Comment on: Price Discovery in High Resolution*

Giuseppe Buccheri[1], Giacomo Bormetti[2], Fulvio Corsi[3,4], and Fabrizio Lillo[2,5]

[1]Scuola Normale Superiore, Italy

[2]University of Bologna, Italy

[3]University of Pisa, Italy

[4]City University of London, UK

[5]CADS, Human Technopole, Milan, Italy

March, 2019

Abstract

This note is commenting on Hasbrouck (2018). The paper investigates the problem of price discovery on markets with trades recorded at sub-millisecond frequencies. The application of the popular information share measure of Hasbrouck (1995) to such data faces several difficulties, as the underlying VECM would need a huge number of lags to capture dynamics at different time-scales. The problem is handled by imposing a set of restrictions on parameters inspired by the Heterogeneous Autoregressive model for realized volatility. We illustrate some potential drawbacks of the information share measure adopted in the paper and propose a modelling strategy aimed at dealing with such limitations. In particular, we introduce a structural multi-market model with a lagged adjustment mechanism describing lagged absorption of information across markets. The advantages of the method are shown in simulations.

Keywords: High-resolution, High-frequency trading, Information share, HAR, Lagged-adjustment

JEL codes: C32, D58, G14

*Corresponding author: fulvio.corsi@unipi.it, University of Pisa, Department of Economics and Management, Via C. Ridolfi, 10 56124 Pisa, Italy.

1 Motivation and main contributions

Modern financial markets are heavily driven by algorithms which operate at very short time scales. To allow for such high speed strategies, new trading venues record quotes and trades up to the microsecond or nanosecond precision. The econometric analysis of these ultra-high-frequency data is challenging, since models need to capture both the fast dynamics of algorithms and the slower dynamics of human agents. In this high resolution framework, Hasbrouck (2018) examines the traditional problem of determining the contribution of an individual market to the price formation of homogeneous or closely-linked securities. The methodology is based on the pioneering work of Hasbrouck (1995), who introduced the notion of Information Share (IS) and employed vector error correction models (VECM) to investigate price discovery on fragmented markets.

The application of the methodology of Hasbrouck (1995) to data recorded at sub-millisecond frequencies faces several difficulties. Apart from the standard issue of non-uniqueness of information share, for which only lower and upper bounds are available, traditional VECM analysis requires a large number of lags to describe dynamics at different time-scales and thus a huge number of parameters to be estimated.

The paper provides two main contributions to the market microstructure literature on price discovery at ultra-high-frequencies. First, the problem of dealing with a large number of lags is handled by imposing a set of restrictions on VECM coefficients inspired by the Heterogeneous Market Hypothesis of Muller et al. (1993) and the Heterogeneous Autoregressive (HAR) model of Corsi (2009). These restrictions follow a step-function scheme, with coefficients being equal on predetermined time-scales. The resulting VECM specification remains parsimonious in presence of many lags (from 10 up to 10^6 in the applications reported in the paper) and provides a unified approach to modelling price dynamics across a wide range of frequencies.

Second, the paper shows that increasing the resolution at which data are analyzed is essential to determine in which market price discovery occurs. To illustrate this interesting result, three different empirical analysis are performed. The first analysis studies the information content of the traditional source of market data, the consolidated tape, and that of prices known to market participants who directly subscribe to feeds of individual exchanges. The second analysis is related to price formation in listing and non-listing exchanges. Finally, the third investigates information in quotes, lit trades and dark trades. The main advantage of the method is that IS bounds progressively shrink as the resolution increases. At the time scale of one second, information shares are indeterminate. As the resolution increases (up to 10 microseconds in the empirical application), a clearer picture emerges: (i) the information content of direct subscribers is substantially larger than that available from the consolidated tape; (ii) listing exchanges are slightly more informative than non-listing exchanges; (iii) quotes and lit trades are substantially more informative than dark trades. Results from the impulse response function (IRF) analysis are also reported.

