University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Brugler, J;Comerton-Forde, C

Title:
Comment on: Price Discovery in High Resolution

Date:
2021-01-01

Citation:
Brugler, J. & Comerton-Forde, C. (2021). Comment on: Price Discovery in High Resolution. Journal of Financial Econometrics, 19 (3), pp.431-438. https://doi.org/10.1093/jjfinec/nbz005.

Persistent Link:
https://hdl.handle.net/11343/258561

Comment on: Price Discovery in High Resolution

James Brugler and Carole Comerton-Forde*

Introduction

The microstructure literature comprises a rich set of papers that seek to understand pricing dynamics at a granular level, commonly exploring the joint dynamics of bids, asks and last sale prices. Its focus is on identifying innovations in prices and separating permanent price impacts from transient effects. Hasbrouck (1995) provides a tool that has been extensively utilized in the literature to examine these dynamics in many different market contexts over the last two decades.[1] However, the evolution of markets over this period, most notably the exponential growth in the volume of data and the increasing importance of trading speed has made the application of Hasbrouck's (1995) method and other related tools discussed in Hasbrouck (2018) more computationally and econometrically challenging. Hasbrouck (2018) offers a new approach to help overcome these challenges.

In this comment we briefly describe the evolution of markets and detail the challenges that these changes create for microstructure researchers and highlight the solution that Hasbrouck (2018) offers for these problems. We survey the literature that uses linear multivariate time-series models to understand high frequency markets. We focus on three examples from the literature to discuss how estimation constraints have affected their modelling choices, describe the potential drawbacks of these choices and how Hasbrouck's 2018 method can alleviate these constraints. We deliberately select papers that cover different asset classes: cash equities, fixed income and equity options. We hope that our discussion will help provide guidance about the costs and benefits of different modelling choices for future researchers confronted with a variety of methods to answer related research questions. We conclude by considering the implications of Hasbrouck's 2018 for the current policy debate on market data costs.

*Brugler is at the Department of Finance, University of Melbourne (james.brugler@unimelb.edu.au). Comerton-Forde is at the School of Banking and Finance at the UNSW Business School (c.comerton-forde@unsw.edu.au).

[1] Examples include Booth, Lin, Martikainen, and Tse (2002), Hendershott and Jones (2005), Anand and Subrahmanyam (2008) and Brogaard, Hendershott, and Riordan (2018).

Evolution of modern markets and the resultant research challenges

Technology and regulation have fundamentally altered the operation of financial markets over the last two decades.[2] The changes are most dramatic in equity markets. For example, in the US, regulatory change significantly increased the competition for equity trading services with activity now being split across 13 exchanges, approximately 40 Alternative Trading Systems (ATS) and over 200 broker-dealer internalizers. Automation and fragmentation contributed to the rise of high frequency trading (HFT) and substantially increased the emphasis on the speed of trading with interactions occurring at very high frequencies. Specialist HFT firms invest heavily in low latency connections to both trading venues and data providers as well as processing power to react to changes in market conditions at extremely fast speeds. HFT firms have become important players in most financial markets. They can observe changes in market conditions, including the state of the order book or the release of new information and update their desired trading strategy and alter their existing orders across a variety of trading venues on a micro-second time scale. One consequence of the presence of HFT firms is a dramatic increase in the volume of data produced by financial markets.[3]

Associations between financial variables are also characterized by dynamics that are potentially very long-lived relative to the highest frequency of interactions. Investors who want to trade in to or out of large positions have to perform a series of trades over a time-scale measured potentially in hours or days, or alternatively look for a counterparty off-exchange in the "upstairs" market. Toth, Palit, Lillo, and Farmer (2015) show that the direction of order-flow is positively autocorrelated over time-periods covering multiple days of trading, a fact that is primarily due to order splitting. Volatility forecasts that take into account slow-moving components at the weekly and monthly levels tend to outperform other models that do not account for long-lived dynamics (Corsi, 2009).

