Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Disentangling Sources of High Frequency Market Microstructure Noise

Simon Clinet & Yoann Potiron

To cite this article: Simon Clinet & Yoann Potiron (2021) Disentangling Sources of High Frequency Market Microstructure Noise, Journal of Business & Economic Statistics, 39:1, 18-39, DOI: 10.1080/07350015.2019.1617158

To link to this article: https://doi.org/10.1080/07350015.2019.1617158

Disentangling Sources of High Frequency Market Microstructure Noise

Simon CLINET
Faculty of Economics, Keio University, 2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan (*clinet@keio.jp*)

Yoann POTIRON
Faculty of Business and Commerce, Keio University, 2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan (*potiron@fbc.keio.ac.jp*)

Employing tick-by-tick maximum likelihood estimation on several leading models from the financial economics literature, we find that the market microstructure noise is mostly explained by a linear model where the trade direction, that is, whether the trade is buyer or seller initiated, is multiplied by the dynamic quoted bid-ask spread. Although reasonably stable intraday, this model manifests variability across days and stocks. Among different observable high frequency financial characteristics of the underlying stocks, this variability is best explained by the tick-to-spread ratio, implying that discreteness is the first residual source of noise. We determine the bid-ask bounce effect as the next source of noise.

KEY WORDS: Efficient price; High frequency data; Market microstructure noise; Mid price; Trade direction.

1. INTRODUCTION

In financial econometrics, a major topic is the estimation of volatility gauging measures from high frequency asset (log-)price data. A common assumption is that the vector of raw prices consisting of the transaction, the best ask, the best bid and the mid prices is generated by a latent efficient Itô-semimartingale process which is contaminated by noise typically accounting for market frictions inherent in the trading process such as: bid-ask spread, whether the trade is buyer or seller initiated, discreteness due to the fact that trades lie on the tick grid, bid-ask bounce effects, that is, the fact that transactions are observed at the bid and ask prices albeit the mid price stays constant, limited volume available, volume imbalance, etc.

Hasbrouck (1995) applied cointegration to relate the vector of raw prices on multiple markets to the efficient price common to all markets. However, the model is initially restricted to the transaction price. Subsequently, Hansen and Lunde (2006) employed cointegration in a model incorporating the full vector of prices. Exploiting Granger representation used in Hasbrouck (2002), they construct an efficient price related to the vector, which has the desired martingale property.

Another simple and familiar model related to the raw prices vector is Roll (1984). In spite of being primarily stipulated with no related efficient price, a generalized version incorporating the latent price is specified in Hasbrouck (2002). In that model, the mid price is de facto equal to the efficient price, and the single source of market microstructure noise is the signed spread, that is, the association of the effective spread along with the trade direction. As a matter of fact, it is common practice in the financial econometrics literature to use the mid price as a measure of the efficient price. While this measure is noisy, it is generally closer to the latent price than is the transaction price since it does not suffer from bid-ask bounce effects.

In conjunction with the information on the trade direction, prominent extensions include and are not limited to the information about the traded volume (Glosten and Harris 1988), the duration time between two trades (Almgren and Chriss 2001), the quoted depth (Kavajecz 1999), and the quoted spread. Estimation with high frequency data on models incorporating trading information has been studied in a recent strand of papers. Li, Xie, and Zheng (2016) and Chaker (2017) provided general methods to estimate more efficiently volatility based on such models. The general model they considered can be described as

$$\underbrace{Z_{t_i}}_{\text{transaction price}} = \underbrace{X_{t_i}}_{\text{efficient price}} + \underbrace{\underbrace{\phi(Q_{t_i},\theta_0)}_{\text{explicative part}} + \underbrace{\epsilon_{t_i}}_{\text{residual noise}}}_{\text{market microstructure noise}}, \quad (1)$$

where Q_{t_i} correspond to the aforementioned (e.g., trade direction, traded volume, etc.) observable variables and ϕ is a function known to the econometrician, such as the linear trade direction in case of the Roll model. In a first step, they pre-estimated the explicative part of the noise. They removed it from the transaction price in a second step, thus reducing the overall uncertainty about the efficient price. They also proposed several pre-estimated price based volatility estimators. Although the method works remarkably well in theory and in numerical simulation to reduce the error related to volatility estimation, it unfortunately requires that the econometrician chooses a specific ϕ and a null/nonzero residual noise scenario. Among other things, Clinet and Potiron (2019) developed tests

Journal of Business & Economic Statistics
January 2021, Vol. 39, No. 1
DOI: 10.1080/07350015.2019.1617158
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jbes.

for the presence of residual noise, thus solving partially the problem. Yet, a specification of ϕ remains necessary, which is the concern of this paper. Finally, Clinet and Potiron (2017) can be seen as an extension of the original problem when considering other high frequency quantities, such as high frequency covariance, powers of volatility or volatility of volatility, and thus shares the same drawback.

Accordingly, our first objective in this article is to provide a selection criterion to discriminate members of a class of models considered. This is obtained working out the Bayesian information criterion (BIC), for which we show the consistency. Naturally this allows for a two-step volatility estimation method, that is, model selection in a first step and aforementioned general methods in a second step. We document in our numerical study that this BIC-based method outperforms the concurrent methods to estimate volatility. Note that the BIC has already been used in Da and Xiu (2017) for the same purpose, but with a different structure on the noise process. Indeed, the authors consider a non-observable noise process which follows a generalized moving average distribution, and use the BIC to determine the order of the process, whereas in the present paper the noise is assumed observable with no particular structure (apart from some ergodicity properties).

Our second objective in this article is to look at how the several models based on the aforementioned trading information perform. We find that the market microstructure noise is mostly explained by the signed quoted spread. This corroborates the empirical findings on French stocks of Clinet and Potiron (2019). The BIC selects most of the time a larger model that adds information about the signed effective spread and the duration time between two trades to that of signed quoted spread, indicating that there is meaningful information in those two variables too.

Our third objective in this article is to study the empirical properties of the signed quoted spread model. Roughly speaking, it is reasonably stable intraday but manifests variability across days and stocks. Thus our goal is to investigate (in the model) the relation between the efficient price and the vector of raw prices and to disentangle residual (in the sense that they are not intraday, but rather explains variability in days and stocks) sources of high frequency market microstructure noise. As discussed above, one would expect that the efficient price stays quite close to the mid price. In reality, how close from each other are the efficient and the mid price, and more generally what is its relation with the full vector of raw prices? What are the residual prevailing causes of noise?

Finding empirically such a high role for the signed quoted spread should not be surprising as its magnitude is classified as the leading financial measure of liquidity in the high frequency market microstructure noise regression from Aït-Sahalia and Yu (2009). The difference between our approach and that of the cited article essentially lies in the model employed for noise. The cited authors considered a common noise with no information, whereas we use the richer signed quoted spread model which notably includes the trade direction. Without this essential information, they estimated the daily noise variance but not the noise itself. Then, they worked with the daily noise variance as a predicted value in the linear regression model. In contrast with their method, we manage to estimate directly the noise and we actually insert the effective-over-quoted-spread parameter, that is, that of the signed quoted spread model, in the regression.

It is widely acknowledged that the market microstructure noise variance displays a U-shaped intraday pattern, that is, noise variance is high at the open of the trading day, low around noon, and somewhere in-between at the close of the trading day. Such pattern was first observed on stock returns, where the related empirical evidence seems to be traced back to Wood, McInish, and Ord (1985) and Harris and Gurel (1986). It is also well-known that the quoted spread exhibits similar intraday patterns (see McInish and Wood 1992), and this is also true for the noise variance (see, e.g., Chen and Mykland 2017, although more recently the pattern could be more adequately described as a half U-shape with no increasing at the end of the trading day). In the signed quoted spread model, we notice that the parameter is relatively constant intraday while the spread unquestionably manifests the (half) U-shaped pattern, suggesting that the spread is solely responsible for this pattern in the market microstructure noise variance. These findings are complementary to that of Christensen, Hounyo, and Podolskij (2018), where diurnal pattern are found to be accounting for a rather significant fraction of intraday variation in volatility. Moreover, we will see that the parameter stays also fairly stable whether we look at daily or stock variation.

We can sum up the main empirical findings on the signed quoted spread model as follows. The mid price is consistently quite close to the efficient price. There is additional information in trade direction in the sense that the efficient price is systematically between the mid price and the transaction price yet not equal to the mid price. The variability in the parameter is relatively small intraday. The variability across days and stocks is best explained by the tick-to-spread ratio, hinting that the first residual cause of noise is the discreteness. We determine the proportion of bid-ask bounce as the second residual source of noise. All those points will be carefully examined using a dataset consisting of all tick-by-tick transactions and quotes recorded between January 1, 2009, and December 31, 2017 from 50 constituents randomly selected from the S&P 500.

There are a few more prior studies related to our work, which stands at the intersection between the literature of financial econometrics and that of market microstructure. Diebold and Strasser (2013) introduced leading financial economic models to volatility estimation. On the grounds that the one-lag autocorrelation in mid price returns is often found positive empirically, Andersen, Cebiroglu, and Hautsch (2017) extended the usual martingale-plus-noise setting to allow for positivity in the one-lag serial autocorrelation.

We proceed as follows. The general model including several leading financial economics environment as submodels, along with maximum likelihood based strategies and goodness of fit are described in Section 2. Selection criterion and BIC-based volatility estimation are discussed in Section 3. An extensive numerical study is provided in Section 4. The empirical study including in particular data description, high frequency estimates of the parameter, volatility, and residual noise, a comparison between five models using goodness of fit, an interpretation

of the parameter in the signed quoted spread model in terms of efficient and raw prices relation, stability of the parameter, a description of the financial characteristics, an identification of the residual sources of noise via linear regression of the parameter on high-frequency financial characteristics and an investigation of the existence of a common market-wide factor in parameter measurements is detailed in Section 5. We conclude in Section 6. A proof of the BIC consistency can be found in the supplementary materials available online.

2. MODEL, ESTIMATION AND GOODNESS OF FIT

In this section, we introduce the general model. Furthermore, we discuss about two complementary estimation strategies based on maximum likelihood and goodness of fit to compare submodels of this general model.

2.1. Model

The market microstructure noise is frequently stated as a white noise in the financial econometrics literature following the model specification of Hasbrouck (1993). Beyond the bid-ask spread and transaction price effects in the Roll (1984) model, additional sources of noise have been examined. For example, the implications of the discreteness as a source of measurement error are studied in Gottlieb and Kalay (1985), Harris (1990a), Jacod (1996), and Delattre and Jacod (1997). Harris (1990b) used additional effects (e.g., adverse selection effects as in Glosten and Milgrom (1985), Glosten (1987), and Glosten and Harris (1988), see also Madhavan, Richardson, and Roomans 1997) in the market microstructure noise. Transaction costs are investigated in, for example, Huang and Stoll (1996), Chan and Lakonishok (1997), and Cao, Choe, and Hatheway (1997). In what follows, we will be considering the trade direction, the traded volume, the duration time between two trades, the quoted depth, and the quoted spread as observed variables.

Following Li, Xie, and Zheng (2016), Chaker (2017), and Clinet and Potiron (2019), we incorporate the observed microstructure variables explicitly in the noise. If we define the possibly non regular observation times as $0 \leq t_0, \ldots, t_N \leq T$, where T corresponds to the horizon time—typically 1 day—and N stands for the possibly random number of observations, the general model can be written as

$$\underbrace{Z_{t_i}}_{\text{transaction price}} = \underbrace{X_{t_i}}_{\text{efficient price}} + \underbrace{\underbrace{\phi(Q_{t_i},\theta_0)}_{\text{explicative part}} + \underbrace{\epsilon_{t_i}}_{\text{residual noise}}}_{\text{market microstructure noise}},$$

where Q_{t_i} are the observed variables, and ϕ is a function known to the econometrician. In this paper, we explicitly consider a five dimensional linear form of ϕ defined as

$$\phi(Q_{t_i},\theta_0) = I_{t_i}\theta_0^{(1)} + \frac{1}{2}S_{t_i}I_{t_i}\theta_0^{(2)} + V_{t_i}I_{t_i}\theta_0^{(3)} + \frac{I_{t_i}}{1+\Delta t_i}\theta_0^{(4)} + D_{t_i}I_{t_i}\theta_0^{(5)}, \quad (2)$$

where I_{t_i} corresponds to the trade direction, that is, 1 if the transaction at time t_i is buyer-initiated and -1 if seller-initiated, S_{t_i} stands for the dynamic quoted spread (of log-price), V_{t_i} is the traded volume, $\Delta t_i := t_i - t_{i-1}$ is the duration time between two successive observation times, and D_{t_i} is the quoted depth.[1] Table 1 provides a summary of some key submodels of (2) used throughout the article. In particular, the specification (1) rules out the possibility that the trade direction affects the efficient price. For example, Madhavan, Richardson, and Roomans (1997) incorporated this very realistic economic updating by market makers.

The submodel where $\theta_0^{(2)} = \theta_0^{(3)} = \theta_0^{(4)} = \theta_0^{(5)} = 0$ corresponds to the Roll model, in which $\theta_0^{(1)}$ stands for the effective half-spread. As Roll (1984) pointed out, there is no reason for the effective spread to be equal to the (daily averaged) quoted spread. As a point of fact, it has been frequently observed that the effective spread is smaller than the average quoted spread (see, e.g., Petersen and Fialkowski 1994, who points out the relatively high proportion of trades executed inside the quoted spread as a possible reason to interpret the gap). Compared to its form in Hasbrouck (2002), note that there is an extra error term in the market microstructure noise which corresponds to external shocks not captured by the spread component. Moreover, the submodel where $\theta_0^{(2)} = \theta_0^{(4)} = \theta_0^{(5)} = 0$ corresponds to the Glosten–Harris model.

The case where $\theta_0^{(1)} = \theta_0^{(3)} = \theta_0^{(4)} = \theta_0^{(5)} = 0$ will be referred as signed quoted spread model (or simply signed spread model). The parameter can be interpreted as the effective-to-quoted spread ratio and there is no (theoretical) restriction in its value. One obvious limitation of the model is that this ratio is constant. Figure 1 contains an example of the model. For present purposes, we assume that the residual error is null and provide some key features of the signed spread model in that case. First, the parameter can also be interpreted as the noise-over-signed spread ratio. In particular, note that the variance of the market microstructure noise is equal to the product of one quarter of the spread variance and $(\theta_0^{(2)})^2$. Second, there are three distinct regimes associated to the parameter value. The most relevant regime—we will see that this regime incontestably stands out on data—is when $0 \leq \theta_0^{(2)} \leq 1$, which corresponds to an environment where the efficient price lies between the mid price and the transaction price. This regime is such that the effective spread is smaller or equal to the quoted spread, which is in line with the aforementioned findings discussed on the Roll model. When $\theta_0^{(2)}$ equals unity, there is no additional informational content in the trade direction as the efficient price coincides with the mid price and the market microstructure noise is equal to the signed quoted half-spread. The closer $\theta_0^{(2)}$ gets to 0, the more informational content in trade direction in the sense that the efficient price is replicating the transaction price's moves around the mid price with a proportional coefficient equal to $(1-\theta_0^{(2)})$. When $\theta_0^{(2)}$ is null, the efficient price is equal to the transaction price, so that the market microstructure noise is zero. Intuitively, we expect that the parameter stays close to unity. If $\theta_0^{(2)} > 1$, then the efficient price is on the side of the mid price opposite to the transaction price. In case when $\theta_0^{(2)} < 0$, then the efficient price

[1] The ask (bid) depth specifies the volume available at the best ask (bid).

