrom, P.R., Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *J. Financ. Econ.*, 1985, **14**, 71–100.

Gould, M.D. and Bonart, J., Queue imbalance as a one-tick-ahead price predictor in a limit order book. *Market Microstructure Liquidity*, 2016, **2**, 1650006.

Hamilton, J.D., *Time Series Analysis*, Vol. 1994 (Princeton University Press: Princeton).

Hasbrouck, J., Measuring the information content of stock trades. *J. Finance*, 1991, **46**, 179–207.

Hasbrouck, J., *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*, 2007 (Oxford University Press).

Hautsch, N. and Huang, R., The market impact of a limit order. *J. Econ. Dyn. Control*, 2012, **36**, 501–522.

Huang, R. and Polak, T., LOBSTER: The limit order book reconstructor. School of Business and Economics, Humboldt Universität zu Berlin, Tech. Rep., 2011.

Kyle, A.S., Continuous auctions and insider trading. *Econometrica*, 1985, **53**, 1315–1335.

Lillo, F. and Farmer, J.D., The long memory of the efficient market. *Stud. Nonlinear Dyn. Econ.*, 2004, **8**(3).

Lillo, F. and Farmer, J.D., The key role of liquidity fluctuations in determining large price changes. *Fluctuation Noise Lett.*, 2005, **5**, L209–L216.

Lillo, F., Farmer, J.D. and Mantegna, R.N., Econophysics: Master curve for price-impact function. *Nature*, 2003, **421**, 129.

Madhavan, A., Richardson, M. and Roomans, M., Why do security prices change? A transaction-level analysis of NYSE stocks. *Rev. Financ. Stud.*, 1997, **10**, 1035–1064.

Müller, H.G., Sen, R. and Stadtmüller, U., Functional data analysis for volatility. *J. Econom.*, 2011, **165**, 233–245.

Pástor, L. and Stambaugh, R.F., Liquidity risk and expected stock returns. *J. Polit. Econ.*, 2003, **111**, 642–685.

Pereira, J.P. and Zhang, H.H., Stock returns and the volatility of liquidity. *J. Financ. Quant. Anal.*, 2010, **45**, 1077–1110.

Saar, G., Price impact asymmetry of block trades: An institutional trading explanation. *Rev. Financ. Stud.*, 2001, **14**, 1153–1181.

Stroud, J.R. and Johannes, M.S., Bayesian modeling and forecasting of 24-hour high-frequency volatility. *J. Am. Stat. Assoc.*, 2014, **109**, 1368–1384.

Xu, K., Gould, M. and Howison, S., Multi-level order-flow imbalance in a limit order book. Available at SSRN 3479741, 2019.

Appendices

The present appendix is divided into five parts. Section 1 describes the data processing, Section 2 discusses some technical details for the estimation procedure, Section 3 discusses the performance of the model when dealing with small tick stocks, section 5 reports a simulation study, and section 4 reports sample auto-correlations of the one-step ahead forecast errors and of its squares, as well as the corresponding quantile-quantile plots.

Appendix 1. Data processing

The logarithmic returns and OFI measures are obtained as follows. Each day, we remove the first and last 30 min of trading activity to filter out unusual events that are likely to occur around the opening and closing auctions. We split the day into slots of 60 seconds starting at 0930EST and finishing at 1600EST. Slots are of semi-open form $[t_{\ell-1}, t_\ell[$, with t_ℓ indicating the ℓth bound. We do not abide by this rule in the last bin of the day, where a close interval is instead used to cover the events that exactly occur during the closing time-stamp (1600 EST).

To construct the OFI, we first detect the submissions of limit orders, executions of market orders, and cancellations/deletions of previously posted orders occurring within the considered time interval. Then, in line with section 4, we compute the following two metrics: (1) cumulative volume of submitted buy limit orders (L_ℓ^b), executed buy market orders (M_ℓ^b), and cancelled/deleted sell limit orders (C_ℓ^b); (2) cumulative volume of submitted sell limit orders (L_ℓ^s), executed sell market orders (M_ℓ^s), and cancelled/deleted buy limit orders (C_ℓ^s). Finally, OFI_ℓ is defined as the difference between (1) and (2).