1.1 VECM with heterogeneous structure

In this section we briefly illustrate the information share measure introduced by Hasbrouck (1995) and the methodology adopted by Hasbrouck (2018) to deal with high resolution data. Let $p_t \in \mathbb{R}^n$ be a vector of log-prices related to a single security (e.g. transaction prices in n markets or bid and ask quotes in $n/2$ markets). The fact that the prices in p_t pertain the same security implies that they are cointegrated, meaning that they cannot move "too far away" from each other. Let ι_n denote an n-dimensional vector of ones. Formally, the vector $[p_t^{(1)} - p_t^{(2)}, p_t^{(1)} - p_t^{(3)}, \ldots, p_t^{(1)} - p_t^{(n)}]' \in \mathbb{R}^{n-1}$ is covariance stationary and the dynamics of first differences of p_t can

be represented by the VECM model:

$$\Delta p_t = \boldsymbol{\alpha}\boldsymbol{\beta}' p_{t-1} + \boldsymbol{\phi}_1 \Delta p_{t-1} + \boldsymbol{\phi}_2 \Delta p_{t-2} + \cdots + \boldsymbol{\phi}_k \Delta p_{t-k} + \epsilon_t \tag{1}$$

where $\mathrm{Cov}[\epsilon_t] = \boldsymbol{\Omega}$, $\boldsymbol{\alpha} \in \mathbb{R}^{n\times(n-1)}$ is a loading matrix, $\boldsymbol{\beta}' = [\iota_{n-1}, -\mathbf{I}_{n-1}]$ is a matrix of cointegrated vectors and $\boldsymbol{\phi}_i \in \mathbb{R}^{n\times n}$, $i = 1, \ldots, k$ are autoregressive matrix coefficients. The Engle-Granger representation theorem implies that the VECM has both a vector moving average (VMA) and a common trend representation. The VMA representation is given by:

$$\Delta p_t = \boldsymbol{\theta}(L)\epsilon_t \tag{2}$$

where $\boldsymbol{\theta}(L) = \sum_{t=0}^{\infty} \boldsymbol{\theta}_i L^i$ is a polynomial lag operator. The common trend representation can be written as:

$$p_t = p_0 + \boldsymbol{\Xi} \sum_{i=1}^{t} \epsilon_i + \boldsymbol{\Xi}^* \epsilon_t \tag{3}$$

Here the second term is a random walk component common to all prices that can be regarded as the efficient price process. The last term is a stationary process describing deviations of prices observed on different markets from the efficient price process. The matrix $\boldsymbol{\Xi}$ coincides with the sum of VMA coefficients:

$$\boldsymbol{\Xi} = \sum_{t=0}^{\infty} \boldsymbol{\theta}_i \tag{4}$$

It can also be computed explicitly as (see e.g. Kilian and Lütkepohl 2017):

$$\boldsymbol{\Xi} = \boldsymbol{\beta}_\perp \left[\boldsymbol{\alpha}'_\perp \left(\mathbf{I}_n - \sum_{i=1}^{k} \boldsymbol{\phi}_i \right) \boldsymbol{\beta}_\perp \right]^{-1} \boldsymbol{\alpha}'_\perp \tag{5}$$

where $\boldsymbol{\alpha}_\perp$, $\boldsymbol{\beta}_\perp \in \mathbb{R}^n$ denote the orthogonal complement[1] of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. The particular structure of $\boldsymbol{\beta}$ implies that $\boldsymbol{\beta}_\perp$ is proportional to ι_n and thus the rows of $\boldsymbol{\Xi}$ are identical.