Table 1. Five models for the information process

	$M1$ Roll	$M2$ Signed Spread	$M3$ Signed spread plus Roll	$M4$ Signed spread Plus Roll Plus duration time	$M5$ General
$\phi(Q_{t_i},\theta)$	$I_{t_i}\theta^{(1)}$	$\frac{1}{2}I_{t_i}S_{t_i}\theta^{(2)}$	$I_{t_i}\theta^{(1)}+\frac{1}{2}I_{t_i}S_{t_i}\theta^{(2)}$	$I_{t_i}\theta^{(1)}+\frac{1}{2}I_{t_i}S_{t_i}\theta^{(2)}$ $+\frac{I_{t_i}}{1+\Delta t_i}\theta^{(4)}$	$I_{t_i}\theta^{(1)}+\frac{1}{2}I_{t_i}S_{t_i}\theta^{(2)}$ $+I_{t_i}V_{t_i}\theta^{(3)}+\frac{I_{t_i}}{1+\Delta t_i}\theta^{(4)}$ $+I_{t_i}D_{t_i}\theta^{(5)}$

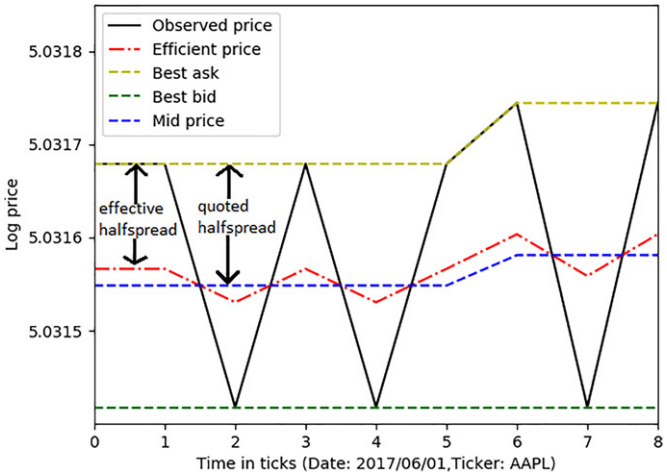

Figure 1. An example of the signed quoted spread model.

stays on the side of the transaction price opposite to the mid price.

In what follows, the latent log-price X_t is assumed to be an Itô-semimartingale of the form

$$dX_t = b_t dt + \sigma_t dW_t + dJ_t, \tag{3}$$

$$d\sigma_t = \widetilde{b}_t dt + \widetilde{\sigma}_t^{(1)} dW_t + \widetilde{\sigma}_t^{(2)} d\widetilde{W}_t + d\widetilde{J}_t, \tag{4}$$

where $(W_t, \widetilde{W}_t)$ is a 2 dimensional standard Brownian motion, the drift $(b_t, \widetilde{b}_t)$ is componentwise locally bounded, $(\sigma_t, \widetilde{\sigma}_t^{(1)}, \widetilde{\sigma}_t^{(2)})$[2] is componentwise locally bounded, itself an Itô process and

$$\inf_t(\min(\sigma_t, \widetilde{\sigma}_t^{(2)})) > 0$$

a.s. We further assume that $(J_t, \widetilde{J}_t)$ is a 2 dimensional pure jump process[3] of finite activity.

We now give the assumption on the observation times. In developing limit theory, we require a latent index n which will tend to infinity in our asymptotics. This is similar to, for example, the remark after Assumption 1 in Li, Zhang, and Zheng (2013). We consider the random discretization scheme which can be found in Clinet and Potiron (2018a, sec. 4) and which is adapted from Jacod and Protter (2011, sec. 14.1). We assume that there exists an Itô-semimartingale $\alpha_t > 0$ which satisfies Assumption 4.4.2, p. 115 in Jacod and Protter (2011) and is locally bounded and locally bounded away from 0, and iid $U_i > 0$ that are independent with each other and from other quantities such that

$$t_0 = 0, \tag{5}$$

$$t_i = t_{i-1} + \frac{T}{n}\alpha_{t_{i-1}} U_i. \tag{6}$$

We also assume that $\mathbb{E}U_i = 1$, and that for any $q > 0$, $m_q := \mathbb{E}U_i^q < \infty$, is independent of n. If we define $\pi_t := \sup_{i\geq 1} t_i - t_{i-1}$ and the number of observations before t as $N(t) = \sup\{i \in \mathbb{N} | 0 < t_i \leq t\}$ we have that $\pi_t \to^{\mathbb{P}} 0$ and that[4]

$$\frac{N(t)}{n} \to^{\mathbb{P}} \frac{1}{T}\int_0^t \frac{1}{\alpha_s} ds. \tag{7}$$

When there is no room for confusion, we sometimes drop T in the expression, that is, we use $N := N(T)$.

2.2. The Two Complemental Maximum-Likelihood Based Strategies

High frequency data based estimation on models including trading information such as the Roll model has been studied recently. We consider two completing maximum likelihood approaches based on two distinct scenarios, whether assuming that the residual error is null or nonzero vanishing asymptotically. The first approach provides an estimate of the parameter along with volatility, whereas the second one is also naturally equipped with an additional estimator of residual error variance. One might think that we could rule out the first method, but operating with the second method we find that the error related to the signed quoted spread model is very small, to the extent that it is virtually impossible to discriminate between the two scenarios. For clarity of exposition, we focus explicitly on the signed spread model hereafter, and emphasize that formula can adapt straightforwardly for the other submodels.

2.2.1. First Approach. The first maximum likelihood approach takes its roots in this simple model of returns where the error is null

$$\Delta Z_{t_i} = \Delta X_{t_i} + \frac{1}{2}\Delta(IS)_{t_i}\theta_0^{(2)}, \tag{8}$$

[2] A very good review on the use of stochastic volatility in financial mathematics is available in Ghysels, Harvey, and Renault (1996).

[3] Jumps in volatility have been observed in, for example, Todorov and Tauchen (2011).

[4] Actually the convergence is *u.c.p*, that is, uniformly in probability on $[0,t]$ for any $t \in [0,T]$. Equation (7) can be shown using Lemma 14.1.5 in Jacod and Protter (2011). The uniformity is obtained as a consequence of the fact that N_n and $\int_0^{\cdot}\frac{1}{\alpha_s}ds$ are increasing processes and Property (2.2.16) in Jacod and Protter (2011).

where $\Delta Y_{t_i} := Y_{t_i} - Y_{t_{i-1}}$, $X_t = \sigma_0 W_t$ with W_t a standard Brownian motion, that is, the volatility is constant and the drift is null, and the observations are regular $\Delta t_i = T/N$ with $N = n$. Only for the sake of statistical purposes and undeniably economically untrue, we can consider the spread component as the signal polluted by shocks in the efficient price. In that case, we obtain that the maximum likelihood estimator, which coincides with the least square estimator, is given by

$$\widehat{\theta}_0^{(2)} = 2\frac{\sum_{i=1}^N \Delta(IS)_{t_i}\Delta Z_{t_i}}{\sum_{i=1}^N \Delta(IS)_{t_i}^2}, \text{ and} \tag{9}$$

$$\widehat{\sigma}_0^2 = T^{-1}\sum_{i=1}^N \Delta\widehat{X}_{t_i}^2, \text{ where } \widehat{X}_{t_i} = Z_{t_i} - \frac{1}{2}I_{t_i}S_{t_i}\widehat{\theta}_0^{(2)}. \tag{10}$$

In in-fill asymptotics, that is, when $\Delta t_i \to 0$, Chaker (2017, Theorems 1 and 4) showed the consistency and the central limit theory of both estimators when the price is a continuous Itô-semimartingale, and the error possibly not null. Li, Xie, and Zheng (2016) demonstrated the consistency of the parameter estimate (Theorem 1) along with the central limit theory of the volatility estimator (Theorem 2) when the efficient price incorporates possible jumps and observations are not regular. As a corollary to our work, we prove in Clinet and Potiron (2019) the central limit theory related to the parameter $\theta_0^{(2)}$.

2.2.2. Second Approach. The second approach is based on the same model, while incorporating additional residual noise, which can be defined as

$$\Delta Z_{t_i} = \Delta X_{t_i} + \frac{1}{2}\Delta(IS)_{t_i}\theta_0^{(2)} + \Delta\epsilon_{t_i}. \tag{11}$$

In case when there is no spread component in this model, that is, $\theta_0^{(2)} = 0$, and the volatility process is assumed to be a constant σ_0, it has long been observed that the observed returns exhibit a MA(1) form. If we postulate that the error is normally distributed with variance a_0^2, the related log-likelihood is

$$l(\sigma^2, a^2) = -\frac{1}{2}\log\det(\Omega) - \frac{N}{2}\log(2\pi) - \frac{1}{2}\Delta Z^T\Omega^{-1}\Delta Z, \tag{12}$$

where $\Delta Z := (\Delta Z_{t_1}, \ldots, \Delta Z_{t_N})$ and Ω is the matrix

$$\Omega = \begin{pmatrix} \sigma^2 T/N + 2a^2 & -a^2 & 0 & \cdots & 0 \\ -a^2 & \sigma^2 T/N + 2a^2 & -a^2 & \ddots & \vdots \\ 0 & -a^2 & \sigma^2 T/N + 2a^2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -a^2 \\ 0 & \cdots & 0 & -a^2 & \sigma^2 T/N + 2a^2 \end{pmatrix}. \tag{13}$$

(14)

When a_0^2 is not shrinking to 0, Aït-Sahalia, Mykland, and Zhang (2005) established the consistency and central limit theory of the maximum likelihood estimator of σ_0^2 and a_0^2 and its robustness to a nonzero drift, which is anyway economically irrelevant at the frequencies we consider, and a nonnormally distributed noise with constant variance. Subsequently, Xiu (2010) demonstrated the robustness of the approach when the price is a continuous Itô-semimartingale and the volatility is changed to integrated volatility. Other papers related to this maximum likelihood estimator and inclusive of the specific case where a_0^2 goes to 0 asymptotically are available in the literature. For example, Aït-Sahalia and Xiu (2016) tested for the presence of market microstructure noise based on Hausman statistics constructed with such estimation procedure. Clinet and Potiron (2018a) showed that the estimators are efficient when performed locally, and robust to jumps and nonregular sampling times. Potiron and Mykland (2016) also considered a local version of this estimator in their examples. Da and Xiu (2017) considered a more general setting where in particular noise is no longer iid.

If we define $\widetilde{Z}_{t_i}(\theta) := Z_{t_i} - \frac{1}{2}I_{t_i}S_{t_i}\theta_0$, we can easily see that $\Delta\widetilde{Z}(\theta_0)$ displays a MA(1) dynamic so that we can specify the partial log-likelihood as

$$\widetilde{l}(\sigma^2, \theta, a^2) = -\frac{1}{2}\log\det(\Omega) - \frac{N}{2}\log(2\pi) - \frac{1}{2}\Delta\widetilde{Z}(\theta)^T\Omega^{-1}\Delta\widetilde{Z}(\theta). \tag{15}$$

We show in Clinet and Potiron (2019) the consistency and the central limit theory of the estimator related to (15) in case when the efficient price is an Itô-semimartingale with possible jumps and observations are nonregular, and the information process *IS* is stationary and satisfies reasonable stability conditions.

2.3. A Measure of Goodness of Fit to Compare Submodels

To assess the performance of the signed spread model or any other submodel in terms of residual error magnitude, we consider a measure of goodness of fit already employed in Li, Xie, and Zheng (2016, Remark 8, p. 37). Here again we leave out the other submodels, although all the subsequent definitions can adapt directly. Assuming that $I_{t_i}S_{t_i}$ and ϵ_{t_i} are centered and that $\mathbb{E}I_{t_i}^2S_{t_i}^2 = \mathbb{E}S_{t_i}^2 = a_{IS}^2$ and $\mathbb{E}\epsilon_{t_i}^2 = a_0^2$, we define the proportion of variance explained as

$$\pi := \frac{a_{IS}^2(\theta_0^{(2)})^2/4}{a_{IS}^2(\theta_0^{(2)})^2/4 + a_0^2}. \tag{16}$$

Roughly speaking, the closer to 1 the proportion of variance explained, the smaller the impact of the residual noise. An estimator of the proportion of variance explained is naturally given by

$$\widehat{\pi} = \frac{\widehat{a}_{IS}^2(\widehat{\theta}_0^{(2)})^2/4}{\widehat{a}_{IS}^2(\widehat{\theta}_0^{(2)})^2/4 + \widehat{a}_0^2}, \tag{17}$$

where the empirical variance of the signed spread is defined as

$$\widehat{a}_{IS}^2 := \frac{1}{N+1}\sum_{i=0}^N S_{t_i}^2, \tag{18}$$

and the parameter and noise variance parameters are obtained with the second aforementioned approach.

3. SELECTION CRITERION AND VOLATILITY ESTIMATION

In this section, we provide a selection criterion to discriminate between the submodels of (2), and selection criterion-based volatility estimation.

3.1. Bayesian Information Criterion

We now introduce the BIC related to the likelihood function with no residual noise (i.e., adopting the first approach)

$$l_{\exp}(\sigma^2,\theta) := \widetilde{l}(\sigma^2,\theta,0), \tag{19}$$

where $\widetilde{l}$ was introduced in (15). In what follows, we prove that when there is no residual noise, the BIC-based model selection scheme is consistent, in the sense that given a set of competing models containing the actual model that generated the observations, the correct model is picked by the procedure with probability tending to 1. Note that we limit ourselves to the case $a_0^2 = 0$ because our empirical study suggests that in practice the best submodels of (2) feature an estimated proportion of variance explained $\hat{\pi}$ very close to 100%.

Let $\Theta \subset \mathbb{R}^{\bar{d}}$, $\bar{d} \geq 1$, represent the maximal model for the information process (more precisely, Θ is the set of admissible values for $\widehat{\theta_0}$). We then consider a family of submodels $\mathcal{M}$ of the form $m = \{\theta = (\theta^{(1)},\dots,\theta^{(\bar{d})}) \in \Theta | \exists (i_1,\dots,i_p), \theta^{(i_1)} = \cdots = \theta^{(i_p)} = 0\}$. Then we say that m has $\bar{d} - p$ parameters, and we often identify m as a subset of $\mathbb{R}^{\bar{d}-p}$ for convenience. By assumption, there exists $m_0 \in \mathcal{M}$ and $d_0 \geq 1$ such that: (i) d_0 is the number of parameters of m_0 (ii) $\theta_0 \in m_0$, (iii) there is no $m \in \mathcal{M}$ with less than d_0 parameters which contains θ_0. We call $l_{\exp}^{(m)}$ the restriction of $l_{\exp}$ to the subset (submodel) m, and accordingly we write $\widehat{\upsilon}^{(m)} = ((\widehat{\sigma}^2)^{(m)}, \widehat{\theta}^{(m)})$ the related maximum likelihood estimator. Finally, we define BIC for each model $m \in \mathcal{M}$ with d parameters as

$$\mathrm{BIC}(m) = d\log N - 2l_{\exp}^{(m)}(\widehat{\upsilon}^{(m)}).$$

A selected model is thus defined as

$$\widehat{m}_{\mathrm{BIC}} \in \mathrm{argmin}_{m\in\mathcal{M}}\mathrm{BIC}(m).$$

Similarly we call $\widehat{d}_{BIC}$ the estimated number of parameters.

Proposition 1. Assume [A] and [B]. Moreover, assume that there is no residual noise ($\epsilon_t = 0$). We have

$$\mathbb{P}\left[\widehat{m}_{\mathrm{BIC}} = m_0\right] \to 1.$$

In particular,

$$\widehat{d}_{\mathrm{BIC}} \to^{\mathbb{P}} d_0.$$

The detailed proof along with assumptions **[A]** and **[B]** are relegated to the appendix in the supplementary material.

Remark 1 (Generalized information criterion). A careful inspection of the proof of Proposition 1 shows that the consistency property remains true if BIC is replaced by any generalized information criterion of the form

$$\mathrm{GIC}(m) = dg(N) - 2l_{\exp}^{(m)}(\widehat{\upsilon}^{(m)}),$$

where g is any function such that $g(N) \to +\infty$ and $g(N)/N \to 0$ as $n \to +\infty$.