To construct returns, we first have to define a notion of mid-price. The latter is calculated by taking the mid point between the best bid and ask prices reported in the LOB file provided by LOBSTER. In this work, we look at each slot and sample the last recorded mid price before the upper bound of the considered interval. Finally, we take the first order difference of the sampled time series in order to obtain the price changes statistics ΔP_ℓ.

In rare circumstances (e.g. slow trading or data issues) it may happen that no LOB events are recorded within a considered slot, say $[t_{\ell-1}, t_\ell[$. In this case, the OFI for that particular slot is set to zero, whereas the sampled mid price is pegged to the value sampled from the previous time interval, thus reflecting the fact that no transactions has occurred within $[t_{\ell-1}, t_\ell[$.

Finally, in order to construct the depth we introduce, for sake of convenience, the following definitions.

Definition A.1 *We define an $\mathbb{R}$-valued irregularly spaced time series Q to be a sequence $(t_i, q_i)_{i\in I}$, where $I \subset \mathbb{N}$ is a set of indices, $t_i \in [0, T] \subset \mathbb{R}$, with $t_i \le t_{i+1}$, and $q_i \in \mathbb{R}$. We denote t_i the time stamps and q_i the values of the time series.*

Definition A.2 *Let Q be a time series over $[0, T]$, $\tau \in \mathbb{R}$ be the length of a time interval, $n = T/\tau$. We define the τ-time grid* *associated to Q to be the set $G = \{p_n\}_{n\in N}$ where*

$$p_n := \sup\{i \in I \mid t_i < n\tau\}$$

Suppose we partition $[0, T]$ in N uniform intervals of length τ. We observe that p_n is the index of the last time-stamp before the end of the nth time interval.

DEFINITION A.3 *Given an irregularly spaced time series Q and an associated τ-time grid G, we define the averaged bin time series to be the series $\mathcal{Q} = (\mathcal{Q}_n)_{n\leq N}$ defined as*

$$\mathcal{Q}_1 = \frac{1}{\tau}\left(q_1 t_1 + \sum_{k=1}^{p_1-1} q_k(t_{k+1} - t_k) + q_{p_1}(\tau - t_{p_1})\right)$$

$$\mathcal{Q}_{n+1} = \frac{1}{\tau}\left(q_{p_n}(t_{(p_n+1)} - n\tau) + \sum_{k=p_n+1}^{p_{(n+1)}-1} q_k(t_{k+1} - t_k)\right.$$
$$\left. + \, q_{p_{(n+1)}}((n+1)\tau - t_{p_{n+1}})\right) \quad n \geq 1$$

In our case, given the time series (t_i, b_i) and (t_i, a_i), with b_i and a_i the bid and the ask size at time t_i respectively, we define the depth as the uniform time series (D_ℓ), $\ell = 1, \ldots, L$ defined as $D_\ell = (B_\ell + A_\ell)/2$, where B_ℓ and A_ℓ are the averaged bin time series of (t_i, b_i) and (t_i, a_i). In other words, we define the depth over each the time interval $[t_{\ell-1}, t_\ell[$ as the weighted average of the size at the best quotes with weight proportional to the time a given size was available.

Appendix 2. Technical details on the estimation procedure

Kalman filter implementation is performed with the R package FKF (Fast Kalman Filter). The package is mainly written in C and uses the FORTRAN LAPACK package to deliver fast computation of Kalman filtering and forecasting. Kalman smoothing was performed with an implementation of the forward-backward smoothing algorithm given in equation (11). Before performing the optimization of log-likelihood (12), as suggested in Durbin and Koopman (2012), we re-parametrize the volatilities in equation (8) as $s_{\cdot,\epsilon} = \log\sigma_{\cdot,\epsilon}$ and $s_{\cdot,\eta} = \log\sigma_{\cdot,\eta}$ to get an unconstrained optimization problem. In order to perform the optimization, we need a prior for the state space and an initial value for the vector of parameters Θ. Initial values for $\beta_\cdot$ and $\sigma_{\cdot,\epsilon}$ are computed by performing an Ordinary Least Squares (OLS) regression $\Delta P_\ell = \beta_\cdot \pi_\ell \mathrm{OFI}_{\cdot,\ell} + \epsilon_{\cdot,\ell}$, $\epsilon_{\cdot,\ell} \sim \mathrm{NID}(0, \sigma^2_{\cdot,\epsilon})$; we set their initial values to the corresponding OLS estimates. Instead, for $\rho_\cdot$ and $\sigma_{\cdot,\eta}$ we perform a grid search with $\rho_\cdot \in (0, 1)$ and $\sigma_{\cdot,\eta} \in (0.1, 2)$. Finally, the prior for the state space is chosen to be deterministic and equal to $\mathsf{q}_{\cdot,1} = 1$ and $\mathsf{P}_{\cdot,1} = 2$; the former is the unconditional mean of the process, the latter is chosen on empirical observations.