The information share of the i-th market was defined by Hasbrouck (1995) as the fraction of variance of the random walk component in eq. (3) that is explained by the i-th market. If $\boldsymbol{\Omega}$ is diagonal, the latter can be computed as:

$$S_i = \frac{\xi_i^2 \boldsymbol{\Omega}_{ii}}{\xi \boldsymbol{\Omega} \xi'} \tag{6}$$

where ξ denotes the common row of $\boldsymbol{\Xi}$ and ξ_i is the i-th component of ξ. If $\boldsymbol{\Omega}$ is not diagonal, one can write the innovations of the VECM as:

$$\epsilon_t = \mathbf{C} z_t \tag{7}$$

where $\mathbf{C}$ is the Cholesky decomposition of $\boldsymbol{\Omega}$ and z_t are orthogonal innovations. S_i can therefore be computed based on z_t. However, the allocation of the total variance among markets depends on the particular order with which they appear on the price vector p_t. Consequently, only lower and upper bounds of S_i can be computed by performing the VECM analysis for all possible permutations of variables.

When studying ultra-high-frequency data, the VECM model in eq. (1) should contain a number k of lags sufficient to incorporate dynamics at different time-scales, from the microsecond resolution of algorithms up to the scales of human agents. Accounting for this wide variety of scales would lead to a huge number of parameters to be estimated. The solution adopted by Hasbrouck (2018) is to impose an HAR structure (Corsi 2009) on the VECM coefficients. The heterogeneous VECM reads:

$$\Delta p_t = \boldsymbol{\alpha}\boldsymbol{\beta}' p_{t-1} + \boldsymbol{\phi}^{(1)} \Delta p_{t-1|t-k_1} + \boldsymbol{\phi}^{(2)} \Delta p_{t-k_1-1|t-k_2} + \cdots + \boldsymbol{\phi}^{(d)} \Delta p_{t-k_{d-1}-1|t-k_d} + \epsilon_t \tag{8}$$

[1] Given a matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$, $p \geq q$, the orthogonal complement $\mathbf{A}_\perp$ is any $p \times (p-q)$ matrix such that $\mathbf{A}'\mathbf{A}_\perp = 0$.

where $\Delta p_{p|q} = \sum_{i=p}^{q} \Delta p_i$. Here d denotes the number of time-scales and the k_i's, $i = 1, \ldots, d$ establish the range of each time-scale. The k coefficients $\boldsymbol{\phi}_s$ appearing in eq. (1) are constrained to be equal to $\boldsymbol{\phi}^{(i)}$ for $s \in [t-k_{i-1}-1, t-k_i]$. For instance, at the frequency of 10 millisecond, Hasbrouck (2018) sets $d = 4$, $k_1 = 1$, $k_2 = 10$, $k_3 = 100$, $k_4 = 1000$, corresponding to time-scales of $[0.01, 0.01]$, $[0.02, 0.1]$, $[0.11, 1]$, $[1.01, 10]$ seconds. The VECM with heterogeneous structure thus provides a unified framework to model price dynamics across a wide range of time-scales.

2 Potential limitations of the methodology

The application of the VECM methodology to data recorded at sub-millisecond frequencies faces several difficulties. First, as underlined by the author, at these frequencies trades are highly sparse and a lot of zero returns are produced by previous-tick interpolation. A consequence of zero returns is that high-frequency prices strongly deviate from the standard semimartingale assumption (Bandi et al. 2017). Neglecting this essential feature of data can lead to several forms of distortions. For instance, zero-returns are the main determinant of the Epps effect (Epps 1979), i.e. the downward bias of sample covariances as the sampling frequency increases (see e.g. Hayashi and Yoshida 2005 and Buccheri et al. 2018b). Due to the high level of sparsity, the VECM is clearly misspecified and OLS estimates are potentially biased.

Second, market microstructure effects, which are relevant at ultra-high frequencies, are not included in the VECM specification. This leads to a classical error-in-variables model and thus constitutes an additional source of bias for OLS estimates. Nguenang (2016) and Dias et al. (2018) discussed different methods to account for microstructure noise on information share measures. It would be interesting examining the impact of both sparsity and measurement errors on OLS estimates, in order to exclude that the observed shrinkage of IS bounds is not due to spurious factors.