Remark 2 (BIC in the presence of residual noise). It is possible to take into account a possible nonzero residual noise ϵ_t by simply replacing in the BIC formula the term $l_{\exp}(\widehat{\upsilon}^{(m)})$ by the general log-likelihood $\widetilde{l}(\widehat{\xi}^{(m)})$ where $\widehat{\xi}^{(m)}$ is one maximizer of $\widetilde{l}$ restricted to the submodel m. Following a similar proof as that of Proposition 1 and using Theorem 4.1 from Clinet and Potiron (2019), the consistency of the BIC could be adapted in that case.

3.2. Volatility Estimation

In this section, the object of interest is the so-called quadratic variation defined as

$$T\overline{\sigma}_0^2 := \int_0^T \sigma_s^2 ds + \sum_{0<s\leq T} \Delta J_s^2,$$

where $\Delta J_s = J_s - J_{s-}$. In addition to the two maximum likelihood based procedures described in Section 2.2, Li, Xie, and Zheng (2016) also provide a small noise robust estimation method and Chaker (2017) employs a modification of the two-scale realized volatility (TSRV) estimator from Zhang, Mykland, and Aït-Sahalia (2005). If the econometrician knows a priori the model for the market microstructure noise (e.g., null, including explicative part of a known specific model but no residual noise, including both explicative part of a known specific model and residual noise). In practice, this is not the case.

Subsequently, Clinet and Potiron (2019) provide in Section 4.3 a sequence to estimate volatility $\widehat{\sigma}_{\mathrm{seq}}^2$ defined as

$$\widehat{\sigma}_{\mathrm{seq}}^2 := \widehat{\sigma}_{\mathrm{RV}}^2 \text{ if no market microstructure noise,} \tag{20}$$

$$\widehat{\sigma}_{\mathrm{seq}}^2 := \widehat{\sigma}_{\mathrm{MLE1}}^2 \text{ if explicative part with no residual noise,} \tag{21}$$

$$\widehat{\sigma}_{\mathrm{seq}}^2 := \widehat{\sigma}_{\mathrm{MLE2}}^2 \text{ if both explicative part and residual noise,} \tag{22}$$

where $\widehat{\sigma}_{\mathrm{RV}}^2 = T^{-1}\sum_i (Z_{t_i} - Z_{t_{i-1}})^2$ is the realized volatility estimator, $\widehat{\sigma}_{\mathrm{MLE1}}^2$ the volatility estimator related to the first maximum likelihood approach, and $\widehat{\sigma}_{\mathrm{MLE2}}^2$ the volatility estimator related to the second maximum likelihood approach. The three possible scenarios are discriminated as follows. The econometrician should implement the Hausman tests from Aït-Sahalia and Xiu (2016) to test for the presence of market microstructure noise. If those tests indicate the presence of noise, adaptation of those tests should also be implemented to test for the presence of residual noise with a given explicative part. This still requires in practice to choose a submodel of the explicative part of the market microstructure noise, and this choice is not discussed in the article. To see the problem more formally, we have that

$$\widehat{\sigma}_{\mathrm{seq}}^2 := \widehat{\sigma}_{\mathrm{seq}}^2(m_{\mathrm{sel}}), \tag{23}$$

for some ad-hoc $m_{\mathrm{sel}} \in \mathcal{M}$ chosen by the econometrician. Thus the simple adaptation we consider consists in picking a submodel in a first step by BIC, and then plugging it in (23). The resulting quadratic variation estimator can be thus defined as

$$\widehat{\sigma}_{\mathrm{sel}}^2 := \widehat{\sigma}_{\mathrm{seq}}^2(\widehat{m}_{\mathrm{BIC}}). \tag{24}$$

4. FINITE SAMPLE

We carry out a Monte Carlo exercise to assess finite sample performance of the two maximum likelihood approaches, estimation and confidence intervals of the proportion of variance explained, model selection, and volatility estimation.

4.1. Setup

We draw $M = 1000$ Monte Carlo paths of high-frequency returns, for which $T = 1/252$ is annualized. One working day stands for 6.5 hr of trading activity, that is, $n = 23{,}400$ sec.

4.1.1. The Efficient Price. We bring forward the Heston model with U-shape intraday seasonality component and jumps in both price and volatility, which is defined as

$$dX_t = bdt + \sigma_t dW_t + dJ_t,$$
$$\sigma_t = \sigma_{t-,U}\sigma_{t,SV},$$

where

$$\sigma_{t,U} = C + Ae^{-at/T} + De^{-c(1-t/T)} - \beta\sigma_{\tau-,U}\mathbb{1}_{\{t\geq\tau\}},$$
$$d\sigma^2_{t,SV} = \alpha(\bar{\sigma}^2 - \sigma^2_{t,SV})dt + \delta\sigma_{t,SV}d\bar{W}_t,$$

with $b = 0.03$, $dJ_t = \nabla S_t dN_t$, $\nabla = T\bar{\sigma}^2$, the signs of the jumps $S_t = \pm 1$ are iid symmetric, N_t is a homogeneous Poisson process with parameter $\bar{\lambda} = T$ so that the contribution of jumps to the total quadratic variation of the price process is about 50%, $C = 0.75$, $A = 0.25$, $D = 0.89$, $a = 10$, $c = 10$, the volatility jump size parameter $\beta = 0.5$, the volatility jump time τ follows a uniform distribution on $[0, T]$, $\alpha = 5$, $\bar{\sigma}^2 = 0.1$, $\delta = 0.4$, $\bar{W}_t$ is a standard Brownian motion such that $d\langle W, \bar{W}\rangle_t = \overline{\phi}dt$, $\overline{\phi} = -0.75$, $\sigma^2_{0,SV}$ is sampled from a Gamma distribution of parameters $(2\alpha\bar{\sigma}^2/\delta^2, \delta^2/2\alpha)$, which corresponds to the stationary distribution of the CIR process. One can consult Clinet and Potiron (2018a) to obtain more information about the model, which strongly takes its roots in Andersen, Dobrev, and Schaumburg (2012) and Aït-Sahalia and Xiu (2016).

4.1.2. The Observation Times. We consider two levels of sampling: tick-by-tick and 5 sec. For the latter, the observation times are generated regularly. For the tick-by-tick case, we assume that $\alpha_t = 1/(e^{\beta_1} + \{e^{\beta_2} + e^{\beta_3}\}^2(t/T - e^{\beta_2}/(e^{\beta_2} + e^{\beta_3}))^2)$, and that U_i are drawn from an exponential distribution with parameter $2T/23400$. We have that the rate of arrival times α_t^{-1} manifests a usual U-shape intraday pattern, as discussed in Engle and Russell (1998, see discussions in secs. 5 and 6 and Figure 2) and Chen and Hall (2013, sec. 5, pp. 1011–1017). We fix $\beta_1 = -0.84$, $\beta_2 = -0.26$, and $\beta_3 = -0.39$ following the empirical values exhibited in Clinet and Potiron (2018b), thus the sampling frequency is on average slightly faster than 1 sec.

4.1.3. The Information. We implement five variables: trade direction, traded volume, the duration time between two trades, the quoted depth, and the quoted spread. The trade direction I_{t_i} is simulated featuring a Bernoulli process with parameter $p = 1/2$ and with a serial autocorrelation chosen equal to 0.3. The traded volume is simulated as an AR(1) process with mean 100, variance equal to 165,000 and autocorrelation parameter set to 0.017. The duration time between two trades is simulated as described in the above section. The quoted depth follows an AR(1) process with corresponding mean equal to 180, a variance equal to 27,500, and a serial autocorrelation parameter set to 0.47. For each path, the quoted spread is simulated as

$$S_i := TS(1 + B_i^{(S)}), \tag{25}$$
$$p_i := \max(0, \min(1, p_{i-1} + \rho(2B_i^{(p)} - 1)), \tag{26}$$

where TS, which stands for "tick size," is fixed to 0.0001,[5] the size of the spread in ticks $B_i^{(S)}$ follows a binomial distribution of max size 4 and probability p_i, the probability p_i is a random walk bounded between 0 and 1 with increment equal to $\rho = 0.1$, and $B_i^{(p)}$ is a Bernoulli distribution with probability parameter 0.5. In (25) and (26), the closer p_i to 1, the closer the quoted spread to 4, and the closer p_i to 0, the closer the quoted spread to 1. The parameter ρ is related to the volatility of the spread.

We implement three "true" models: the signed spread model ($M2$), the signed spread plus Roll model ($M3$), and the signed spread plus Roll plus duration time model ($M4$). The other two variables from the general model (2), that is, quoted depth and traded volume, are consistently reported as nonsignificant for most days and stocks in our empirical study, which is the reason why we do not incorporate them in possible components of the "true" model. In $M2$, we fix $\theta_0^{(2)} = 0.86$. In $M3$, we fix $\theta_0^{(1)} = -10^{-5}$ and $\theta_0^{(2)} = 0.90$. In $M4$, we fix $\theta_0^{(1)} = -10^{-5}$, $\theta_0^{(2)} = 0.90$, and $\theta_0^{(4)} = 10^{-5}$. The values of the parameters correspond roughly to the fitted values.[6] Note that although the quoted depth and the traded volume are not used for generating the "true" model, they are nonetheless required to implement the maximum likelihood approaches on the general model.

4.1.4. The Residual Noise. When there is residual noise in the model, we fix the proportion of variance explained to 90%. In terms of residual variance, we have that the residual variance a_0^2 is ranging from 1.4×10^{-9} to 1.6×10^{-9} depending on the model considered, which is in line with the empirical findings.

4.1.5. Truncation Method to Estimate the Asymptotic Variance. Despite the fact that both maximum likelihood estimation approaches are jump-in-price-robust in view of Theorem 3.1 in Clinet and Potiron (2019), the asymptotic variance estimators require truncation (see the expression of $\widehat{V}_2$ in (3.14) in the cited paper). If we introduce $\widetilde{k}$ which is random and satisfies $\widetilde{k}T/N \to^{\mathbb{P}} 0$ and $\widetilde{u}_i = \widetilde{\alpha}(t_i - t_{i-1})^{\omega}$, the asymptotic variance estimators use truncated expressions such as $\mathbf{1}_{\{|\Delta\widehat{X}_i| \leq \widetilde{u}_i\}}$, and spot volatility estimation on the block of size $\widetilde{k}$. We choose $\omega = 0.48$, $\widetilde{\alpha} = \alpha_0\widehat{\sigma}_{\text{exp}}$, $\alpha_0 = 4$, and $\widetilde{k} = \lfloor N^{1/2} \rfloor$, consistently with the cited paper.

[5] Here, the tick size is much smaller than the typical value of 0.01 units for stocks, because it refers to the tick size in log price. For instance a 0.01 tick size on a 30 units price gives a tick on log-price equals to $\log(30) - \log(29.99) \approx 0.0001$, which corresponds roughly to our setting. Of course our setting is not realistic in the sense that the tick in log-price is fixed whereas it is not the case in practice, but this does not affect estimation procedure as the price intraday price variation is very small.

[6] This is nonetheless not fully reported in the empirical study.

Table 2. Finite sample properties of the two maximum likelihood approaches when residual noise is null

Samp. freq.	Param.	MLE	Bias	SD	0.50%	2.50%	5.00%	95.00%	97.50%	99.50%
Unfeasible statistics										
Tick-by-tick	σ_0^2	1	0.049	1.401	0.08%	0.79%	1.90%	99.54%	99.89%	99.99%
5-sec	σ_0^2	1	−0.041	1.020	0.38%	2.12%	4.46%	95.67%	97.98%	99.69%
Tick-by-tick	θ_0	1	0.040	0.770	2.93%	7.05%	10.79%	90.51%	94.04%	98.18%
5-sec	θ_0	1	−0.024	0.857	1.40%	4.90%	7.38%	91.74%	95.60%	98.95%
Tick-by-tick	σ_0^2	2	0.041	1.150	0.34%	1.58%	3.48%	97.54%	99.04%	99.97%
5-sec	σ_0^2	2	0.000	1.025	0.38%	1.82%	4.59%	95.11%	97.71%	99.69%
Tick-by-tick	θ_0	2	0.043	0.825	2.14%	5.78%	8.96%	91.75%	95.13%	98.69%
5-sec	θ_0	2	−0.025	0.859	1.29%	4.65%	7.47%	91.90%	95.78%	98.94%
Tick-by-tick	a_0^2	2	−0.018	1.032	0.16%	1.78%	4.00%	95.09%	97.82%	99.28%
5-sec	a_0^2	2	−0.029	1.002	0.32%	2.13%	4.95%	94.99%	96.97%	99.41%
Feasible statistics										
Tick-by-tick	σ_0^2	1	0.005	1.452	0.12%	0.38%	1.16%	99.46%	99.89%	99.99%
5-sec	σ_0^2	1	−0.087	1.164	0.10%	0.73%	2.28%	96.01%	98.39%	99.73%
Tick-by-tick	θ_0	1	0.030	0.809	2.41%	6.13%	9.36%	90.75%	94.57%	99.25%
5-sec	θ_0	1	−0.019	0.906	0.93%	4.30%	7.13%	92.36%	96.15%	99.52%
Tick-by-tick	σ_0^2	2	0.015	1.212	0.13%	0.89%	2.15%	97.74%	99.01%	99.98%
5-sec	σ_0^2	2	−0.025	1.135	0.03%	0.76%	2.96%	96.10%	98.38%	99.70%
Tick-by-tick	θ_0	2	0.043	1.174	0.19%	1.29%	2.86%	97.63%	99.15%	99.94%
5-sec	θ_0	2	−0.030	1.223	0.04%	0.99%	2.28%	97.41%	99.14%	99.98%
Tick-by-tick	a_0^2	2	−0.016	1.094	0.08%	1.66%	3.40%	96.22%	98.40%	99.78%
5-sec	a_0^2	2	−0.031	1.097	0.17%	1.67%	3.79%	96.11%	98.35%	99.90%

NOTES: This table shows summary statistics and empirical quantiles benchmarked to the $N(0,1)$ distribution for the infeasible and feasible Z-statistics related to the two maximum likelihood approaches when residual noise is null. The simulation design is $M2$ with $M = 1000$ Monte Carlo paths. The column "SD" reports the standard errors.

4.1.6. Concurrent Volatility Estimators and Simulated Data for Comparison. To assess the performance of the BIC-based volatility estimator, we consider a group of eight concurrent alternative volatility estimators which is a mix of estimators considered in Clinet and Potiron (2019) and leading estimators from the literature. First, we have $\widehat{\sigma}^2_{\mathrm{seq}}(M1)$, $\widehat{\sigma}^2_{\mathrm{seq}}(M2)$, $\widehat{\sigma}^2_{\mathrm{seq}}(M3)$, $\widehat{\sigma}^2_{\mathrm{seq}}(M4)$, and $\widehat{\sigma}^2_{\mathrm{seq}}(M5)$. We also implement some popular estimators: QMLE of Aït-Sahalia, Mykland, and Zhang (2005), pre-averaging estimator (PAE) from Jacod et al. (2009), realized kernels (RK) in Barndorff-Nielsen et al. (2008).[7]

The simulated model for the market microstructure noise is picked uniformly from $\{M2, M3, M4\}$, and then the presence of residual noise in the simulated model is further given uniformly by a Bernoulli variable. Finally, as in Clinet and Potiron (2019) we consider volatility estimation in the absence of jumps as most methods are actually not robust to this feature.

4.2. Results and Discussion

4.2.1. Finite Sample Properties of the Two Maximum Likelihood Approaches. In this part, we study nonfeasible and feasible version of Theorem 3.1 in Clinet and Potiron (2019). In particular, in the notation of the cited paper, we estimate the quarticity, where the formal definition is given in Theorem 3.1 in the cited paper, $\mathcal{Q}$ with $\widehat{V}_2$ (whose formal definition can be found in (3.14) in the cited paper) and $U_{\theta_0}^{-1}$ as $U_{\widehat{\theta}_0}^{-1}$ (whose formal definitions can be found in Theorem 3.1).

Table 2 reports the finite sample properties of the two maximum likelihood approaches when residual noise is null. Both approaches are robust to this framework. Given all the misspecification of the likelihood function—time-varying volatility, nonregular observation times, the results are quite decent. The bias is relatively small for all the statistics, and the standard deviation is ranging from around 0.80 to 1.40, which is reasonably close to unity.