Appendix 3. The performance of the model for small tick stocks

So far, we have analysed the performance of the large tick stocks. In this section we investigate the performance of the model for small tick stocks which, as said, present features that are at odds with the hypotheses of the model. In particular, we provide evidence for the latter thesis. The stocks we consider are Amgen, Inc. (AMGN) Amazon.com Inc. (AMZN) and Alphabet Inc. Class C (GOOG). Tables A1 and A2 collect some descriptive statistics for these stocks.

Figure A1 displays the analogous quantities of figure 1 for AMZN. We focus on the right panel, which shows that the goodness of fit of the model for small tick stock is significantly smaller than the one observed for large tick stocks.

Also, we conduct the VAR analysis of section 6. Table A3 reports the regression coefficients of model (12) along with the corresponding t-statistics between brackets. Again, since the estimation is applied to each day in our sample, the outcome is five parameters values. Contrary to the large tick stocks case, for two out of three stocks, both the coefficient c_1—which captures the effect of past mid-price changes on the current value of OFI—and d_1—which captures the effect of past OFI on the current value of OFI—are highly significant. We comment on each quantity.

(1) (c_1) The coefficients of lagged price change on the current level of OFI are negative and significant. The intuition goes as follows. Suppose, for instance, that a cancellation order at the best ask occurs by causing the depletion of the best ask queue and, as a consequence, a positive price change. This leads to an increase of sell market orders, cancellations of buy limit orders, and an increase of sell limit orders, i.e. to a decrease of the OFI. Moreover, even if the initial cancellation order at the best ask does not cause the depletion of the queue, the same reasoning applies since it is well known

Table A1. Descriptive statistics of investigated stocks over the sample period. The sample period is from January 1, 2016, to June 30, 2016. Mid-price and Spread are reported in dollar unit. The Relative Spread is reported in basis point unit. Stocks are sorted by average price (or by spread), i.e. inversely by relative tick size.

		Mid-price		Spread		Relative Spread	
Stock	Ticker	Avg.	Std	Avg.	Std	Avg.	Std
Amgen, Inc.	AMGN	151.998	5.513	0.086	0.022	5.67	1.52
Amazon.com Inc.	AMZN	622.819	69.093	0.416	0.107	6.84	2.22
Alphabet Inc. Class C	GOOG	717.521	21.643	0.473	0.129	6.61	1.88

Table A2. Main sample statistics of the limit order book averaged over the sample period. The sample period is from January 1, 2016, to June 30, 2016. The amount of limit orders at the best bid (Bid $Q_{1,\ell}$) and ask (Ask $Q_{1,\ell}$), of cancellations (Bid $Q_{1,c}$ and Ask $Q_{1,c}$) and of market orders (Bid $Q_{1,m}$ and Ask $Q_{1,m}$) for each stock is reported. Quantities are characterized in term of both number of events (#Ev.) and average volume (Avg.Vol.) measured in number of shares.