This combination of interactions taking place at sub-second time-frames with long-lived dynamics poses particular challenges to the estimation of linear multivariate time-series models such as vector autoregressions (VARs) and vector error correction models (VECMs). These have become workhorse models for understanding how modern financial markets function and various financial

[2] O'Hara (2015) describes in detail the changes in the trading landscape.

[3] For example, today the US equity markets generate over 10 billion messages per day, representing a more than 20 fold increase over the last decade. (See transcript of SEC Roundtable on Market data products, market access services and their associated fees held on 25 June, 2018 (page 54)).

variables interact with one another. When estimating VARs and VECMs in clock time, these models must simultaneously feature a very high resolution of time intervals (to capture high frequency interactions) and also an extremely large number of lags to capture dynamics over the order of multiple minutes, hours or days.

Unfortunately estimating models that have both features quickly becomes computationally infeasible. For example, suppose we wish to estimate a VAR for three variables: signed orderflow, the midquote to midquote return and the bid-ask spread. Suppose further that we wish to estimate the model at the millisecond level (to capture high-frequency interactions) and include a sufficient number of lags to cover one hour of clock time. In each equation, there are over 10 million parameters to estimate and more than 30 million parameters to estimate in the entire system. The parameter covariance matrix contains approximately $30 \times 31/2$ million parameters. Researchers working with these models often use one (or more) of three techniques to circumvent these estimation issues. The first is to aggregate time into more manageable intervals like seconds, minutes or hours. The cost to doing this is to induce a greater degree of simultaneity between the variables in a system. In other words, when time is aggregated to more coarse intervals, assumptions that structural shocks to each variable do not contemporaneously affect other variables in the system become less likely to hold, complicating inference.

The second is to retain a high resolution of lags but restrict the number of lags to a manageable number. As mentioned above, despite the high frequency of some interactions, many relevant financial time-series still display very long-lived serial dependence. Failure to incorporate sufficiently long-lived lags can induce serial correlation in the residuals and biases in relevant estimates.

The third is to estimate the model not in calendar time but in event time (also known as tick time) where time is updated only when a trade occurs, a quote is updated or another relevant event takes place, rather than at at regular chronological intervals. Estimating models in this way has the drawback of changing the calendar time span covered by the model in non-random ways. During particularly active periods, such as new information arriving in the market or during market stress, a fixed amount of event time covers a shorter amount of calendar time than during periods of relative inactivity and vice versa. Dynamics over longer calendar time periods are therefore over-sampled during periods of calm and under-sampled during periods of relative activity. The amount of calendar time between events has been shown to be important in the price formation

process (Dufour and Engle, 2000).

Hasbrouck's solution

Hasbrouck (2018) presents an alternative solution to this problem. Specifically, Hasbrouck suggests following Corsi (2009) and constraining model coefficients to a much smaller number of values than the total number of lags in each equation. In this setup, coefficients that capture the effect of one endogenous variable on another and that lie in pre-specified non-overlapping time-spans are restricted to be equal. Corsi (2009) uses similar restrictions in developing his heterogeneous autoregressive volatility models. This formulation has the advantage of allowing extremely high resolution in calendar time without the need to limit the maximum number of lags for computational reasons alone.

Using quote data for two large US stocks sourced directly from the trading venues that match orders and also from the consolidated tape, Hasbrouck (2018) demonstrates that models that aggregate time to coarse time intervals yield very different conclusions as to which record contributes the most to price discovery from systems of equations using sub-millisecond resolution. Models that use coarse intervals cannot distinguish between the contribution to the price discovery process of the consolidated tape from direct feeds. Using the proposed model with high resolution clearly demonstrates that the direct feeds dominate. Other applications from primary listing exchange vs. non-listing exchange and lit vs. dark venues further demonstrate the importance of retaining high resolution. At low resolution the information of the primary listing exchange is indistinguishable from non-listing exchanges, but at high resolution the listing exchange accounts for above one-half of the information shares despite accounting for only one fifth of trading volume. Further, at low resolution dark trades appear to have a small but discernible information contribution, but at high resolution they make almost no contribution to price discovery.