Table 3 reports the finite sample properties of the second likelihood approach when residual noise is nonzero, that is, with related proportion of variance explained set to 90%. We do not report the results of the first approach as it is not designed for this framework. In that case, the studentized statistics standard deviation and bias explode, and this is most likely due to the fact that the residual noise is so small (i.e., with variance $a_0^2 \approx 1.60 \times 10^{-9}$, which is in line with the values exhibited in the empirical study) to the extent that the large noise asymptotic is not appropriate to explain the finite sample of the normalized statistics.

4.2.2. Estimation and Confidence Intervals of the Proportion of Variance Explained. Table 4 reports summary statistics and confidence intervals for the noise variance and the proportion of variance explained when residual noise is null or nonzero. The three confidence intervals are computed using the central limit theory with three different asymptotics: no noise, small noise, and large noise. The theoretical validation of the confidence intervals in the no noise and large noise case are

[7]Details on the choice of tuning parameters for the PAE and the RK can be obtained upon request to the authors.

Table 3. Finite sample properties of the second maximum likelihood approach when nonzero residual noise

Samp. freq.	Param.	MLE	Bias	SD	0.50%	2.50%	5.00%	95.00%	97.50%	99.50%
Unfeasible statistics										
Tick-by-tick	σ_0^2	2	0.055	1.972	0.00%	0.01%	0.13%	99.96%	99.99%	100.00%
5-sec	σ_0^2	2	−0.025	2.335	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
Tick-by-tick	θ_0	2	0.040	2.459	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
5-sec	θ_0	2	−0.057	4.441	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
Tick-by-tick	a_0^2	2	−0.338	13.185	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
5-sec	a_0^2	2	−0.174	39.401	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
Feasible statistics										
Tick-by-tick	σ_0^2	2	0.176	4.846	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
5-sec	σ_0^2	2	0.269	2.727	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
Tick-by-tick	θ_0	2	0.046	2.459	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
5-sec	θ_0	2	−0.057	4.441	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
Tick-by-tick	a_0^2	2	−2.486	19.140	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
5-sec	a_0^2	2	−121.5	1969	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%

NOTES: This table shows summary statistics and empirical quantiles benchmarked to the $N(0,1)$ distribution for the infeasible and feasible Z-statistics related to the second maximum likelihood approach when residual noise is nonzero. The simulation design is $M2$ with $M = 1000$ Monte Carlo paths. The column "SD" reports the standard errors.

Table 4. Estimation and confidence intervals for the proportion of variance explained

Sp. freq.	Par.	Mod.	Mean	SD	CI1 size	CI1 cvr.	CI2 size	CI2 cvr.	CI3 size	CI3 cvr.
No residual noise										
Tick-by-tick	a_0^2	$M2$	$-2.83e^{-12}$	$1.88e^{-10}$	$8.79e^{-10}$	92.1%	$8.84e^{-10}$	93.0%	$1.93e^{-11}$	3.0%
5-sec	a_0^2	$M2$	$-1.66e^{-11}$	$1.24e^{-9}$	$5.77e^{-9}$	92.9%	$5.84e^{-9}$	94.7%	$1.38e^{-10}$	5.5%
Tick-by-tick	π	$M2$	1.000	0.014	0.063	91.9%	0.063	92.8%	0.001	3.0%
5-sec	π	$M2$	1.018	0.136	0.594	93.0%	0.621	96.2%	0.013	5.5%
Tick-by-tick	a_0^2	$M3$	$-3.10e^{-12}$	$1.87e^{-10}$	$8.79e^{-10}$	92.1%	$8.83e^{-10}$	92.8%	$1.92e^{-11}$	3.0%
5-sec	a_0^2	$M3$	$-2.33e^{-11}$	$1.24e^{-9}$	$6.82e^{-9}$	97.8%	$6.88e^{-9}$	98.2%	$1.62e^{-10}$	4.8%
Tick-by-tick	π	$M3$	1.000	0.014	0.066	91.9%	0.063	92.8%	0.001	2.9%
5-sec	π	$M3$	1.001	0.132	0.598	97.7%	0.621	98.8%	0.017	4.4%
Tick-by-tick	a_0^2	$M4$	$-3.19e^{-12}$	$1.24e^{-9}$	$6.83e^{-9}$	92.2%	0.063	92.9%	0.001	2.9%
Tick-by-tick	π	$M4$	1.000	0.014	0.066	91.9%	0.063	92.8%	0.017	4.4%
Nonzero residual noise										
Tick-by-tick	a_0^2	$M2$	$1.60e^{-9}$	$1.92e^{-10}$	$8.79e^{-10}$	91.5%	$8.91e^{-10}$	92.7%	$1.93e^{-11}$	3.0%
5-sec	a_0^2	$M2$	$1.55e^{-9}$	$1.27e^{-9}$	$5.76e^{-9}$	92.0%	$5.92e^{-9}$	94.1%	$1.38e^{-10}$	5.5%
Tick-by-tick	π	$M2$	0.900	0.010	0.061	94.4%	0.061	95.0%	0.001	2.2%
5-sec	π	$M2$	0.880	0.100	0.572	95.2%	0.643	98.2%	0.013	4.9%
Tick-by-tick	a_0^2	$M3$	$1.60e^{-9}$	$1.92e^{-10}$	$8.79e^{-10}$	91.5%	$8.91e^{-10}$	92.8%	$1.93e^{-11}$	3.0%
5-sec	a_0^2	$M3$	$1.45e^{-9}$	$1.24e^{-9}$	$6.82e^{-9}$	97.9%	$6.93e^{-9}$	98.4%	$1.75e^{-10}$	3.8%
Tick-by-tick	π	$M3$	0.900	0.010	0.063	94.4%	0.063	95.0%	0.001	2.2%
5-sec	π	$M3$	0.877	0.099	0.569	95.3%	0.641	98.3%	0.013	5.0%
Tick-by-tick	a_0^2	$M4$	$1.60e^{-9}$	$1.92e^{-10}$	$8.79e^{-10}$	91.5%	$8.91e^{-10}$	92.8%	$1.93e^{-11}$	3.0%
Tick-by-tick	π	$M4$	0.900	0.010	0.063	94.4%	0.063	95.0%	0.001	2.2%

NOTES: This table shows summary statistics and confidence intervals for the noise variance and the proportion of variance explained when residual noise is null or nonzero. In the no residual noise case, $a_0^2 = 0$ and $\pi = 1$. In the nonzero residual noise case, $a_0^2 = 1.60e^{-9}$ (on average) and $\pi = 0.90$. The confidence intervals are obtained, respectively, with asymptotics "no noise", "small noise" and "large noise". The columns "cvr." denotes the proportion of trajectories where the 95% confidence interval overlapped with the target value. The simulation design is $M2$, $M3$, and $M4$ with $M = 1000$ Monte Carlo paths. $M4$ does not include 5 sec regular setting because the duration time is included in that model. The column "SD" reports the standard errors.

straightforward corollaries of, respectively, Theorems 3.1 and 3.3 in Clinet and Potiron (2019). The small noise case is also a corollary of Theorem 3.1 but would require a longer proof. The main motivation behind implementing the confidence intervals is the fact that the proportion of variance explained is often estimated above 100% in our empirical study. We have chosen not to focus on the intervals in the theoretical part of our paper. Note that the quantities used to estimate the confidence intervals have already been discussed in Section 4.2.1.

The results of the estimation procedures are in line with the theory. As expected from the limit theory, we can see that overall the tick-by-tick based estimation procedure perform better than the 5 sec based estimation procedures. In addition, the presence of residual noise does not seem to be affecting

Table 5. Finite sample fit of the second maximum likelihood estimation approach

$\theta_0^{(1)}$ mean	$\theta_0^{(1)}$ SD	$\theta_0^{(2)}$ mean	$\theta_0^{(2)}$ SD	$\theta_0^{(3)}$ mean	$\theta_0^{(3)}$ SD	$\theta_0^{(4)}$ mean	$\theta_0^{(4)}$ SD	$\theta_0^{(5)}$ mean	$\theta_0^{(5)}$ SD
No residual noise, $M2$									
$-3.66e^{-8}$	$2.72e^{-6}$	0.86	0.01	$2.96e^{-12}$	$4.88e^{-12}$	$6.73e^{-8}$	$7.00e^{-7}$	$4.47e^{-13}$	$1.60e^{-10}$
No residual noise, $M3$									
$-1.05e^{-5}$	$2.74e^{-6}$	0.90	0.01	$-3.07e^{-12}$	$6.75e^{-12}$	$-7.07e^{-8}$	$7.21e^{-7}$	$1.74e^{-12}$	$1.62e^{-10}$
No residual noise, $M4$									
$-1.02e^{-5}$	$2.75e^{-6}$	0.90	0.01	$-2.76e^{-12}$	$7.67e^{-12}$	$1.07e^{-5}$	$6.92e^{-7}$	$1.26e^{-11}$	$1.61e^{-10}$
Nonzero residual noise, $M2$									
$6.92e^{-8}$	$3.19e^{-6}$	0.86	0.01	$4.53e^{-12}$	$2.50e^{-11}$	$-4.92e^{-8}$	$1.23e^{-6}$	$-1.06e^{-11}$	$1.84e^{-10}$
Nonzero residual noise, $M3$									
$-9.82e^{-6}$	$3.27e^{-6}$	0.90	0.01	$3.77e^{-12}$	$2.65e^{-11}$	$-6.15e^{-8}$	$1.43e^{-6}$	$-1.99e^{-11}$	$2.03e^{-10}$
Nonzero residual noise, $M4$									
$-9.76e^{-6}$	$3.14e^{-6}$	0.90	0.01	$2.64e^{-12}$	$2.81e^{-11}$	$1.04e^{-5}$	$1.31e^{-6}$	$1.16e^{-12}$	$1.88e^{-10}$

NOTES: This table shows finite sample fit of the second maximum likelihood estimation approach when residual noise is null or nonzero. The results are produced fitting the general model ($M4$) whereas the data are generated with "true" models $M2$, $M3$, and $M4$. For $M2$, we have $\theta_0^{(2)} = 0.86$ and $\theta_0^{(1)} = \theta_0^{(3)} = \theta_0^{(4)} = \theta_0^{(5)} = 0$. For $M3$, we have $\theta_0^{(1)} = -10^{-5}$, $\theta_0^{(2)} = 0.90$ and $\theta_0^{(3)} = \theta_0^{(4)} = \theta_0^{(5)} = 0$. For $M4$, we have $\theta_0^{(1)} = -10^{-5}$, $\theta_0^{(2)} = 0.90$, $\theta_0^{(4)} = 10^{-5}$ and $\theta_0^{(3)} = \theta_0^{(5)} = 0$. The simulation design is tick-by-tick with $M = 1000$ Monte Carlo paths. The table does not show 5 sec setting because the duration time is included in the general model. The column "SD" reports the standard errors.

Table 6. BIC

Model	$M2$ mean	$M2$ sel.	$M3$ mean	$M3$ sel.	$M4$ mean	$M4$ sel.	$M5$ mean	$M5$ sel.
No residual noise								
$M2$	−354,073	100.0%	−354,064	0.0%	−354,055	0.0%	−354,037	0.0%
$M3$	−354,003	5.2%	−354,064	94.8%	−354,055	0.0%	−354,037	0.0%
$M4$	−354,036	35.2%	−354,040	6.8%	−354,055	58.0%	−354,013	0.0%
Nonzero residual noise								
$M2$	−348,525	100.0%	−348,516	0.0%	−348,507	0.0%	−348,488	0.0%
$M3$	−348,751	6.6%	−348,783	93.4%	−348,773	0.0%	−348,749	0.0%
$M4$	−348,390	37.5%	−348,388	8.7%	−348,391	53.8%	−348,359	0.0%

NOTES: This table shows BIC mean and proportion of model selected for $M2$, $M3$, $M4$, and $M5$ when the data are generated with $M2$, $M3$, and $M4$. Each row includes a different setting which corresponds to the simulated model. Each column corresponds to different model used for selection. In particular the columns "mean" report the BIC mean, and "sel." the proportion of time the was model selected. The simulation design is tick-by-tick with $M = 1000$ Monte Carlo paths. The table does not show 5 sec setting because the duration time is included in the general model.

the performance of the procedure. In particular, the proportion of variance explained—which is equal to 100% in case of no residual noise and 90% when there is nonzero residual noise—is estimated just fine in both cases. As for the confidence intervals, the no noise and small noise cases produce almost the same results, whereas the large noise case is far off. This is in line with the results reported in Table 3. The large noise asymptotics is not adapted for residual noise whose variance is as small as $a_0^2 \approx 1.60 \times 10^{-9}$. Yet this is the typical value that we find empirically.

4.2.3. Finite Sample Fit of the Second Maximum Likelihood Estimation Approach on Different Models. Table 5 reports the finite sample fit of the second maximum likelihood estimation approach when fitted on the general model ($M5$). We can see that the procedure is working well overall. For $M2$, the signed spread parameter is nicely estimated and all the other parameters are close to 0, and nonsignificantly different from 0. For $M3$, the signed spread parameter and the Roll parameter are estimated very closely from their respective values, and the other parameters are not significantly different from 0. This also works similarly for $M4$.

4.2.4. Model Selection. Table 6 reports the BIC mean and proportion of model selected for $M2, M3, M4$, and $M5$ when the simulated model is $M2, M3$, or $M4$. When the simulated model is $M2$, $M2$ is selected all the time. When $M3$ is the simulated model, it is selected with a proportion around 0.95. When the simulated model is $M4$, the BIC is not as good, although $M4$ remains selected most of the time, around 55% of the days, and $M2$ around 35%. This is still reasonable. The reason why the BIC does not perform that good in that case might be that the trade direction and the signed duration time are highly correlated.

4.2.5. Comparison of the Performance of Volatility Estimators. Table 7 reports the performance of the BIC-based volatility estimator along with eight alternatives. Undoubtedly, the former performs better than the alternatives. The gain is small but not negligible when compared to the non BIC-based estimators of Clinet and Potiron (2019). The gain is much more substantial when compared to common volatility estimators from the literature such as QMLE, PAE, and RK, consistently with the numerical studies of Li, Xie, and Zheng (2016) and Clinet and Potiron (2019). In particular, this indicates that the

Table 7. Comparison of the performance of volatility estimators

Est.	Bias	SD	RMSE
$\widehat{\sigma}^2_{\text{sel}}$	2.04	8.69	8.93
$\widehat{\sigma}^2_{\text{seq}}(M1)$	3.29	8.78	9.38
$\widehat{\sigma}^2_{\text{seq}}(M2)$	2.12	8.73	8.98
$\widehat{\sigma}^2_{\text{seq}}(M3)$	2.11	8.73	8.98
$\widehat{\sigma}^2_{\text{seq}}(M4)$	2.09	8.73	8.98
$\widehat{\sigma}^2_{\text{seq}}(M5)$	2.06	8.76	9.00
QMLE	8.44	9.43	12.66
PAE	1.57	9.99	10.11
RK	8.44	9.83	12.96

NOTES: This table shows the bias and the standard deviation of several volatility estimators. Values are scaled by a factor 10^5. The simulation design is a mix of $M2$, $M3$, $M4$ for the explicative with a mixing of residual noise, tick-by-tick with $M = 1000$ Monte Carlo paths. The table does not show 5 sec setting because the duration time is included in the general model. The column "SD" reports the standard errors. The column "RMSE" reports the root mean square errors.

relative imprecision of the BIC in some cases do not impact too badly the BIC-based volatility estimation procedure.

5. EMPIRICAL ANALYSIS

We are now ready to analyze estimation results obtained on real data, and to compare five models based on the goodness of fit measure $\hat{\pi}$ and BIC. We will see that while more general models are often selected, the signed spread model already explains the major part of the market microstructure noise with a single parameter. Accordingly, we will work with this model to interpret its parameter mean value in terms of raw prices vector relation, to look at its parameter variability compared to the signed spread variability, and to look for the residual sources of noise by relating its parameter to various observable financial characteristics of the stocks.