In this section we show that, even in absence of sparsity and microstructure noise, the information share measure of Hasbrouck (1995) can be subject to other potential limitations. The possibility of retaining a parsimonious VECM specification at high resolution is one of the main contributions of the paper. The matrix coefficients $\boldsymbol{\phi}^{(i)}$, $i = 1, \ldots, d$, in the heterogeneous VECM capture price dynamics on a wide range of time-scales and can easily be estimated by OLS. However, they do not enter directly in the expression of information shares. This is shown by noticing that the term:

$$\boldsymbol{\alpha}'_{\perp} \left(\mathbf{I}_n - \sum_{i=1}^{k} \boldsymbol{\phi}_i \right) \boldsymbol{\beta}_{\perp} \tag{9}$$

appearing in eq. (5) is scalar and thus it cancels out when taking the ratio in eq. (6). Thus, information share depends on the loading parameters $\boldsymbol{\alpha}$, on the covariance matrix $\boldsymbol{\Omega}$ of the efficient price innovations but *not* on VECM coefficients $\boldsymbol{\phi}^{(i)}$. This implies that potentially relevant dynamic features captured by the long lag structure are not reflected on information shares.

Autoregressive coefficients play a relevant role in determining the dynamics of the VECM process. Markets with same loading parameters in $\boldsymbol{\alpha}$ and same variances in $\boldsymbol{\Omega}$ can behave very differently if a lag structure exists. Let us consider a bivariate system of two markets obeying the VECM process in eq. (1) and, for simplicity, let us assume $k = 1$, i.e. that only one lag exists. We set VECM parameters as:

$$\boldsymbol{\alpha} = \begin{pmatrix} 0.1 \\ 0.1 \end{pmatrix}, \quad \boldsymbol{\Omega} = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}, \quad \boldsymbol{\phi}_1 = \begin{pmatrix} -0.6 & -0.8 \\ 0.4 & 0.1 \end{pmatrix} \tag{10}$$

We generate $T = 1000$ observations of this VECM model. Figure 1 shows the paths of the two markets and their common trend in the subsample comprising the first 200 observations. While the second market moves closely to

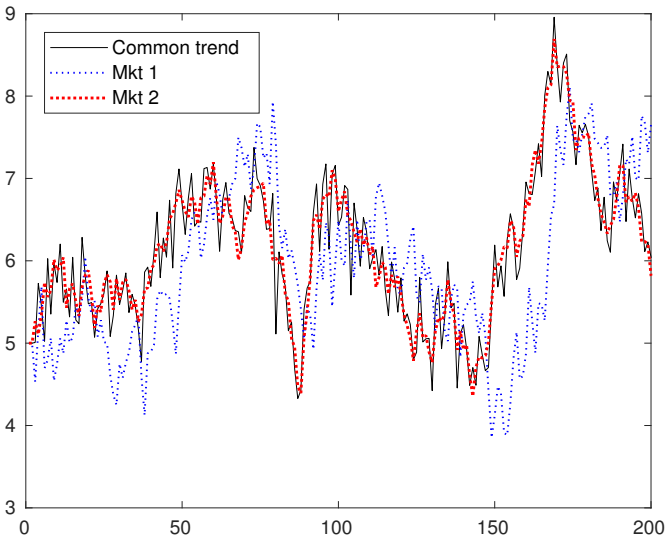

Figure 1: Simulated paths of a bivariate VECM process with one lag and the associated common trend. The coefficient matrix $\boldsymbol{\phi}_1$ induces a significant delay on the dynamic of the first market.

the common trend, the first market exhibits a significant delay. The second market is clearly more informative, as it quickly incorporates new information. However, as indicated by eq. (5) and (6), information share is the same in the two markets, since they have identical loading parameters and variances. Indeed, replacing $\boldsymbol{\alpha}$ and $\boldsymbol{\Omega}$ with OLS estimates, eq. (6) gives $S_1 = 0.503$, $S_2 = 0.496$. Note that the two information shares are uniquely defined, as $\boldsymbol{\Omega}$ is diagonal. Due to the fact that S_i does not depend on $\boldsymbol{\phi}_1$, we would find the same result for *any* choice of $\boldsymbol{\phi}_1$. Even adding further lags $\boldsymbol{\phi}_2, \boldsymbol{\phi}_3, \ldots, \boldsymbol{\phi}_k$ would still provide $S_1 = S_2$, as their effect would cancel out when computing S_i. Yet, the new lags can further affect the dynamics of the two markets, determining e.g. a longer delay in the first market or delays in both markets.