Revisiting some existing literature in the context of Hasbrouck (2018)

We next consider three important papers in the literature: Fleming, Mizrach, and Nguyen (2018), Groß-Klußmann and Hautsch (2011), and Chan, Chung, and Fong (2002), and consider how Hasbrouck (2018) may help to alleviate estimation constraints. We stop short of applying this approach to the some of the settings in the existing literature but encourage other authors to explore these

issues empirically.

Fixed income: Fleming, Mizrach, and Nguyen (2018)

Fleming et al. (2018) examine price discovery for on-the-run US treasury securities trading on the BrokerTec electronic communication network using a structural VAR approach. The base model consists of two endogenous variables: mid-quote to mid-quote return and orderflow for the tth transaction. Identification is achieved by a standard theoretical ordering of the contemporaneous causality between variables (order flow affects returns contemporaneously but the reverse is not true). The lag length is set to five.

In an important extension of the model, the endogenous variables are augmented with the net change in the number of first-tier limit buy and sell orders between trades. These variables are designed to capture changes in the state of the order book between trades. This complicates the estimation of the model in event time because the number of changes in the order book (4.7 million changes per day at the best five levels) are orders of magnitude larger than the number of trades (approximately 12,000 trades per day). Rather than reformulate the model in event time where an event is defined as a change in the order book of any type (including a trade), event time in Fleming et al. (2018) continues to update when an actual trades occur.

Augmenting the model in this way requires restrictions such as a full ordering of all endogenous variables in the model. Fleming et al. (2018) make the assumption that the changes in the number of buy and sell orders between trades are not contemporaneously correlated. In other words, a shock to the arrival or cancellation rate of buy (sell) orders is assumed to have no effect on the quantity of sell (buy) orders in the order book between trades.

Noting that this assumption is made purely for estimation purposes, it is worth considering reasons why it might fail. It is plausible that market participants monitor the state of the order book and adjust their resting limit orders accordingly. For example, upon observing a large buy limit order enter the order book, an algorithmic trader that is executing a sell order may choose to increase the price of their existing sell orders, decreasing the depth at the best ask and implying that net changes in bids and asks are not independent between trades. Indeed if limit orders are contributing to the price formation process, some kind of effect like this is necessarily occurring.

Alternatively, formulating the problem using Hasbrouck's (2018) method, Fleming et al. (2018)

could reformulate the model in calendar time but at the millisecond level (the highest frequency of time stamps in the data) or in event-time but defining an event as a change in the state of the orderbook. Using the distributed lag model would allow feasible estimation of the relevant parameters with sufficiently long-lived lags. This would allow inference without resorting to such strong assumptions about causal ordering of order book changes between trades.

Cash equities: Groß-Klußmann and Hautsch (2011)

In our second example, computational considerations directly affect the modelling choice made by the authors. Groß-Klußmann and Hautsch (2011) examine the role that non-scheduled news arrivals have on high frequency returns, volatility and liquidity in the UK equity market. Using one and a half years of data across 40 stocks, the authors estimate a VAR using standard volume, volatility and liquidity variables as well as dummies for the arrival of relevant machine readable news with variables aggregated to 20 second intervals. An additional modelling decision is that the VAR model is estimated using only periods when a trade has taken place (i.e. the sample used for estimation conditions on trade arrivals).

Regarding the aggregation issue and while noting that 20 seconds is a relatively high frequency, it is still possible that aggregation to such a time frame can cloud the inference that can be drawn from these models. This has been as discussed above and evidence in Hasbrouck (2018) demonstrates the potential importance of this issue.

Here we focus on the second modelling choice: to estimate the model using only periods with non-zero trading. Groß-Klußmann and Hautsch (2011) choose to do this explicitly because of the high frequency and sparsity in their data. The authors then think carefully about conditions under which this is a valid approach. They derive the likelihood function for a general VAR data generating process and decompose it into additive components that condition on a trade occurring and a trade not occurring. Groß-Klußmann and Hautsch (2011) demonstrate that so long as the parameters that govern the distribution of the endogenous variables conditional on a trade occurring or not occurring and those that govern the probability of a trade occurring are disjoint, the likelihood components can be maximized separately. Under this assumption, their modelling choice will not affect the inference drawn from their estimation of the VAR.