5.1. Data Description and High Frequency Estimates of the Parameters, Volatility, and the Residual Noise Variance in Five Models

Our empirical study is conducted on a sample of 50 stocks randomly selected from the S&P 500 during the period between January 1st, 2009 and December 31st, 2017. Our dataset consists of intraday transaction price, bid price, ask price, trade time-stamps, traded volumes, and volumes standing at the first limits of the order book collected from the tick-by-tick Trade and Quote (TAQ) database. Following the prominent algorithm described in Lee and Ready (1991), we reconstruct the trade direction. The time resolution is equal to the microsecond.

We compute the likelihood (15) of the second approach based on sampling the data at the highest frequency available, that is, tick-by-tick. All along this empirical study, we consider five competing models, that is, $M1$–$M5$ described in Table 1.

For each model, we estimate for each stock i and day t the integrated volatility $\sigma_{i,t}$, the vector of parameters $\theta_{i,t}$ and the residual noise variance $a^2_{i,t}$. In addition, we estimate the daily variance of the signed spread defined in (18) as $a^2_{IS,i,t}$ and the noise-to-signal ratio for the signed spread model as $\xi^2_{i,t} := (a^2_{i,t} + (\theta^{(2)}_{i,t})^2 a^2_{IS,i,t})\sigma^{-2}_{i,t}$. We exclude stock-day combinations with fewer than 500 intraday transactions or with problems in the data (such as price values that are missing or dropping suddenly to 0, or bid price recorded higher than the ask price).

Table 8 reports the descriptive statistics related to the full period 2009–2017, number of days included and daily trades. 38 stocks over 50 are included during the full period, whereas the period of the others typically does not cover several years of the full period. This is due to the fact that S&P Dow Jones Indices updates the components of the S&P 500 periodically, mainly in response to acquisitions, or to keep the index up to date as various companies grow or shrink in value. The number of days included is on average equal to 1905, ranging from 543 to 2239. The total number of days included across the 50 stocks amounts to 95,255. The average number of daily trades is 18,400, with a minimum of 3900 for ARG and a maximum of 100,700 for AAPL.

Table 9 reports the basic summary statistics for the volatility $\sigma_{i,t}$, the noise-to-signal ratio $\xi^2_{i,t}$, the parameter in the signed spread model $\theta^{(2)}_{i,t}$ and the parameter in the Roll model $\theta^{(1)}_{i,t}$. The volatility (when normalized by the mid price) is estimated to 3.10×10^{-3} on average. The noise-to-signal ratio averages to 3.97×10^{-4}. The parameter in the signed spread model equals 0.861, while the parameter in the Roll model (scaled by the daily average half quoted spread) is slightly smaller, equal to 0.816. For the sake of conciseness, we do not report the parameters related to the concurrent models, although we document a model selection study in the next section.

We will employ the first maximum likelihood based estimates as control variables for the other approach results, by rerunning the regressions with the estimates coming for the first approach. By doing so, we make sure that unusual results are due to the data itself, and not a recurrent problem in the optimization procedure of the second likelihood method. The results are very similar and do not imply economically meaningful differences. For the sake of brevity, we do not report the related results obtained with the first approach.

5.2. Goodness of Fit and Model Selection for Five Competing Models

It is of interest to assess whether on real data a market microstructure noise model systematically stands out or not. Accordingly, we start our empirical investigation with a comparative study of the five explicative noise models, that is, $M1$–$M5$ described in Table 1. To do so, we examine them through the lenses of BIC-based model selection, proportion of market microstructure noise variance explained $\hat{\pi}$ and significance of parameters of the general model including all five variables.

We first consider our goodness-of-fit criterion $\hat{\pi}$ for each model. The mean and standard deviation (across stocks and days) of the proportion of variance explained are reported in Table 11. Apart from the Roll model ($M1$), not only the average $\hat{\pi}$ are all extremely close to 100%,[8] but their standard deviation

[8]Note that the presence of estimated proportions larger than 100% is due to the fact that the residual noise variance may be estimated negative when the

Table 8. Descriptive statistics: period, number of days included and daily trades in 2009–2017

Ticker	Period	Number of days included	Number of daily trades
AAPL	2009–2017	2232	100,700
ACE	2009–2016	1711	8900
ALTR	2009–2015, 2017	1690	14,000
ALXN	2009–2017	1772	8300
AMAT	2009–2017	2208	29,700
ARG	2009–2016	1813	3900
BCR	2009–2017	2188	4300
BMY	2009–2017	2218	26,200
CELG	2009–2017	2120	20,400
CHRW	2009–2017	2175	8400
CL	2009–2017	2236	13,700
CMI	2009–2017	2172	12,000
CMS	2009–2017	2216	9000
CPB	2009–2017	2219	8500
DHI	2009–2017	2057	16,900
EOG	2009–2017	2174	17,600
ES	2009–2017	1230	7200
F	2009–2017	1937	41,500
FB	2013–2017	1360	94,000
FIS	2009–2017	2196	8300
FISV	2009–2017	2193	6200
FLIR	2009–2017	2106	5400
GT	2009–2017	1855	12,900
HCBK	2009–2015	1599	8700
HES	2009–2017	2148	17,300
HRS	2009–2017	2183	5400
KHC	2015–2017	608	21,600
KIM	2009–2017	1691	10,700
KMX	2009–2017	2090	10,100
LH	2009–2017	2188	5700
LYB	2010–2017	1698	17,100
MAS	2009–2017	2087	13,800
MAT	2009–2017	2158	14,900
MCK	2009–2017	2146	10,300
MWV	2009–2015	1526	5200
NAVI	2011, 2014–2017	797	13,200
NE	2009–2017	2025	19,700
NFLX	2009–2017	2101	31,200
NLSN	2011–2017	1605	8900
PYPL	2015–2017	610	43,500
QCOM	2009–2017	2239	40,000
RHI	2009–2017	2098	6300
SRE	2009–2017	2194	7500
SYY	2009–2017	2209	13,200
TRV	2009–2017	2197	13,000
TSO	2009–2017	1986	15,300
TXN	2009–2017	2193	24,700
WRK	2015–2017	543	11,000
WU	2009–2017	2187	14,400
YHOO	2009–2017	2086	38,100
Mean		1905	18,400
SD		451	19,100
5%		694	5300
Median		2103	13,100
95%		2226	42,600

are quite small, indicating a strong stability of their explanatory power. On the other hand, the Roll model only achieves around 80% of explained variance and can be reasonably ruled out. We note that the proportion found for the Roll model is in line with the empirical results of Li, Xie, and Zheng (2016, Tables 4–7) and our own work Clinet and Potiron (2019, Table 6). We notice that the second simplest model, the signed spread model, already features an estimated average proportion of 100.3%. Although not reported, it turns out that, all the other submodels with one parameter feature estimated proportions of variance explained below 85%, making the signed spread incontestably the most explicative variable.

These findings have two consequences. First, the signed spread variable stands out with a real explanatory power very close to 100%. Second, the other three bigger models only allow for (at most) very small improvements in terms of goodness of fit (100.29%, 100.00%, 100.46%). We now give additional results for the signed spread model. In Table 10, we have reported details about confidence intervals for the real proportion of variance explained π using two distinct methods. We recall that the first method is an application of Theorem 3.1 from Clinet and Potiron (2019) along with the delta method, assuming no residual noise. The second method consists in the same protocol except that we now assume a small residual noise variance of order $1/N$. We can see that both average confidence intervals intercept 1 and are quite close to each other. We have also reported the proportion of days where 1 lies within their boundaries. We found that more than half of the time the estimated level of variance explained for the signed spread model is consistent with a perfect fit $\pi = 1$. Finally, unsurprisingly, a model selection scheme based on $\hat{\pi}$ largely advantages more complex models such as $M3$ (26.3%) and $M4$ (56.7%), whereas the signed spread model ($M2$) was selected only 8% of the time as documented in Table 11. This may be due to the fact that $\hat{\pi}$ tends to be mechanically higher for models with more parameters than for submodels (although, unlike the R^2 of a linear regression, it does not necessarily increase when adding new parameters).

We now turn to the BIC model selection as introduced in Section 3.1. In Table 11, we can see that the larger model $M4$ is selected most often (54.25%), followed by the global model $M5$ (45.14%). The signed spread models $M2$ and $M3$ are selected less than 1% of the days, while the Roll model is systematically ruled out by the procedure. In practice, such findings indicate that there is some nonnegligible residual information in models $M4$ and $M5$ that the signed spread model fails to capture. In particular, the signed trade inter-arrival time variable unequivocally boosts both $M4$ and $M5$ for the selection process. This finding is largely corroborated in Table 12, where we have reported the t-statistics of each parameter in the global model ($M5$). Indeed, the parameter related to the signed trade inter-arrival is significantly different from zero 86.4% of the days. This is less clear for the other two variables that appear in $M5$. While the signed quoted depth is significant 28% of the time, the signed volume is significant less than 1% of the days

residual noise is very close or equal to 0. Such extension of the parameter space is necessary to get a consistent likelihood optimization procedure. We refer the reader to the discussion before Theorem 3.1, p. 9 in Clinet and Potiron (2019).

Table 9. Daily estimates of volatility, noise-to-signal ratio, parameter in the signed spread model ($M2$), and parameter in the Roll model ($M1$) obtained using the second maximum likelihood based approach

	Mean	SD	5%	Median	95%
Volatility $\sigma_{i,t}$ scaled by mid price ($\times 10^2$)	0.310	0.176	0.121	0.260	0.685
Noise-to-signal ratio $\xi^2_{i,t}$ ($\times 10^3$)	0.397	0.596	0.052	0.255	1.193
Signed spread parameter $\theta^{(2)}_{i,t}$	0.861	0.073	0.726	0.869	0.961
Roll parameter $\theta^{(1)}_{i,t}$ over daily half-spread	0.816	0.084	0.669	0.821	0.943

Table 10. Confidence intervals for the proportion of variance explained in the signed spread model

$\hat{\pi}$	1.0030
$[\hat{\pi}_{\text{low}}, \hat{\pi}_{\text{up}}]$ (Method 1)	[0.9972, 1.0087]
$[\hat{\pi}_{\text{low}}, \hat{\pi}_{\text{up}}]$ (Method 2)	[0.9970, 1.0089]
Prop. "$1 \in [\hat{\pi}_{\text{low}}, \hat{\pi}_{\text{up}}]$" (Method 1)	52.7%
Prop. "$1 \in [\hat{\pi}_{\text{low}}, \hat{\pi}_{\text{up}}]$" (Method 2)	53.5%

NOTES: This table shows the average confidence intervals for the proportion of variance explained in the signed spread model across stocks and time, and computed according to two distinct methods. Method 1 corresponds to a consistent estimate assuming no residual noise. Method 2 assumes a small residual noise variance proportional to N^{-1}. "Prop. $1 \in [\hat{\pi}_{\text{low}}, \hat{\pi}_{\text{up}}]$" (Method i) stands for the proportion of days where 1 belongs to the corresponding confidence interval.

so that it can probably be ruled out with very few changes in the regression results. Finally, in terms of interpretability, the larger models $M4$ and $M5$ perform poorly as they suffer from an important instability of their parameters across days and stocks (For instance, the five parameters are positive between 25% and 62% of the time). This is due to the high-level of collinearity between variables.

Two conclusions can be drawn from the above investigation. First, according to BIC and the market microstructure noise variance explained, larger models such as $M4$ and $M5$ are undoubtedly closer to the truth than are submodels $M1$, $M2$, and $M3$. In particular, for statistical applications such as volatility estimation as described in Section 3.2, we advocate the practitioner to select first a model by BIC, and then keep the estimated volatility from this specific model. Second, in contrast with BIC and pure goodness-of-fit criteria, it is also clear from the above results that the model only incorporating the signed spread ($M2$) captures already a proportion of MMN that is estimated very close to 100% (Table 13). As we will see later, it also features a reasonable stability across time and stocks which makes it a good candidate to work with on empirical data whereas larger models feature unstable parameters. Although we wanted to use the BIC based exhibited models for our empirical analysis, the subsequent analysis was not working very well. This is a clear limitation and drawback of the BIC method. Therefore, we focus on a thorough analysis of the signed spread model ($M2$) for our empirical analysis in the remainder of this paper.

5.3. Interpreting the Effective-to-Quoted-Spread Parameter in the Signed Spread Model in Terms of Efficient and Vector of Raw Prices Relation

From now on, we reduce our study to the case of the signed spread model where

$$\phi(Q_{t_i}, \theta_0) = \frac{1}{2} S_{t_i} I_{t_i} \theta_0^{(2)}.$$

For ease of notation, we hereafter denote the single parameter of the model $\theta_0 \in \mathbb{R}$, also called effective-to-quoted-spread parameter. While not reported, the daily estimates $\theta_{i,t}$ range from 0.5 to 1.01 where the threshold value 1 is crossed over only for a few days over 95,255 points. It is thus reasonable to

Table 11. $\hat{\pi}$-based and BIC-based model selection for the five models across days and stocks

	$M1$	$M2$	$M3$	$M4$	$M5$
Mean $\hat{\pi}$	0.7976	1.0030	1.0029	1.0000	1.0046
SD $\hat{\pi}$	0.126	0.008	0.008	0.0003	0.027
Prop. best model ($\hat{\pi}$)	0.00%	8.00%	26.90%	56.70%	8.30%
Prop. best model (BIC)	0.00%	0.29%	0.31%	54.25%	45.14%

NOTES: This table reports for each model the mean proportion of variance explained and its standard deviation, along with the proportions of days where it was selected by the goodness of fit measure $\hat{\pi}$ and BIC.

Table 12. t-statistics for each parameter of the global model across days and stocks

	$\theta_0^{(1)}$	$\theta_0^{(2)}$	$\theta_0^{(3)}$	$\theta_0^{(4)}$	$\theta_0^{(5)}$
Mean $\lvert t\text{-stat}\rvert$	31.44	207.5	0.069	7.928	2.471
Prop. $\lvert t\text{-stat}\rvert > 1.96$	96.9%	100%	0.026%	86.4%	28.0%

NOTES: This table shows the mean value of each parameter of $M5$, the average absolute value of the t-statistic for each parameter of the global model, along with the proportion of days where the absolute value was above the threshold 1.96.

Table 13. Signed spread model in 2009–2017

Ticker	$\hat{\pi}_{i,t}$
AAPL	1.002
ACE	1.020
ALTR	1.003
ALXN	1.028
AMAT	0.997
ARG	1.035
BCR	1.027
BMY	0.997
CELG	1.016
CHRW	1.03
CL	1.014
CMI	1.015
CMS	0.990
CPB	1.004
DHI	0.994
EOG	1.024
ES	1.008
F	1.000
FB	0.998
FIS	1.008
FISV	1.022
FLIR	1.014
GT	0.999
HCBK	0.999
HES	1.027
HRS	1.026
KHC	1.003
KIM	0.987
KMX	1.021
LH	1.024
LYB	1.017
MAS	0.995
MAT	0.994
MCK	1.026
MWV	1.017
NAVI	0.997
NE	1.012
NFLX	1.007
NLSN	1.013
PYPL	0.998
QCOM	0.997
RHI	1.022
SRE	1.019
SYY	0.995
TRV	1.008
TSO	1.007
TXN	0.993
WRK	1.010
WU	0.997
YHOO	0.996

say that the spread parameter θ_0 virtually satisfies $0 \leq \theta_0 \leq 1$, that is, the first regime discussed in Section 2. This implies that the efficient price stays systematically between the mid price and the transaction price. Moreover, for almost every day there is additional information (in the sense that the efficient price is different from the mid price, i.e., $\theta_0 \neq 1$) in trade direction.

5.4. The Effective-to-Quoted-Spread Parameter Variability

Prior to introducing the financial characteristics and to relate the signed spread model parameter to those characteristics, we document in this section about the reasonable stability of θ_0 (compared to the quoted spread size) whether we look at the variation within a day, daily or across stocks. This corroborates the predominance of the bid-ask spread among the possible sources of market microstructure noise.