	Bid $Q_{1,\ell}$		Ask $Q_{1,\ell}$		Bid $Q_{1,c}$		Ask $Q_{1,c}$		Bid $Q_{1,m}$		Ask $Q_{1,m}$	
Symbol	#Ev.	Avg.Vol.	#Ev.	Avg.Vol.	# Ev.	Avg.Vol.	#Ev.	Avg.Vol.	#Ev.	Avg.Vol.	#Ev.	Avg.Vol.
AMGN	16412.88	234.23	16186.08	232.55	10223.09	137.75	9922.26	134.91	4375.17	48.93	7486.83	83.74
AMZN	17863.07	780.85	17629.69	793.85	11593.60	432.32	11186.28	436.26	5108.44	178.90	11444.78	392.32
GOOG	19005.85	914.45	24914.89	1011.14	12355.18	514.16	17500.40	565.43	3580.73	136.56	7653.50	291.92

Figure A1. Left panel represents averages across days of the price impact coefficient estimates $\hat{\beta}_k$ as in equation (4) for AMZN as a function of the time of the day for $\Delta_K = 30$ min and Δ_ℓ equal to: 0.5 s (red), 5 s (blue), 10 s (green), 20 s (magenta), 30 s (gold), 1 min (cyan) and 5 min (black). Right panel represents the contour plot of the averages R^2 across days as a function of the regression window Δ_K and sampling frequency Δ_ℓ.

Table A3. VAR parameter estimates. Coefficients which result to be significant at 10%, 5% and 1% confidence levels are marked, respectively, with one, two and three stars.

	Parameters' values (*t*-stats)				
Ticker	b_0	a_1	b_1	c_1	d_1
AMGN	− 2.87e − 05	0.102	− 8.97e − 06	− 5419.17	0.34
	(− 19.74***)	(0.78)	(− 1.24)	(− 3.22***)	(5.09***)
AMZN	− 1.04e − 04	0.092	− 4.06e − 05	− 1265.3	0.42
	(− 15.20***)	(1.22)	(− 2.86**)	(− 3.44***)	(7.64***)
GOOG	− 3.31e − 07	− 0.01	1.84e − 08	− 112226.99	0.01
	(− 27.78***)	(− 0.003)	(0.15)	(− 0.46)	(0.14)

that the bid/ask queue imbalance in a LOB provides significant predictive power for the direction of the next mid-price movement (see, for instance Gould and Bonart 2016, and reference therein). Finally, we note that if one neglects the presence of limit orders and cancellations, the negative lagged ΔP_ℓ coefficient in the OFI_ℓ specification implies Granger-Sims causality running from quote revisions to trades, and it is consistent with findings in Hasbrouck (1991).

(2) (d_1) This parameter is related to the auto-correlation of the (aggregated) order flow imbalance measure. To better understand why this coefficient for AMGN and AMZN is (statistically) significant we first define the following variables: $\Delta M_\ell \equiv M_\ell^b - M_\ell^s$, $\Delta L_\ell \equiv L_\ell^b - L_\ell^s$, $\Delta C_\ell \equiv C_\ell^b - C_\ell^s$. Then, we run the following regression:

Table A4. Regression parameter estimates. Coefficients which result to be significant at 10%, 5% and 1% confidence levels are marked, respectively, with one, two and three stars.

	Parameters' values (*t*-stats)		
Ticker	β_1	β_2	β_3
AMGN	0.212	0.13	− 0.05
	(2.02**)	(0.06)	(− 0.05)
AMZN	0.406	0.215	− 0.171
	(4.90***)	(0.06)	(− 2.02**)
GOOG	0.008	0.030	− 0.060
	(0.03)	(0.06)	(− 0.63)

$$\mathrm{OFI}_\ell = \beta_1\,\Delta M_{\ell-1} + \beta_2\,\Delta L_{\ell-1} + \beta_3\,\Delta C_{\ell-1} + \eta_\ell, \tag{A1}$$

where η_ℓ is noise term. The following table A4 summarizes the results (results for large tick stocks are in line with those of GOOG). Therefore, for small tick stocks, the auto-correlation of the OFI is induced from the Granger-Sims causality running from market orders (and cancellations) to order flow imbalance.

In conclusion, in order to extend the model to small tick stocks, one should account for these additional dependencies. A comprehensive exploration of this direction is beyond the scope of the present paper, and is left for future work. For sake of completeness, we estimate also for AMGN, AMZN and GOOG the proposed model: the following table A5 reports the results of the estimation.