A potential concern with this assumption is that because value traded in a given period is also a

variable in the VAR, any non-zero autocorrelation in this variable (which would be captured by the parameters of the VAR) must also affect the unconditional probability of observing non-zero trading value in a given period. To see this, consider the case where trade arrivals have zero autocorrelation and compare it with a case where trade arrivals are highly persistent. The unconditional probability that a given period contains a trade is higher in the second scenario relative to the first, all else being equal. This implies that the parameter sets are arguably not disjoint and that these parameters should instead be estimated jointly using all time periods.

The computational burden of doing so be alleviated using the method proposed by Hasbrouck (2018). The model could also be estimated at higher frequencies, which may also be beneficial given that the VAR is written in reduced form, effectively abstracting from contemporaneous effects between the trading variables.

Equity options: Chan, Chung, and Fong (2002)

Chan et al. (2002) examine the price discovery roles of quote revisions and trades in stock and option markets respectively. To do this, Chan et al. (2002) obtain second-by-second quotes and trades for stock options listed on the Chicago Board Options Exchange and stocks listed on the NYSE, Nasdaq and AMEX. They estimate a VAR with the midquote-to-midquote return and signed volume traded for the cash equity, calls and puts (six variables) and is estimated using data aggregated to five minute intervals.

A key identifying assumption discussed in the model is the ordering such that trades affect returns but the reverse is not true (similar to Fleming et al., 2018). At sufficiently high frequencies, this assumption is well-founded (in Hasbrouck, 1991 this assumption is derived for a model estimated in event time). When trades and quotes are aggregated over sufficiently coarse time periods, this assumption becomes potentially more restrictive (leverage effects, for example, would suggest that this condition does not hold outside of event time). Furthermore, returns and volumes across stocks, calls and puts are very likely to be contemporaneously correlated at five minute intervals. For returns, this could be due to arbitrage while for volume this could be due to the practice of delta hedging options positions by market makers. The econometric approach of Chan et al. (2002) abstracts from this by excluding contemporaneous (or alternatively, structural) terms in the relevant VAR equations.

Chan et al. (2002) contains some evidence regarding the potential importance of the aggregation issue. When estimating the model using 5 minute aggregation, they find that net stock volume affects future stock and option prices (with the predicted signs) while call and put volumes tend to predict returns but with the wrong signs. However, when the model is estimated using 100s intervals as a robustness check, this puzzling finding is almost entirely eliminated. This demonstrates that the degree of aggregation has the potential to affect the inference that can be drawn from these models.

Estimating with higher frequency time intervals would be feasible using the method proposed by Hasbrouck (2018) without necessarily sacrificing the ability to capture longer lived dynamics. The assumptions that returns and volumes across the underlying security and its derivatives are not contemporaneously correlated becomes less restrictive as the sampling frequency increase.

Implications for current debate on market data

To conclude, we consider the implications of Hasbrouck's (2018) findings for current regulatory policy discussions relating to market data.

There is intense debate in US equity markets about market data. There are two key concerns: (i) the two-tiered system created by the presence of the Security Information Processor (SIP)[4] and proprietary direct exchange feeds, and (ii) the growing cost of direct exchange feeds.[5] Although exchanges are required to disseminate data to the SIP and direct customers simultaneously, there are differences in the latency of the SIP and direct exchange feeds, which raise issues about fairness.[6] The cost debate centers around the question of whether the cost of data is fair and reasonable.[7]

The US Securities and Exchange Commission (SEC) hosted roundtables in October 2018 to debate these issues. On one side of the debate were the exchanges who argue that market data costs are fair and reasonable, and that brokers and investors can make a commercial decision about whether to pay for direct exchange feeds or to rely on the SIP consolidated tape. On the other

[4]The SIP links the U.S. markets by processing and consolidating all protected bid/ask quotes and trades from every trading venue into a single consolidated data feed. The SIP disseminates important regulatory information including the National Best Bid and Offer (NBBO), Limit Up Limit Down (LULD) price bands, short sale restrictions and regulatory halts.