In some specific cases, information shares can lead to other counterintuitive conclusions. From eq. (5) and (6), note that S_i depends negatively on $\boldsymbol{\alpha}$, meaning that markets with small adjustments to the common trend are more informative. At the same time, S_i depends positively on the diagonal elements of $\boldsymbol{\Omega}$, meaning that markets with larger variance are more informative. However, the dependence of S_i on these two factors is such that a fast adapting market with a small variance can be as informative as a slowly adapting market with a large variance. In order to illustrate this effect, consider a bivariate VECM model with no lags. The elements of $\boldsymbol{\alpha}$, $\boldsymbol{\Omega}$ are set as:

$$\boldsymbol{\alpha} = \begin{pmatrix} 0.01 \\ 0.1 \end{pmatrix}, \quad \boldsymbol{\Omega} = \begin{pmatrix} 0.1 & 0 \\ 0 & 10 \end{pmatrix} \tag{11}$$

Parameters with different magnitudes are intentionally chosen with the aim of emphasizing the effect. As in the previous example, we generate $T = 1000$ observations of the VECM and report in fig. 2 the subsample comprising the first 200 observations. The first market has a smaller loading parameter and thus it rapidly adjusts to the common trend. The second market is slower but has a much larger variance. After estimating the VECM, we find $S_1 = 0.526$, $S_2 = 0.473$, implying that the two markets have almost the same information content. However, one would expect the first market being definitely more informative than the second. Note that nothing prevents increasing further the variance of the second market to have $S_2 > S_1$.

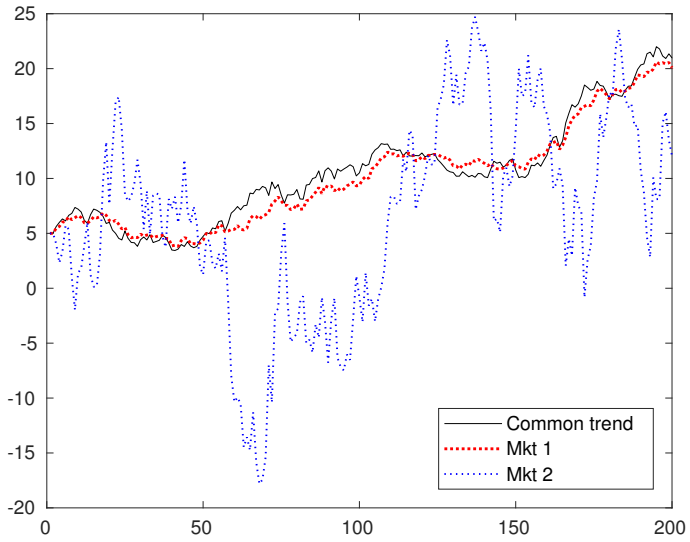

Figure 2: Simulated paths of a bivariate VECM process with no lags and the associated common trend. The second market has a significantly larger variance but same information share as the first market.

3 A multi-market lagged adjustment model

In this section we propose a method to deal with the difficulties underlined above. As in De Jong and Schotman (2010), our information share measure is based on a structural model that can be cast in a state-space representation. In the structural model, prices in different markets deviate from the efficient martingale process because of two factors: (i) market microstructure effects, (ii) lagged absorption of information across markets. The latter is modeled by extending to a multivariate framework the lagged adjustment mechanism of Hasbrouck and Ho (1987). Information shares depend explicitly on lagged adjustments matrices and therefore account for lead-lag effects among markets. Sparse trades arising at high frequencies can be handled as a typical missing values problem in the state-space representation. We therefore avoid the introduction of artificial zero returns, which are a potential source of bias.