Appendix 4. Diagnostic testing for Kalman filter

The following figure A2 report examples of sample auto-correlations of the standardized one-step ahead forecast errors, of its squares, as well as of its quantile–quantile plots.

Appendix 5. Simulation study

The robustness of the two-steps procedure proposed in section 5.1 is tested numerically in the present section. We choose a simulation setting that resembles as close as possible the main features of our data. In particular, we assume a trading day of 6.5 hours and

Table A5. Summary of estimation results of model in equations (1)–(3) on order book data of eight NASDAQ 100 Index stock over 100 days in 2016. For each parameter we report: (1) The median (second column), (2) The mean (third column), (3) The average t-statistic (fourth column), (4) the lower and the upper quartile.

Ticker	Median	Mean	Average t-statistic	Q_1	Q_3
Parameter: β					
AMGN	3.239×10^{-3}	3.287×10^{-3}	42.13	2.775×10^{-3}	3.769×10^{-3}
AMZN	10.92×10^{-3}	11.28×10^{-3}	30.41	9.158×10^{-3}	13.30×10^{-3}
GOOG	11.51×10^{-3}	12.30×10^{-3}	25.07	9.516×10^{-3}	14.64×10^{-3}
Parameter: ρ					
AMGN	0.4262	0.4954	15.42	0.2587	0.7638
AMZN	0.4298	0.4815	13.45	0.2026	0.7821
GOOG	0.3634	0.4259	12.6	0.1454	0.6463
Parameter: σ_ϵ					
AMGN	0.0617	0.0719	22.43	0.0512	0.0864
AMZN	0.2891	0.3369	20.32	0.2433	0.3951
GOOG	0.2545	0.3106	19.09	0.2066	0.3694
Parameter: σ_η					
AMGN	0.4151	0.4114	18.61	0.2553	0.5749
AMZN	0.5576	0.5333	16.34	0.3372	0.6624
GOOG	0.5435	0.5336	18.79	0.2002	0.3758

we simulate $N = 100$ one-minute time-series of $L = 390$ from the model described by equations (1)–(3). We use the following data generating process (DGP). First, we keep from real data a representative time-series for the intraday pattern and the OFI, as well as values for the hyper-parameters vector $\mathbf{\Theta} = \left(\beta, \rho, \sigma_\epsilon, \sigma_\eta\right)$. Then, *at each trading day*, we sample the daily level of market impact from a positive uniform random variable with support $(0, 1.5\,\beta)$, $\mathrm{OFI}_{\cdot,\ell}$, $\ell = 1, \ldots, L$, from a zero-mean Gaussian random variable with a standard deviation consistent with that found on empirical data†, the coefficient ρ from a positive uniform random variable with support $(0.2, 0.4)$. Finally, $\epsilon_{\cdot,\ell}$ (resp. $\eta_{\cdot,\ell}$), $\ell = 1, \ldots, L$, are sampled from independent and identically distributed Gaussian random variables with a standard deviation equals to $\tilde{\sigma}_\epsilon$ (resp. $\tilde{\sigma}_\eta$), where $\tilde{\sigma}_\epsilon$ (resp. $\tilde{\sigma}_\eta$) are sampled from a positive uniform random variable with support $(0.8\,\sigma_\epsilon, 1.2\,\sigma_\epsilon)$ (resp. $\left(0.8\,\sigma_\eta, 1.2\,\sigma_\eta\right)$). For sake of completeness, we perform our simulation study with values of parameters, intraday pattern and OFI consistent with those found in empirical data for both large (e.g. MSFT) and small (e.g. AMZN) tick stocks. We comment now the results. Figure A3 reports the true and estimated time-of-the-day pattern. We see that in the 'large tick stocks' setting the estimated intraday pattern is basically indistinguishable from the true one, whereas in the 'small tick stocks' setting it is estimated with some bias, especially in the last part of the (fictitious) trading day. Incidentally, we notice that in the latter case it may happens that a non-negligible percentage of simulated price change and OFI have opposite signs and estimates of the parameter ρ are negative. Table A6 reports results of parameter estimates (computed by ignoring the just mentioned pathological cases). For sake of completeness, estimates after the first stage are also reported. Parameters are estimated remarkably well and, as observed in real data, the autoregressive coefficient ρ is overestimated if the intraday pattern is not taken into account in the 'large tick stocks' setting. Remarkably, in both cases, accounting for the intraday pattern reduces the dispersions of parameter estimates.