We recall a key relation between the market microstructure noise variance a^2_{MMN}, the parameter and the quoted spread variance varIS discussed in Section 2 as

$$a^2_{\text{MMN}} \approx \frac{\theta_0^2 \text{var}IS}{4}, \tag{27}$$

where the use of "approximation" instead of "equal" in the relation is due to the fact that there may be some (very small) residual noise.

Table 14 reports the parameter estimates and the daily averaged quoted spread across stocks. The parameter mean value ranges from 0.767 for MWV to 0.916 for AAPL, with a mean across stocks equal to 0.862 and a standard deviation equal to 0.042. The quoted spread (when multiplied by 1000) ranges from 0.150 for AAPL to 0.801 for HCBK, with an average equal to 0.394 and a standard deviation of 0.141. Obviously, the variation (when normalized by the mean) of the parameter is only about one tenth that of the daily averaged quoted spread.

We next examine the intraday stability of θ_0 (aggregated over the fifty stocks and over the full period 2009–2017). In Figure 2, we show the average intraday variations of θ_0 (blue, solid), the quoted spread (green, dashed), and the market microstructure noise standard deviation (red, dot-dashed), where we have split the trading day in eight periods of equal length. We have normalized the three variables so that they daily average to 1. We can see that despite the fact that θ_0 exhibits a U-shape over the trading period, it is much smaller than that of the bid-ask spread. Indeed, we find a daily average parameter high-low variation of 9%, whereas the average market microstructure noise standard deviation's deviation is as high as 74%. Furthermore, we can see that the market microstructure noise standard deviation exhibits a half U-shape pattern rather than a full one. This is in line with, for example, the recent findings of Chen and Mykland (2017, Figure 8) where the authors report lower noise levels in the afternoon and around noon than in the morning. The same pattern is observed for the quoted spread, and this is also consistent with the relation (27). Consequently, the aforementioned findings suggest that over a day, the variation of the market microstructure noise is mainly driven by the bid-ask spread.

Now we turn to the daily parameter variation along with the evolution of other quantities related to the microstructure noise over the period 2009–2017. The evolution of the monthly averaged parameter is shown in the top left panel of Figure 3.

Table 14. The effective-to-quoted-spread parameter and daily averaged quoted spread in 2009–2017

Ticker	Signed spread parameter $\theta_{i,t}^{(2)}$	Quoted bid-ask spread ($\times 10^3$)
AAPL	0.916	0.150
ACE	0.866	0.258
ALTR	0.770	0.349
ALXN	0.896	0.603
AMAT	0.901	0.611
ARG	0.864	0.391
BCR	0.887	0.407
BMY	0.874	0.288
CELG	0.904	0.335
CHRW	0.872	0.299
CL	0.856	0.192
CMI	0.910	0.384
CMS	0.826	0.435
CPB	0.836	0.296
DHI	0.816	0.563
EOG	0.897	0.340
ES	0.875	0.265
F	0.895	0.735
FB	0.870	0.210
FIS	0.832	0.324
FISV	0.864	0.311
FLIR	0.797	0.427
GT	0.833	0.568
HCBK	0.914	0.801
HES	0.866	0.305
HRS	0.871	0.382
KHC	0.923	0.277
KIM	0.839	0.529
KMX	0.825	0.380
LH	0.887	0.376
LYB	0.877	0.339
MAS	0.832	0.557
MAT	0.813	0.405
MCK	0.888	0.339
MWV	0.767	0.406
NAVI	0.863	0.671
NE	0.784	0.383
NFLX	0.936	0.528
NLSN	0.844	0.378
PYPL	0.867	0.277
QCOM	0.835	0.196
RHI	0.809	0.392
SRE	0.864	0.298
SYY	0.857	0.309
TRV	0.859	0.222
TSO	0.852	0.501
TXN	0.830	0.291
WRK	0.915	0.425
WU	0.896	0.567
YHOO	0.869	0.442
Mean	0.862	0.394
SD	0.042	0.141
5%	0.790	0.202
Median	0.865	0.379
95%	0.920	0.644

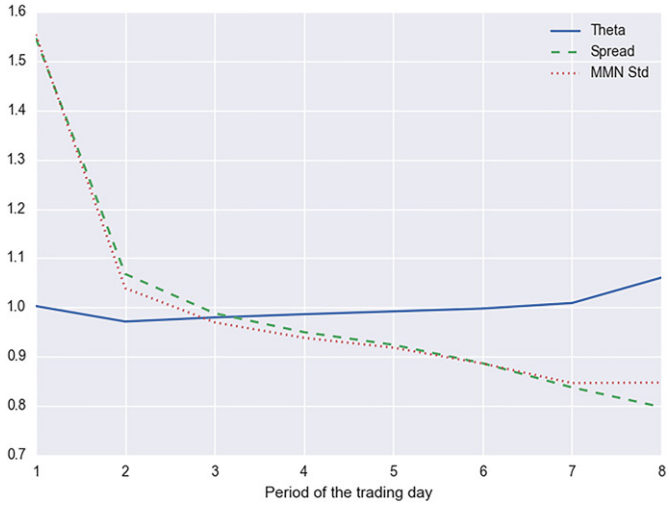

Figure 2. Intraday variations of the effective-to-quoted-spread parameter and the quoted spread in normalized units.

After a first period of decreasing trend between 2009 and the stock markets fall in August 2011 where it reached a minimum value of 0.77, we can see a clear positive trend since 2011 until the end of 2017, where the parameter attained 0.91. Coincidentally, the market noise variance (top right panel) decreased from 2009 to 2014 where it seems to have stabilized on average. While not shown, the evolution of the quoted spread mimicks that of the noise variance. Finally, the noise-to-signal ratio (middle left panel) shows a general tendency to increase over those last 9 years, suggesting that the volatility levels of the market actually decreased faster than the noise variance since the 2008 financial crisis.

We have seen that in terms of intraday, daily or variation across stocks, the parameter is far more stable than the quoted spread. Consequently, we determine the signed quoted spread, that is, that we recall to be the association of the effective spread along with the trade direction effects, as the main source of market microstructure noise. In addition, the high frequency financial characteristics explaining the parameter variability will be classified as supplementary sources in Section 5.6.

Furthermore, in light of the above findings, trade direction seems to have become less informative over time as the efficient-to-mid price deviation has decreased globally, now counting for less than 10% of the quoted bid-ask half-spread. One possible interpretation is that market makers tend to track the efficient price (through their order submissions) more and more precisely, making de facto the mid price stick to the efficient price. In addition, the decreasing of the noise variance over time parallel to that of the spread suggests that the better tracking of the efficient price along with a reduction of transaction costs.

We finally examine the serial autocorrelation of the parameter's daily returns. Figure 4 contains the first 30 lags of the parameter's autocorrelation function. Essentially, the returns feature a nonnegligible negative first lag autocorrelation, whereas further lags are at most barely significant. In particular, there is no evidence of long range dependence in the increments of the effective-to-quoted-spread parameter. The negative first

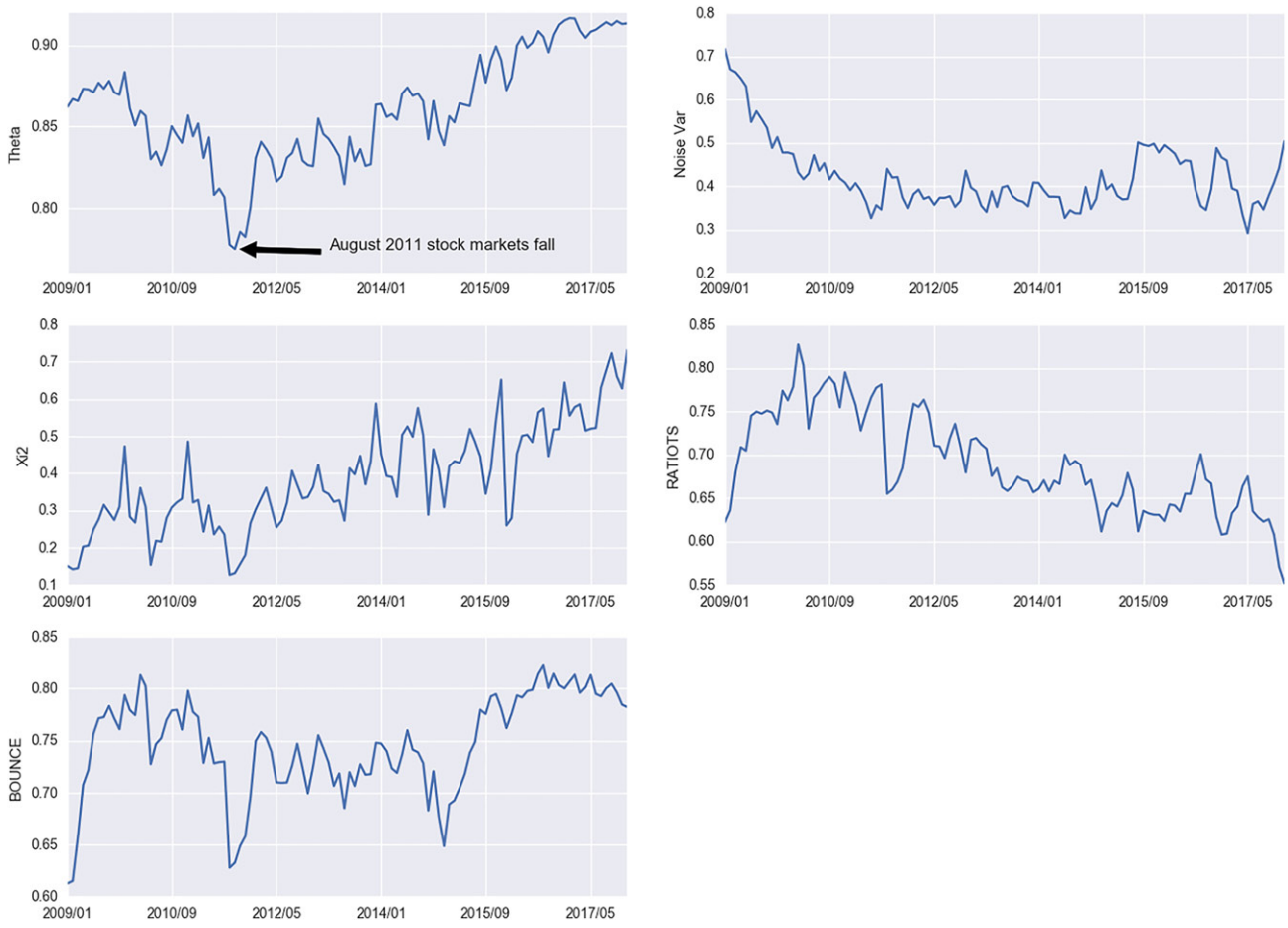

Figure 3. Monthly averages of five quantities over the period 2009–2017. The top left panel documents the evolution of the effective-to-quoted-spread parameter. The top right panel shows the market microstructure noise variance. The middle left panel reports the noise-to-signal ratio whereas the middle right panel shows RATIOTS. Finally, the bottom left panel corresponds to BOUNCE.

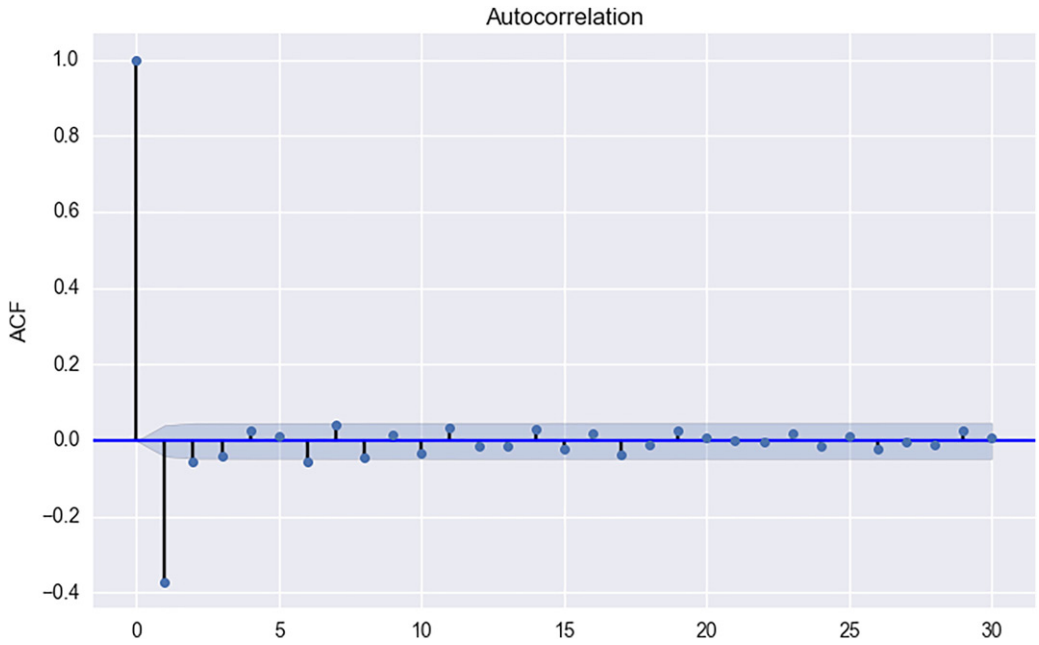

Figure 4. Autocorrelation function of the parameter's daily returns.

Table 15. Summary table of the financial characteristics

Name	Symbol	Mean	SD	5%	Median	95%
Tick/spread ratio	RATIOTS	0.690	0.290	0.155	0.778	0.997
Tick/price ratio ($\times 10^4$)	RATIOTP	2.602	1.891	0.644	1.985	6.835
Spread/price rRatio ($\times 10^3$)	RATIOSP	0.374	0.168	0.169	0.335	0.735
Bid-ask bounce proportion	BOUNCE	0.749	0.163	0.465	0.756	0.973
Log-number of trades	log$N$	9.373	0.857	8.083	9.307	10.934
Log-number of volumes	log$V$	14.724	1.085	13.056	14.652	16.759
Volatility ($\times 10^2$)	σ	0.310	0.176	0.121	0.260	0.685
Trade direction autocorrelation	TDCORR	0.398	0.112	0.242	0.379	0.596
Absolute order flow imbalance ($\times 10^2$)	AOFI	5.741	4.783	0.422	4.580	15.037
Order book asymmetry	OBA	0.374	0.067	0.255	0.382	0.473
Average trade size	ATS	0.804	0.534	0.131	0.700	1.658

lag also suggests a slight mean-reversion effect in θ_0 (and thus in the noise level of the market, by (27)) in time.

5.5. High Frequency Financial Characteristics of the Underlying Stocks

We look at a collection of financial characteristics that are typically identified as potential sources of market frictions in the literature. These characteristics are commonly interpreted as financial measures of liquidity. A summary of these financial characteristics can be found in Table 15.

In addition to the volatility $\sigma_{i,t}$ introduced in Section 5.1, we let $\text{RATIOTS}_{i,t}$ be the daily average ratio of the price tick size over the spread. We see this measure, which by definition lies between 0 and 1, as primarily accounting for the impact of discreteness on the market microstructure noise. Indeed, a stock with $\text{RATIOTS}_{i,t} \approx 1$ (also called large-tick asset), which in particular can be considered as liquid since transaction costs are reduced to their minimum, suffers directly from discreteness effects. Correspondingly, a stock featuring $\text{RATIOTS}_{i,t} \approx 0$ can be considered as relatively free from those effects. Similarly, $\text{RATIOTP}_{i,t}$ (respectively, $\text{RATIOSP}_{i,t}$) stands for the ratio of one tick over the average mid price (respectively, the average ratio of the quoted spread over the mid price). While those three measures are related, there are conceptually very distinct and appeal to distinct audiences. On the one hand, $\text{RATIOTS}_{i,t}$ corresponds to discreteness effects of special interest to the market maker, as its objective is usually to "make the spread." On the other hand, a fundamental investor may be interested in $\text{RATIOSP}_{i,t}$ (or even $\text{RATIOTP}_{i,t}$, there are both very close to each other in practice) to value the impact of discreteness effects in her balance sheet.