3.1 Model

Let us assume that the intraday dynamics of an asset log-price are driven by an efficient process p_t^* evolving over time as a random walk:

$$p_{t+1}^* = p_t^* + \omega_{t+1}, \quad \text{Var}[\omega_t] = \sigma^2 \tag{12}$$

Let $Y_t \in \mathbb{R}^n$ be observations of the asset log-price on n different markets. We write Y_t as:

$$Y_t = P_t + \epsilon_t, \quad \text{Cov}[\epsilon_t] = \mathbf{\Sigma} \tag{13}$$

where P_t is an "adjusted" price process and ϵ_t is a white noise describing market microstructure effects. The adjusted price reflects lagged absorption of information across different markets. For instance, one market can react very quickly to new information and thus the price observed in that market adjusts almost instantaneously. In contrast, if a market reacts slowly, the adjustment process is delayed and the information is absorbed with a certain lag.

In order to incorporate such mechanism into our microstructure model, we introduce a multivariate generalization of the "lagged price adjustment" model of Hasbrouck and Ho (1987). We write:

$$P_{t+1} = P_t + \mathbf{\Psi}(\iota_n p^*_{t+1} - P_t) \tag{14}$$

where $\mathbf{\Psi} \in \mathbb{R}^{n \times n}$ is a lagged adjustment matrix. In order to grasp the intuition behind eq. (14), let us assume that $\mathbf{\Psi} = \mathbf{I}_n$, i.e. $\mathbf{\Psi}$ is equal to the n-dimensional identity matrix. In this particular case, eq. (14) reduces to $P_{t+1} = \iota_n p^*_{t+1}$, meaning that the price instantaneously adjusts to new information in all the n markets. This is the case of a "perfect" market, i.e. a market without frictions where the flow of information immediately reflects into prices. Assume now that the lagged adjustment matrix is diagonal, but different from the identity matrix. In this case, the adjustment process is not instantaneous, since prices adjust with a finite speed depending on the diagonal elements of $\mathbf{\Psi}$. Note that the adjustment processes in different markets are independent, meaning that the price in one market does not affect the price in a different market. Finally, assume that $\mathbf{\Psi}$ has nonzero off-diagonal elements. As before, the adjustment process is not instantaneous and the speed of adjustment of the price in one market depends on the elements on the corresponding row of $\mathbf{\Psi}$. However, the adjustment processes are not independent, since the level of the price in one market can affect the price in a different market and the strength of such lead-lag effects is characterized by the off-diagonal elements of $\mathbf{\Psi}$.

This structural model shares some similarities with that of De Jong and Schotman (2010). The main difference is that we introduce the lagged adjustment mechanism, which allows to describe lagged absorption of information across different markets. The proposed model belongs to the class of "Multi-Asset Lagged Adjustment" (MLA) models introduced by Buccheri et al. (2018a). However, while they considered the case of several assets traded in the same market, we consider a different problem, namely the case of one asset traded on several markets. Compared to a general MLA model, here prices are constrained to move "not too far" from their *common trend*, represented by the unobserved efficient price p^*_t.

In this structural representation, one can easily characterize the main determinants of price discovery. We will say that one market is more informative than another if the price in that market is "closer" to the common trend. Deviations from the latter are due to microstructure effects, described by the covariance matrix $\mathbf{\Sigma}$, and to the lagged adjustment mechanism, characterized by the matrix $\mathbf{\Psi}$. As such, one market is informative if microstructure effects are small and if it adjusts rapidly to new information. In Section 3.2, we will formalize this intuitive notion through a measure of information share depending explicitly on both $\mathbf{\Sigma}$ and $\mathbf{\Psi}$.