† Here, we ignore the temporal correlation of the OFI. However, in first approximation, this is consistent with what observed on real data.

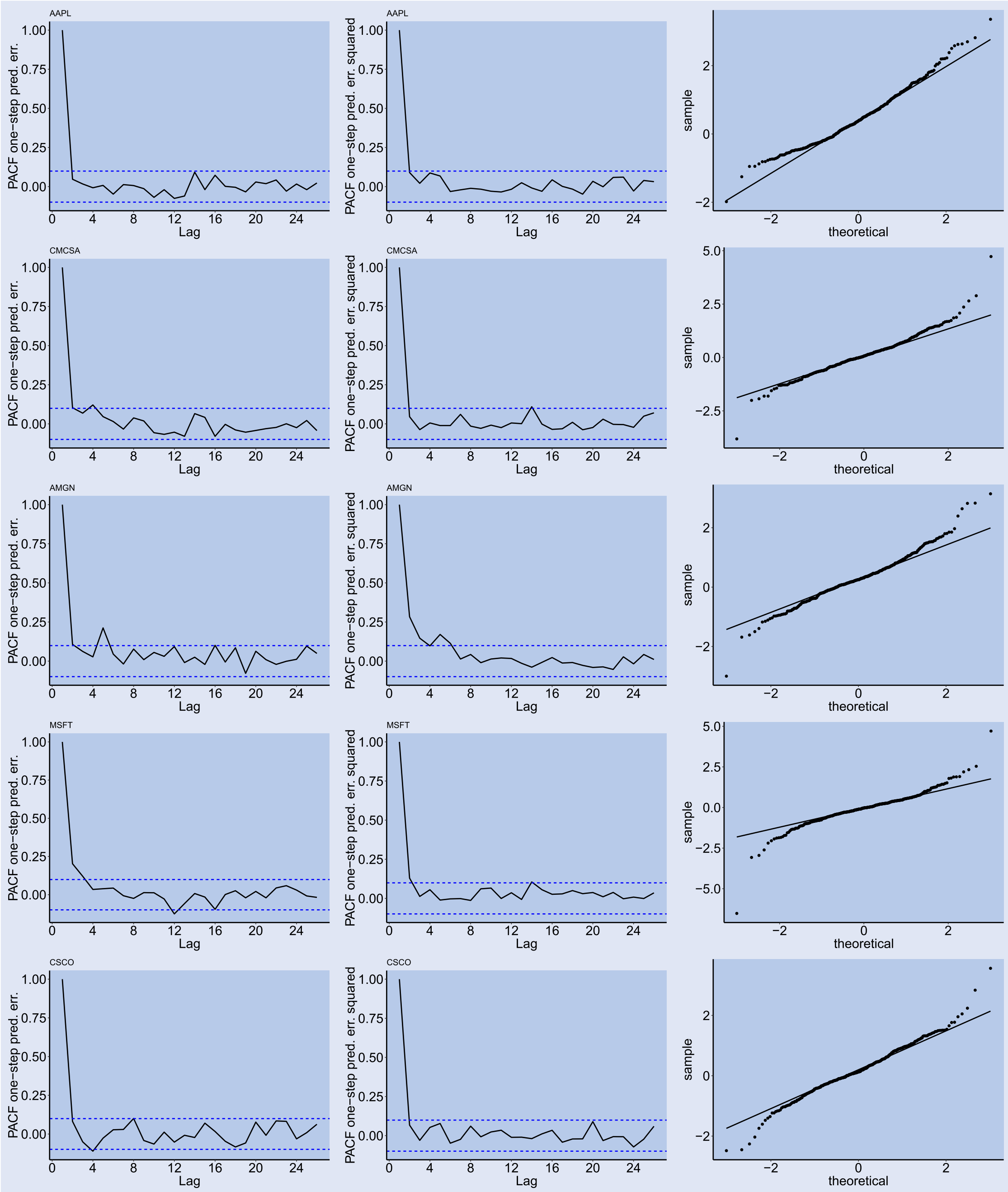

Figure A2. Examples of sample auto-correlations of the standardized one-step ahead forecast errors, of its squares, as well as of its quantile–quantile plots.