We define the proportion of trades that are exclusively due to the bid-ask bounce mechanism, which is defined in the present work as transactions which do not induce a shift in the mid price, as $\text{BOUNCE}_{i,t}$. We also let $\log N_{i,t}$ denote the daily logarithm of the number of trades, and similarly $\log V_{i,t}$ be the daily logarithm of the traded volume. $\text{TDCORR}_{i,t}$ is the daily first lag autocorrelation of the trade direction. Writing $\text{NBID}_{i,t}$ (respectively, $\text{NASK}_{i,t}$) the number of market orders executed at the best bid price (respectively, best ask price), we define the absolute order-flow imbalance $\text{AOFI}_{i,t} = |\text{NASK}_{i,t} - \text{NBID}_{i,t}|/(\text{NASK}_{i,t} + \text{NBID}_{i,t}) \in [0,1]$, which measures the (average) asymmetry in the trade order flow.

Finally, we consider two additional measures, respectively, related to the shape and the depth of the first level of the order book. We define the order book asymmetry $\text{OBA}_{i,t}$ as the daily average of the quantity $|\text{VASK} - \text{VBID}|/(\text{VASK} + \text{VBID}) \in [0,1]$, where VASK (respectively, VBID) stands for the pending volume at the best ask limit (respectively, best bid limit). $\text{OBA}_{i,t}$ is related to the micro-price (see, e.g., Gatheral and Oomen 2010, sec. 4.3, for corresponding definition and interpretations), which is a popular proxy for the efficient price. An $\text{OBA}_{i,t}$ close to 1 indicates that one limit of the order book is nearly empty. It is well-known that the sign of short term mid price increments is negatively correlated with the signed $\text{OBA}_{i,t}$. In fact, this corresponds to the signaling and barrier effects described in Huang and Stoll (1994). The second volume-based measure is defined as the daily average trade size $\text{ATS}_{i,t}$, scaled by the pending volume at the best limit where the trade was executed. $\text{ATS}_{i,t}$ is thus the daily average of $\text{VTRADE}/\text{VBEST} \in [0,1]$ where, for each trade, VTRADE is the size of the order, and VBEST corresponds to VASK if the order was buyer-initiated and VBID otherwise.

5.6. The Effective-to-Quoted-Spread Parameter and Financial Characteristics: Looking for the Residual Sources of Noise

We now proceed to relate our high frequency estimates of θ_0 to the financial characteristics discussed in Section 5.5. By regressing θ_0 on other variables, we aim at looking for residual sources of noise (in the sense that they are not intraday, but rather explains variability in days and stocks). Indeed, recalling the key relation (27), we see that a measure which has an impact on θ_0 directly affects the general variance level of the related MMN. It can therefore be seen as a source of MMN since it contributes to determine the size of the noise on a given day for a given stock. (However, note that in view of the intraday stability of θ_0, such contribution does not seem to vary at high-frequency time scales.) It is also worth noting that, alternatively, the effect of the aforementioned measures could be studied by considering more general models than the additive approach of (1). For instance, Li, Zhang, and Li (2018) and Tang and Zhang

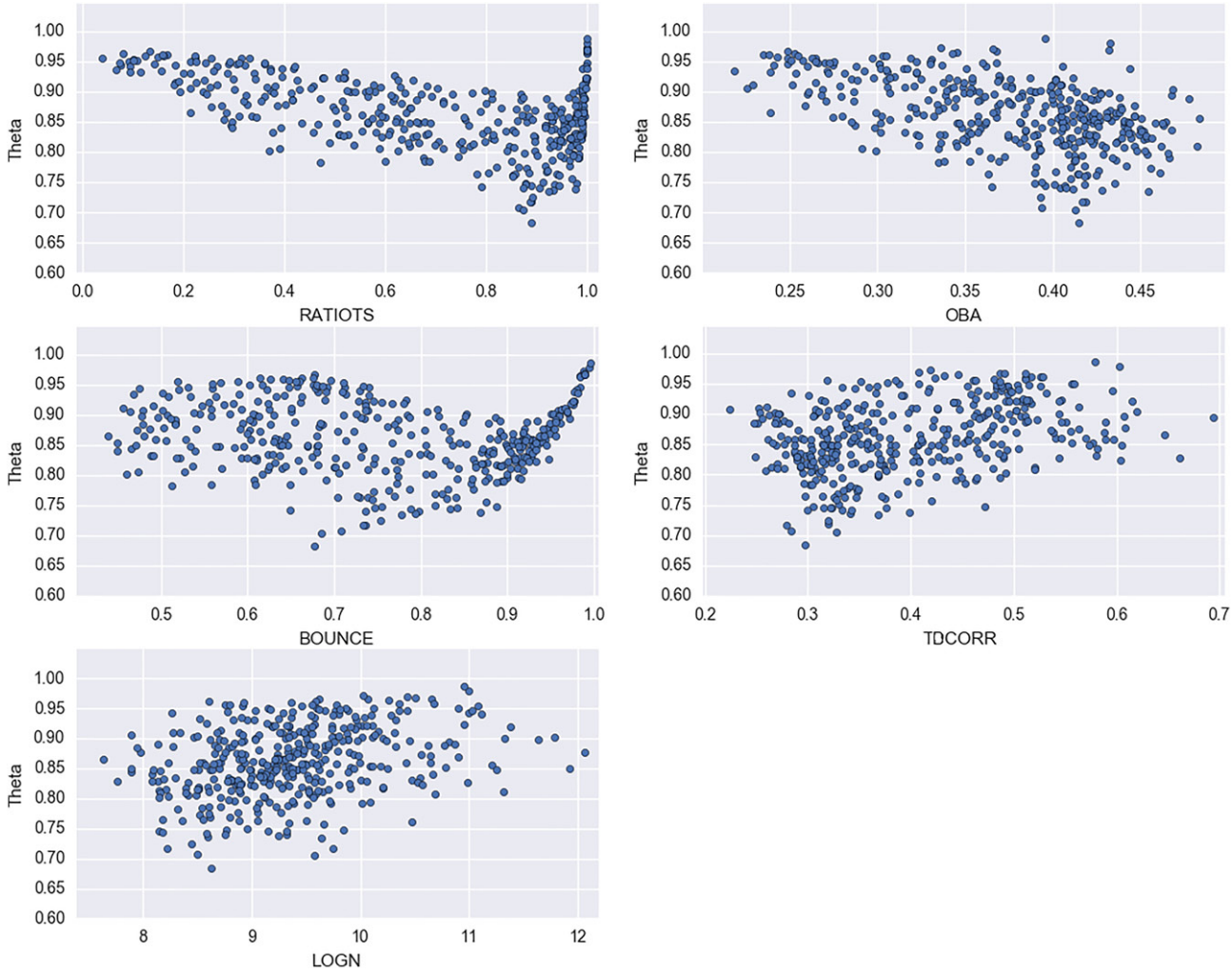

Figure 5. Scatterplots of the effective-to-quoted-spread parameter versus five financial characteristics.

(2018) recently proposed new volatility estimation methods for models which explicitly feature price discreteness in addition to the additive MMN. We consider the set of measures

$$M := \{\text{RATIOTS}, \text{RATIOTP}, \text{RATIOSP}, \text{BOUNCE}, \log N, \log V, \sigma, \text{TDCORR}, \text{AOFI}, \text{OBA}, \text{ATS}\}.$$

Moreover, for computational tractability, we conduct our regression analysis on yearly averaged-values of the effective-to-quoted spread parameter and of the aforementioned financial characteristics.[9] Accordingly, for each $x \in M$, each year $y = 2009, \ldots, 2017$ and each stock i, we look at the yearly averages $\theta_{i,y}$ and $x_{i,y}$ of our daily estimates. Figure 5 shows scatterplots of $\theta_{i,y}$ versus five measures taken from M. We can see that a linear relation seems reasonable for three measures, but $\theta_{i,y}$ seems to exhibit a strong nonlinear relationship with $\text{RATIOTS}_{i,y}$ and $\text{BOUNCE}_{i,y}$. In both cases, the V-shape of the scatterplot suggests a structural break in θ_0 when RATIOTS (respectively, BOUNCE) approaches 0.9, with a negative pre-breakpoint slope, and a positive post-breakpoint slope. Accordingly, we begin our investigation by running the following simple regressions

$$\theta_{i,y} = \beta_0 + \beta_1 x_{i,y} + \epsilon_{i,y} \tag{28}$$

for $x \in M_0 := \{\text{RATIOTP}, \text{RATIOSP}, \log N, \log V, \sigma, \text{TDCORR}, \text{AOFI}, \text{OBA}, \text{ATS}\}$. Moreover, we explicitly incorporate the presence of a structural break in the regressions of θ_0 on BOUNCE and RATIOTS by running the regression

$$\theta_{i,y} = \beta_0 + \beta_1 (x_{i,y} - x^*)\mathbf{1}_{\{x_{i,y} \le x^*\}} + \beta_2 (x_{i,y} - x^*)\mathbf{1}_{\{x_{i,y} > x^*\}} + \epsilon_{i,y} \tag{29}$$

for $x \in M_1 := \{\text{RATIOTS}, \text{BOUNCE}\}$. The estimation of (β_0, β_1) in (28) and the quadruplet $(\beta_0, \beta_1, \beta_2, x^*)$ in (29) is performed by an ordinary least square method (OLS).

The results of the individual OLS regressions can be found in Table 16. We have sorted the individual regressions according to their adjusted regression R-squared. Overall, the tick-over-spread ratio explains most variation in θ_0 (52.1%), followed by the order book asymmetry (26.7%), and the bid-ask bounce proportion (22.9%). We find a negative correlation between RATIOTS and θ_0 and BOUNCE and θ_0 before the breakpoint. The regression slope then turns sharply positive when those variables approach 1. It is difficult to interpret the relationship

[9] Although not reported in the article, we have also considered shorter periods such as 6 or 3 months. Apart from slightly lower R^2, the significance, the rankings of the financial measures and their estimated coefficients in the regressions reported in Tables 16 and 17 are unchanged.

Table 16. Simple linear regressions of the effective-to-quoted-spread parameter on financial characteristics

Variable	Coeff	Breakpoint	Adj R^2	t-stat
RATIOTS	$-0.15, 6.90$	0.98	52.1%	$-19.20, 15.07$
OBA	-0.51		26.7%	-12.10
BOUNCE	$-0.16, 1.62$	0.91	22.9%	$-8.16, 10.32$
TDCORR	0.23		13.4%	7.93
DATE ($\times 10^{-5}$)	2.22		12.2%	7.53
log$N$ ($\times 10^{-1}$)	0.25		10.1%	6.76
RATIOSP ($\times 10^{-1}$)	0.77		4.7%	4.53
σ ($\times 10^{-1}$)	-0.57		2.2%	-3.15
RATIOTP	-36.00		1.3%	-2.48
log$V$ ($\times 10^{-2}$)	0.70		1.0%	2.11
ATS ($\times 10^{-1}$)	-0.14		0.7%	-1.86^{-}
AOFI	0.11		-0.2%	0.54^{-}

NOTES: This table documents the simple linear regressions described in (28) and (29). The column "Coeff" shows the slope coefficients (prior and post breakpoint for RATIOTS and BOUNCE). The column "Breakpoint" contains the values where the structural break occurs. "Adj R^2" corresponds to the adjusted R^2 of the regressions and finally the last column documents the t-statistics of the regressions, where the presence of $^{-}$ on the right indicates that the p-value was above the 5% level.

between θ_0 and both variables individually, especially because they are highly correlated. OBA is negatively correlated with θ_0. This indicates that trades from asymmetric order books tend to feature lower θ_0, or, in other words, more additional informational content. We also find that both variables TDCORR and log$N$ explain, respectively, 13.4% and 10.1% of the variability in θ_0 with a positive correlation. In addition to the other variables, we have also run the regression on time (row DATE in Table 16) to confirm our findings of Section 5.4. Unsurprisingly, we find a positive linear trend in θ_0 over time. Finally, the other variables explain less than 5% of the variation in θ_0.

Due to high correlations across the financial characteristics, the individual regression results are hard to interpret. Accordingly, we now proceed to disentangle the effects of the measures on θ_0 by running the following global multiple linear regression

$$\theta_{i,y} = \beta_0 + \sum_{j|x_j \in M_0} \beta_j x_{j,i,y} + \sum_{j|x_j \in M_1} \Big\{ \beta_{1,j}(x_{j,i,y} - x_j^*)\mathbf{1}_{\{x_{j,i,y} \le x_j^*\}} + \beta_{2,j}(x_{j,i,y} - x_j^*)\mathbf{1}_{\{x_{j,i,y} > x_j^*\}} \Big\} + \epsilon_{i,y}.$$

The results of the above regression are presented in the left panel of Table 17. Overall, we find that the financial characteristics from the literature account for 90.4% (in terms of adjusted R^2) of the variability in θ_0. The two most significant variables are by far RATIOTS and BOUNCE. We note that the order book asymmetry variable is no longer relevant in the multiple regression context, most likely due to its high level of correlation with the two aforementioned variables. Most importantly, we can see that RATIOTS and BOUNCE now present slopes of constant sign on both sides of their respective breakpoints. Other things equal, the tick-over-spread ratio, accounting for the level of discreteness of the stock, affects negatively θ_0, that is, increasing the tick size in spread units tends to shift the efficient price away from the mid price. The negativity of the relation is puzzling as we expect that more liquid stocks are less noisy (see Aït-Sahalia and Yu 2009), but can be interpreted as follows. When the tick-over-spread ratio is smaller and other things are held constant, the market makers placed on both sides of the spread can track the shocks in the efficient price more meticulously since the tick grid is denser, thus reducing the deviation between the efficient price and the mid price. In conjunction with this relation, we find a positive relation between θ_0 and the proportion of bid-ask bounce. The obvious interpretation is that, all other things being equal, the trade direction is less informative in stocks with a bigger proportion of bid-ask bounce. Now, given the high level of correlation between both variables ($\rho \approx 0.87$), the two-regime behavior of θ_0 as a function of discreteness (see Figure 5 along with the individual regression on RATIOTS) can be easily explained as follows. As long as RATIOTS is not too close to unity, increasing the discreteness level mainly negatively affects the way market makers can track the efficient price. However, when the tick-over-spread ratio becomes too close to unity (i.e., hits the breakpoint 0.98 found in Table 16), it becomes virtually impossible to insert new orders inside the spread, which implies a high level of bid-ask bounce and pulls back the efficient-to-mid price deviation to 0.[10]

We next investigate whether it is possible or not to reach a similar level of goodness of fit (here adjusted R^2) with a subset of the regressors in the above regression. It turns out that, as reported in the right panel of Table 17, the combined effects of RATIOTS and BOUNCE already gives an adjusted R^2 of 86.5%, which is quite close to 90.4%.

Finally, we look at the best submodel selected by Akaike's information criterion (AIC). We find that {RATIOTS, BOUNCE, TDCORR, RATIOTP, log$N$, σ} is the most informative model among all possible combinations, suggesting that {TDCORR, RATIOTP, log$N$, σ} should contain additional minor sources of market microstructure noise. The regression results are very similar to the global ones for each variable and have been omitted for the sake of brevity.

[10]Note that in the limit of a pure bid-ask bounce, Equation (9) in the dynamic case immediately yields $\theta_0 = 1$.