[5]The cost of market data is also an issue in other jurisdictions where there is competition for trading services.

[6]The Consolidated Tape Association website reports that the current median latency is about 230 microseconds. Sourced from `https://www.ctaplan.com/indexon1December2018`.

[7]In the US the requirement for providing market data at fair and reasonable costs is set out in the Securities Exchange Act of 1934. Other jurisdictions also have similar requirements set out in legislation.

side of the debate were brokers, high frequency traders and long-term investors. Brokers and long-term investors argue that the need for purchasing direct exchange feeds is critical for fulfilling their best execution obligations and for delivering good outcomes for their customers. Brokers, high frequency traders and long-term investors all argue that not purchasing direct feeds put them at an informational disadvantage making it more difficult to manage trading processes and trading costs.

Hasbrouck's (2018) findings can help inform this debate. Like many others he documents the maximum latency between the SIP consolidated and direct feeds, indicating it ranges from 2.52 to 181.63 milliseconds for trades and 18.32 to 86.17 milliseconds for quotes.[8] He also demonstrates that at one second resolutions direct feeds are indistinguishable from the SIP consolidated tape, but that at 10 and 100 microsecond resolutions the direct feeds dominate the SIP consolidated tape. Together these results provide clear empirical evidence that participants relying purely on the SIP consolidated tape are at an informational disadvantage. Therefore, the decision to purchase a direct feed is more than a just a commercial decision, it should be obligatory for anyone with a best execution obligation or making frequent trading decisions.

Direct feeds and the SIP consolidated tape also differ in the level of information provided. Direct feeds provide access to the full depth of book for each exchange, while the consolidated tape only provides the top of the book quotes. Future research could also use Hasbrouck's (2018) approach to determine the incremental information learned from the depth of book information. This would be helpful to inform decisions about the future development of the SIP consolidated tape.

[8] These estimates exclude ADF trades.

References

Anand, A. and A. Subrahmanyam (2008). Information and the intermediary: Are market intermediaries informed traders in electronic markets? *Journal of Financial and Quantitative Analysis 43*(1), 1–28.

Booth, G., J.-C. Lin, T. Martikainen, and Y. Tse (2002). Trading and pricing in upstairs and downstairs stock markets. *Review of Financial Studies 15*(4), 1111—-1135.

Brogaard, J., T. Hendershott, and R. Riordan (2018). Price Discovery without Trading: Evidence from Limit Orders. *Journal of Finance (Forthcoming).*

Chan, K., Y. P. Chung, and W.-M. Fong (2002). The informational role of stock and option volume. *The Review of Financial Studies 15*(4), 1049–1075.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics 7*(2), 174–196.

Dufour, A. and R. F. Engle (2000). Time and the price impact of a trade. *The Journal of Finance 55*(6), 2467–2498.

Fleming, M. J., B. Mizrach, and G. Nguyen (2018). The microstructure of a US Treasury ECN: The BrokerTec platform. *Journal of Financial Markets 40*, 2–22.

Groß-Klußmann, A. and N. Hautsch (2011). When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance 18*(2), 321–340.

Hasbrouck, J. (1991). Measuring the information content of stock trades. *The Journal of Finance 46*(1), 179–207.

Hasbrouck, J. (1995). One security, many markets: Determining the contributions to price discovery. *The Journal of Finance 50*(4), 1175–1199.

Hasbrouck, J. (2018). Price discovery in high resolution. NYU Working Paper.

Hendershott, T. and C. Jones (2005). Island goes dark: Transparency, fragmentation, and regulation. *Review of Financial Studies 18*(3), 743—793.

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics 116*(2), 257–270.

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