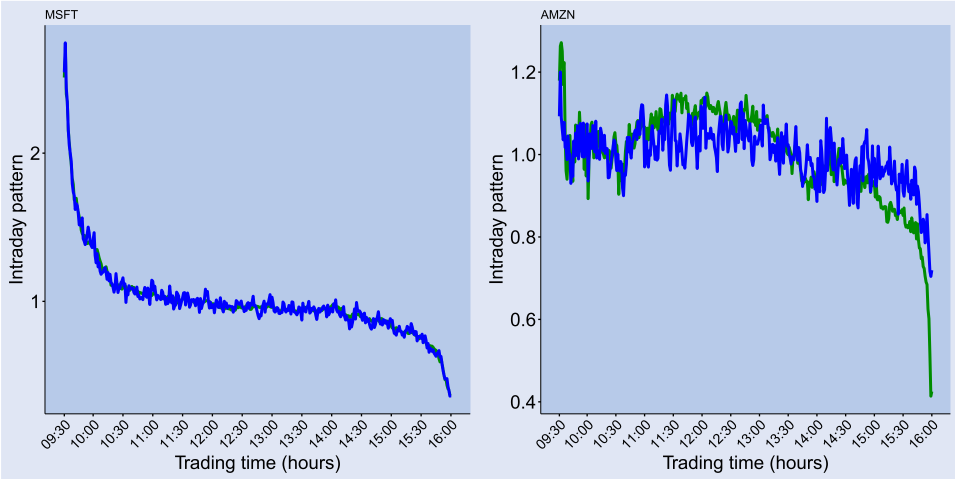

Figure A3. True (green line) and the estimated (blue line) intraday pattern in simulated data. Left panel: typical situation for small tick stocks. Right panel: typical situation for a large small tick stock.

Table A6. Parameter estimates on simulated data. Top panel: Simulation setting for large tick stocks. Bottom panel: Simulation setting for small tick stocks. From left column to right column: True (distributional) value of the parameter; Estimated values of the parameter after the first stage; Estimated values of the parameter after the second stage.

Parameter	True value	Estimated value I	Estimated value II
Large tick setting			
β	$\overset{d}{\sim} U\,(0, 1.5\beta)$	1.140×10^{-6}	1.256×10^{-6}
	$\beta = 1.413 \times 10^{-6}$	(5.784×10^{-7})	(5.283×10^{-7})
σ_ϵ	$\overset{d}{\sim} U\,(0.8\sigma_\epsilon, 1.1\sigma_\epsilon)$	9.938×10^{-3}	9.224×10^{-3}
	$\sigma_\epsilon = 9.628 \times 10^{-3}$	(1.224×10^{-3})	(1.111×10^{-3})
ρ	$\overset{d}{\sim} U\,(0.2, 0.4)$	0.78	0.33
		(0.14)	(0.15)
σ_η	$\overset{d}{\sim} U\,(0.8\sigma_\eta, 1.2\sigma_\eta)$	0.198	0.248
	$\sigma_\eta = 0.275$	(0.115)	(0.05)
Small tick setting			
β	$\overset{d}{\sim} U\,(0, 1.5\beta)$	7.83×10^{-5}	7.87×10^{-5}
	$\beta = 9.739 \times 10^{-5}$	(4.195×10^{-5})	(4.176×10^{-7})
σ_ϵ	$\overset{d}{\sim} U\,(0.8\sigma_\epsilon, 1.1\sigma_\epsilon)$	0.283	0.284
	$\sigma_\epsilon = 0.301$	(0.026)	(0.0262)
ρ	$\overset{d}{\sim} U\,(0.2, 0.4)$	0.36	0.33
		(0.24)	(0.12)
σ_η	$\overset{d}{\sim} U\,(0.8\sigma_\eta, 1.2\sigma_\eta)$	0.532	0.498
	$\sigma_\eta = 0.540$	(0.285)	(0.16)