Note that eq. (12), (14) imply:

$$\Delta P_{t+1} = (\mathbf{I}_n - \mathbf{\Psi})\Delta P_t + \mathbf{\Psi}\iota_n \omega_{t+1} \tag{15}$$

where $\Delta P_{t+1} = P_{t+1} - P_t$ are log-returns. From this VAR(1) representation, it is evident that the returns of the adjusted prices are driven by a common innovation, coinciding with that of the efficient price. Stability requires that all the eigenvalues of $\mathbf{I}_n - \mathbf{\Psi}$ lie inside the unit circle. As shown in Appendix A, the model

$$Y_t = P_t + \epsilon_t \tag{16}$$

$$\Delta P_{t+1} = (\mathbf{I}_n - \mathbf{\Psi})\Delta P_t + \mathbf{\Psi}\iota_n \omega_{t+1} \tag{17}$$

has a linear state-space representation. It can thus be estimated through quasi-maximum likelihood using the Kalman filter. This estimation method is particularly convenient when dealing with sparse high-frequency data, since the Kalman filter can handle with missing observations. We can therefore treat sparsity and asynchronicity as a typical

missing value problem in linear-Gaussian state-space models, in a similar fashion to Corsi et al. (2015). Note that adding further lags in the transitions equation (17) is possible at the expense of increasing the dimensionality of the matrices appearing in the state-space representation in Appendix A.

3.2 Price discovery

In order to define a measure of information share in our structural model, we follow a logic similar to that of De Jong and Schotman (2010). Let us consider the observed price innovations (computed with respect to the efficient price):

$$V_t = Y_t - \iota_n p^*_{t-1} \tag{18}$$

In Appendix B, we prove the following:

Lemma 1. *If all the eigenvalues of* $(\mathbf{I}_n - \mathbf{\Psi})$ *lie inside the unit circle, the observed price innovation* V_t *has the following Wold representation:*

$$V_t = -\sum_{i=2}^{\infty}(\mathbf{I}_n - \mathbf{\Psi})^i \iota_n \omega_{t-i+1} + \mathbf{\Psi}\iota_n\omega_t + \epsilon_t$$

The main difference with the expression of V_t appearing in De Jong and Schotman (2010) is the presence of the series $\sum_{i=2}^{\infty}(\mathbf{I}_n - \mathbf{\Psi})^i \iota_n \omega_{t-i+1}$ and the coefficient $\mathbf{\Psi}$ multiplying the efficient price innovations ω_t. They arise as a consequence of the lagged adjustment mechanism introduced through eq. (14). The unconditional covariance of V_t is given by:

$$\mathbf{\Upsilon} = \text{Cov}[V_t] = \sigma^2 \sum_{i=2}^{\infty}(\mathbf{I}_n - \mathbf{\Psi})^i \iota_n \iota_n' (\mathbf{I}_n - \mathbf{\Psi})^{i\prime} + \sigma^2 \mathbf{\Psi}\iota_n\iota_n'\mathbf{\Psi}' + \mathbf{\Sigma} \tag{19}$$

We now consider the linear regression of V_t onto the innovations of the efficient price:

$$\omega_t = \mathbf{\Gamma}' V_t + \zeta_t \tag{20}$$

where $\mathbf{\Gamma} \in \mathbb{R}^n$ is a vector of coefficients given by:

$$\mathbf{\Gamma} = \mathbf{\Upsilon}^{-1}\text{Cov}[V_t, \omega_t] = \mathbf{\Upsilon}^{-1}\mathbf{\Psi}\iota_n\sigma^2 \tag{21}$$

The R^2 of this regression quantifies the fraction of variance of the efficient log-price process p^*_t explained by the n observed log-prices. It can be written as:

$$R^2 = \frac{\mathbf{\Gamma}'\mathbf{\Upsilon}\mathbf{\Gamma}}{\sigma^2} = \mathbf{\Gamma}'\mathbf{\Psi}\iota_n \tag{22}$$

Setting $\mathbf{\Psi}\iota_n = \gamma$, we see that the R^2 can be decomposed as the sum of n terms:

$$R^2 = \mathbf{\Gamma}'\gamma = \sum_{i=1}^{n}\mathbf{\Gamma}_i\gamma_i \tag{23}$$

In a similar fashion as De Jong and Schotman (2010), we define the information share of the i-th market as $S_i = \mathbf{\Gamma}_i\gamma_i$. Each of these terms quantifies the contribution of the i-th market to the explained variance of the efficient price process. Note that S_i depends explicitly on both $\mathbf{\Sigma}$, which is related to microstructure noise, and on $\mathbf{\Psi}$, which describes the lagged adjustment mechanism.