Table 17. Multiple linear regressions of the effective-to-quoted-spread parameter on financial characteristics

	All characteristics			Main characteristics		
Variable	Coeff	t-stat	Breakpoint	Coeff	t-stat	Breakpoint
RATIOTS	$-0.31, -0.98$	$-15.30, -23.73$	0.76	$-0.27, -0.93$	$-29.54, -28.98$	0.76
BOUNCE	$0.42, 1.30$	$12.37, 17.71$	0.79	$0.39, 1.40$	$19.85, 34.21$	0.79
TDCORR	-0.15	-5.86				
σ	-0.10	-5.08				
RATIOTP	113.38	4.50				
log$N$	0.01	1.99				
ATS	-0.01	-1.97				
RATIOSP	0.035	1.45^{-}				
AOFI	-0.16	-1.39^{-}				
OBA	0.05	1.22^{-}				
log$V$ ($\times 10^{-1}$)	0.03	0.54^{-}				
Intercept	0.76	18.41		0.87	297.10	
Adj R^2	90.4%			86.5%		

NOTES: This table documents two multiple linear regressions of θ_0. The left panel corresponds to the regression of θ_0 on all the financial characteristics. The right panel stands for the regression on the submodel consisting of the set {RATIOTS, BOUNCE}. The columns "Coeff" show the slope coefficients (pre and post breakpoint for RATIOTS and BOUNCE). The columns "Breakpoint" contain the values where the structural break occurs. The columns "t-stat" document the t-statistics of the regressions, where the presence of $^{-}$ on the right indicates that the p-value was above the 5% level. The row "Adj R^2" corresponds to the adjusted R^2 of the regressions.

Table 18. Market factor regression for the effective-to-quoted parameter and market microstructure noise standard deviation

	θ_0			a_{MMN}		
	β_i	$\lvert t\text{-stat}\rvert$	Adj R^2	β_i	$\lvert t\text{-stat}\rvert$	Adj R^2
Mean	0.79	10.76	7.54%	0.57	7.14	3.32%
SD	0.53	6.24	6.47%	0.31	2.96	2.68%
Min	0.07	0.52	-0.10%	-0.05	0.20	-0.05%
Median	0.70	9.59	5.62%	0.63	7.05	3.21%
Max	1.91	24.7	22.5%	1.15	13.2	8.11%
Positive	100%			94%		
Significant	100%			88%		

NOTES: In this table, we have reported the results of the regression (30). The left panel corresponds to the case θ_0, whereas the right panel shows the results for a_{MMN}. The row "Positive" gives the proportion of stocks with a positive β_i. The row "Significant" corresponds to the proportion of stocks for which the two-sided test of the null hypothesis $\beta_i = 0$ was rejected with 5% significance level.

Overall, we deduce that discreteness can reasonably be considered as the main source to explain variability in effective-to-quoted-spread parameter, and accordingly bid-ask bounce as the second source.

5.7. Market Factor in the Effective-to-Quoted-Spread Parameter

We conclude this empirical analysis by investigating whether the parameter and the market microstructure noise standard deviation contain a market factor or not. Co-movements in a_{MMN} have already been reported in Aït-Sahalia and Yu (2009). The authors show that over the period 1995–2005, even though regressing on a common market factor yields poor adjusted R^2 levels, a nonnegligible portion of stocks features a nonzero slope coefficient, suggesting that there is commonality in microstructure noise variation.

To assess the presence of a market wide factor in both variables, we look for each stock i at the regression of the daily relative change of $\theta_{i,t}$ (respectively, $a_{i,t,\text{MMN}}$) on the daily relative change of the (equally weighted) market-wide variable $\theta_{-i,t} = \frac{1}{49}\sum_{j\neq i}\theta_{j,t}$ (respectively, $a_{-i,t,\text{MMN}} = \frac{1}{49}\sum_{j\neq i} a_{j,t,\text{MMN}}$), excluding the stock i itself

$$\log\left(\frac{y_{i,t}}{y_{i,t-1}}\right) = \alpha_i + \beta_i \log\left(\frac{y_{-i,t}}{y_{-i,t-1}}\right) + \epsilon_{i,t}, \qquad (30)$$

where $y_{i,t} \in \{\theta_{i,t}, a_{i,t,\text{MMN}}\}$. By excluding the stock that appears on the left-hand side of the regression from the market index on the right-hand side, we stay away from getting artificially biased coefficients.

Results of both regressions are reported in Table 18. In most cases we find positive slopes, which are statistically significant in 100% of cases (respectively, 88%) for θ_0 (respectively, a_{MMN}). It indicates the presence of a market effect in the daily returns of both variables. Moreover, we can clearly see that such effect is stronger in θ_0 than it is for the market microstructure noise. This is confirmed by a higher level of adjusted regression R^2 for θ_0 (7.54% on average) than for the noise (3.32%). Since price discreteness and bid-ask bounce are the main constituents of θ_0 (Table 17), and θ_0 and the quoted

signed spread variance determine the market microstructure noise variance (by the relation (27)), the fact that θ_0 is more often significantly correlated to its market index counterpart than is a_{MMN} suggests that price discreteness and bid-ask bounce mechanisms are more subject to such a market effect than the quoted signed spread variance.

6. CONCLUDING REMARKS

In this paper, we investigate the consistency of a BIC to discriminate between several financial models of market microstructure noise. We also give a BIC-based volatility estimation procedure, although not considering the according limit theory. We identify the quoted spread as a simple model which explains a very large proportion of the market microstructure noise. We investigate the relation between the efficient price and the vector of raw prices in this model. We find that the efficient price is systematically between the mid price and the transaction price. We also document that the variability of the parameter is low compared to that of the signed spread. We explain the parameter variability with several financial characteristics, and accordingly identify discreteness as the first residual source of noise, and bid-ask bounce effects as the second residual source.

SUPPLEMENTARY MATERIALS

Supplemental materials consist of proofs related to the BIC (Section 3).

ACKNOWLEDGMENTS

We would like to thank Yingying Li, Xinghua Zheng, Torben Andersen, Frederic Abergel and the participants of Quantitative Finance seminar at CentraleSupélec for helpful discussions and advice.

FUNDING

The research of Yoann Potiron is supported by Japanese Society for the Promotion of Science Grant-in-Aid for Young Scientists (B) No. 60781119 and a special grant from Keio University. The research of Simon Clinet is supported by a special grant from Keio University. All financial data are provided by the Chair of Quantitative Finance of the Ecole Centrale Paris.

[Received October 2018. Revised January 2019.]

REFERENCES

Aït-Sahalia, Y., Mykland, P. A., and Zhang, L. (2005), "How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise," *Review of Financial Studies*, 18, 351–416. [22,25]

Aït-Sahalia, Y., and Xiu, D. (2016), "A Hausman Test for the Presence of Market Microstructure Noise in High Frequency Data" (to appear), *Journal of Econometrics*. [22,23,24]

Aït-Sahalia, Y., and Yu, J. (2009), "High Frequency Market Microstructure Noise Estimates and Liquidity Measures," *Annals of Applied Statistics*, 3, 422–457. [19,36,37]

Almgren, R., and Chriss, N. (2001), "Optimal Execution of Portfolio Transactions," *Journal of Risk*, 3, 5–40. [18]

Andersen, T. G., Cebiroglu, G., and Hautsch, N. (2017), "Volatility, Information Feedback and Market Microstructure Noise: A Tale of Two Regimes," Working Paper, available at SSRN 2921097. [19]

Andersen, T. G., Dobrev, D., and Schaumburg, E. (2012), "Jump-Robust Volatility Estimation Using Nearest Neighbor Truncation," *Journal of Econometrics*, 169, 75–93. [24]

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008), "Designing Realized Kernels to Measure the Ex Post Variation of Equity Prices in the Presence of Noise," *Econometrica*, 76, 1481–1536. [25]

Cao, C., Choe, H., and Hatheway, F. (1997), "Does the Specialist Matter? Differential Execution Costs and Intersecurity Subsidization on the New York Stock Exchange," *The Journal of Finance*, 52, 1615–1640. [20]

Chaker, S. (2017), "On High Frequency Estimation of the Frictionless Price: The Use of Observed Liquidity Variables," *Journal of Econometrics*, 201, 127–143. [18,20,22,23]

Chan, L. K., and Lakonishok, J. (1997), "Institutional Equity Trading Costs: NYSE Versus Nasdaq," *The Journal of Finance*, 52, 713–735. [20]

Chen, F., and Hall, P. (2013), "Inference for a Nonstationary Self-Exciting Point Process With an Application in Ultra-High Frequency Financial Data Modeling," *Journal of Applied Probability*, 50, 1006–1024. [24]

Chen, R. Y., and Mykland, P. A. (2017), "Model-Free Approaches to Discern Non-stationary Microstructure Noise and Time-Varying Liquidity in High-Frequency Data," *Journal of Econometrics*, 200, 79–103. [19,31]

Christensen, K., Hounyo, U., and Podolskij, M. (2018), "Is the Diurnal Pattern Sufficient to Explain Intraday Variation in Volatility? A Nonparametric Assessment," *Journal of Econometrics*, 205, 336–362. [19]

Clinet, S., and Potiron, Y. (2017), "Estimation for High-Frequency Data Under Parametric Market Microstructure Noise," Working Paper, available at arXiv no. 1712.01479. [19]

——— (2018a), "Efficient Asymptotic Variance Reduction When Estimating Volatility in High Frequency Data," *Journal of Econometrics*, 206, 103–142. [21,22,24]

——— (2018b), "Statistical Inference for the Doubly Stochastic Self-Exciting Process," *Bernoulli*, 24(4B):3469–3493. [24]

——— (2019), "Testing if the Market Microstructure Noise Is Fully Explained by the Informational Content of Some Variables From the Limit Order Book," *Journal of Econometrics*, 209, 289–377. [18,19,20,22,23,24,25,26,27,29]

Da, R., and Xiu, D. (2017), "When Moving-Average Models Meet High-Frequency Data: Uniform Inference on Volatility," Working Paper, available at Dacheng Xiu's website. [19,22]

Delattre, S., and Jacod, J. (1997), "A Central Limit Theorem for Normalized Functions of the Increments of a Diffusion Process, in the Presence of Round-Off Errors," *Bernoulli*, 3, 1–28. [20]

Diebold, F. X., and Strasser, G. (2013), "On the Correlation Structure of Microstructure Noise: A Financial Economic Approach," *The Review of Economic Studies*, 80, 1304–1337. [19]

Engle, R. F., and Russell, J. R. (1998), "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data," *Econometrica*, 66, 1127–1162. [24]

Gatheral, J., and Oomen, R. C. (2010), "Zero-Intelligence Realized Variance Estimation," *Finance and Stochastics*, 14, 249–283. [34]

Ghysels, E., Harvey, A. C., and Renault, E. (1996), "5 Stochastic Volatility," *Handbook of Statistics*, 14, 119–191. [21]

Glosten, L. R. (1987), "Components of the Bid-Ask Spread and the Statistical Properties of Transaction Prices," *The Journal of Finance*, 42, 1293–1307. [20]

Glosten, L. R., and Harris, L. E. (1988), "Estimating the Components of the Bid/Ask Spread," *Journal of Financial Economics*, 21, 123–142. [18,20]

Glosten, L. R., and Milgrom, P. R. (1985), "Bid, Ask and Transaction Prices in a Specialist Market With Heterogeneously Informed Traders," *Journal of Financial Economics*, 14, 71–100. [20]

Gottlieb, G., and Kalay, A. (1985), "Implications of the Discreteness of Observed Stock Prices," *The Journal of Finance*, 40, 135–153. [20]

Hansen, P. R., and Lunde, A. (2006), "Realized Variance and Market Microstructure Noise," *Journal of Business & Economic Statistics*, 24, 127–161. [18]

Harris, L. (1990a), "Estimation of Stock Price Variances And Serial Covariances From Discrete Observations," *Journal of Financial and Quantitative Analysis*, 25, 291–306. [20]

——— (1990b), "Statistical Properties of the Roll Serial Covariance Bid/Ask Spread Estimator," *The Journal of Finance*, 45, 579–590. [20]

Harris, L., and Gurel, E. (1986), "Price and Volume Effects Associated With Changes in the S&P 500 List: New Evidence for the Existence of Price Pressures," *The Journal of Finance*, 41, 815–829. [19]

Hasbrouck, J. (1993), "Assessing the Quality of a Security Market: A New Approach to Transaction-Cost Measurement," *The Review of Financial Studies*, 6, 191–212. [20]

——— (1995), "One Security, Many Markets: Determining the Contributions to Price Discovery," *The Journal of Finance*, 50, 1175–1199. [18]

——— (2002), "Stalking the 'Efficient Price' in Market Microstructure Specifications: An Overview," *Journal of Financial Markets*, 5, 329–339. [18,20]

Huang, R. D., and Stoll, H. R. (1994), "Market Microstructure and Stock Return Predictions," *The Review of Financial Studies*, 7, 179–213. [34]

——— (1996), "Dealer Versus Auction Markets: A Paired Comparison of Execution Costs on NASDAQ and the NYSE," *Journal of Financial Economics*, 41, 313–357. [20]

Jacod, J. (1996), "La variation quadratique du brownien en presence d'erreurs d'arrondi," *Astérisque*, 236, 155–162. [20]

Jacod, J., Li, Y., Mykland, P. A., Podolskij, M., and Vetter, M. (2009), "Microstructure Noise in the Continuous Case: The Pre-Averaging Approach," *Stochastic Processes and Their Applications*, 119, 2249–2276. [25]

Jacod, J., and Protter, P. E. (2011), *Discretization of Processes*, Berlin, Heidelberg: Springer Science & Business Media. [21]

Kavajecz, K. A. (1999), "A Specialist's Quoted Depth and the Limit Order Book," *The Journal of Finance*, 54, 747–771. [18]

Lee, C., and Ready, M. J. (1991), "Inferring Trade Direction From Intraday Data," *The Journal of Finance*, 46, 733–746. [28]

Li, Y., Xie, S., and Zheng, X. (2016), "Efficient Estimation of Integrated Volatility Incorporating Trading Information," *Journal of Econometrics*, 195, 33–50. [18,20,22,23,27,29]

Li, Y., Zhang, Z., and Li, Y. (2018), "A Unified Approach to Volatility Estimation in the Presence of Both Rounding and Random Market Microstructure Noise," *Journal of Econometrics*, 203, 187–222. [34]

Li, Y., Zhang, Z., and Zheng, X. (2013), "Volatility Inference in the Presence of Both Endogenous Time and Microstructure Noise," *Stochastic Processes and Their Applications*, 123, 2696–2727. [21]

Madhavan, A., Richardson, M., and Roomans, M. (1997), "Why Do Security Prices Change? A Transaction-Level Analysis of NYSE Stocks," *The Review of Financial Studies*, 10, 1035–1064. [20]

McInish, T. H., and Wood, R. A. (1992), "An Analysis of Intraday Patterns in Bid/Ask Spreads for NYSE Stocks," *The Journal of Finance*, 47, 753–764. [19]

Petersen, M. A., and Fialkowski, D. (1994), "Posted Versus Effective Spreads: Good Prices or Bad Quotes?," *Journal of Financial Economics*, 35, 269–292. [20]

Potiron, Y., and Mykland, P. A. (2016), "Local Parametric Estimation in High Frequency Data" (to appear), *Journal of Business & Economic Statistics*. [22]

Roll, R. (1984), "A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market," *The Journal of Finance*, 39, 1127–1139. [18,20]

Tang, Y., and Zhang, Z. (2018), "A Combined Filtering Approach to High-Frequency Volatility Estimation With Mixed-Type Microstructure Noises," *Applied Stochastic Models in Business and Industry*, 1–21, DOI: 10.1002/asmb.2352. [35]

Todorov, V., and Tauchen, G. (2011), "Volatility Jumps," *Journal of Business & Economic Statistics*, 29, 356–371. [21]

Wood, R. A., McInish, T. H., and Ord, J. K. (1985), "An Investigation of Transactions Data for NYSE Stocks," *The Journal of Finance*, 40, 723–739. [19]

Xiu, D. (2010), "Quasi-Maximum Likelihood Estimation of Volatility With High Frequency Data," *Journal of Econometrics*, 159, 235–250. [22]

Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005), "A Tale of Two Time Scales: Determining Integrated Volatility With Noisy High-Frequency Data," *Journal of the American Statistical Association*, 100, 1394–1411. [23]