Figure (3) shows the case where one market is largely affected by microstructure noise and no lagged adjustments are present. Parameters are set as:

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 0.01 \end{pmatrix} \tag{24}$$

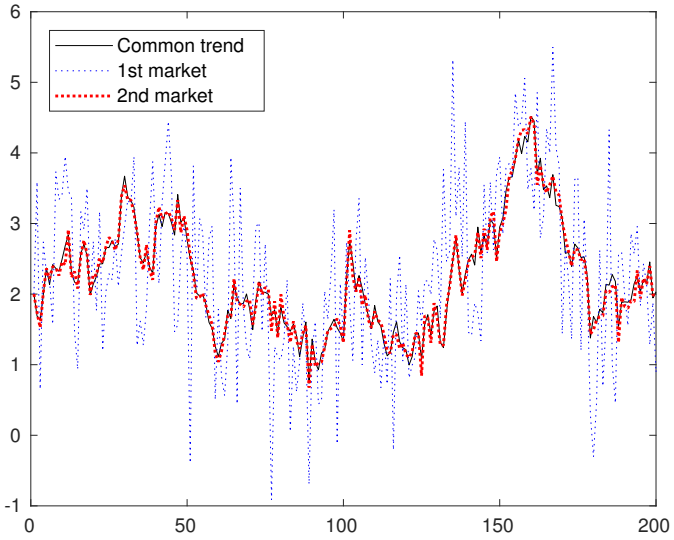

Figure 3: Simulated paths of a bivariate structural model with no lagged adjustment. The price observed in the second market is affected by large measurement errors.

and $\mathbf{\Psi} = \mathbf{I}_2$, $\sigma^2 = 0.1$. The information shares computed through the structural model are $S_1 = 0.009$ and $S_2 = 0.900$, consistently with the fact that the price in the second market is closer to the efficient price process. We then fit a VECM model with one lag. Hasbrouck's information share bounds are $S_1 = [0.010, 0.180]$, $S_2 = [0.819, 0.989]$. Note that bounds are wide even in presence of a diagonal $\mathbf{\Sigma}$. Indeed, the latter represents the measurement noise covariance matrix in the structural model and is not related to the covariance matrix $\mathbf{\Omega}$ of the VECM innovations.

We now consider a case with a lagged adjustment mechanism. The variance of the efficient price innovations is set as before while the noise covariance matrix is chosen as $\mathbf{\Sigma} = 1 \times 10^{-3}\mathbf{I}_2$. This means that the effect of microstructure noise is similar in the two markets and is weak compared to efficient price innovations. The lagged adjustment matrix is chosen as:

$$\mathbf{\Psi} = \begin{pmatrix} 0.3 & 0 \\ 0 & 1 \end{pmatrix} \tag{25}$$

As shown in figure 4, the price in the first market follows the common trend with a lag induced by the coefficient 0.3. The price in the second market has instead no lag, and stays close to the efficient price. The information shares computed through the structural model are $S_1 = 0.001$ and $S_2 = 0.988$. Hasbrouck's information share bounds are instead very wide: $S_1 = [0.001, 0.909]$ and $S_2 = [0.090, 0.998]$ and hardly allow to distinguish the most informative market. This is mainly due to the fact that the estimated covariance $\mathbf{\Omega_{12}} = 0.0288(0.003)$ is statistically different from zero. We conclude that the proposed information share measure consistently accounts for microstructure effects and lagged adjustment, which are the two determinants of market informativeness in our structural model.

4 Conclusions

Hasbrouck (2018) proposes an interesting extension of the traditional VECM approach to deal with data recorded at sub-millisecond frequencies. His approach retains a parsimonious VECM specification in presence of many lags, allowing to model dynamics across a wide range of time-scales. Empirical results show that a clearer description of

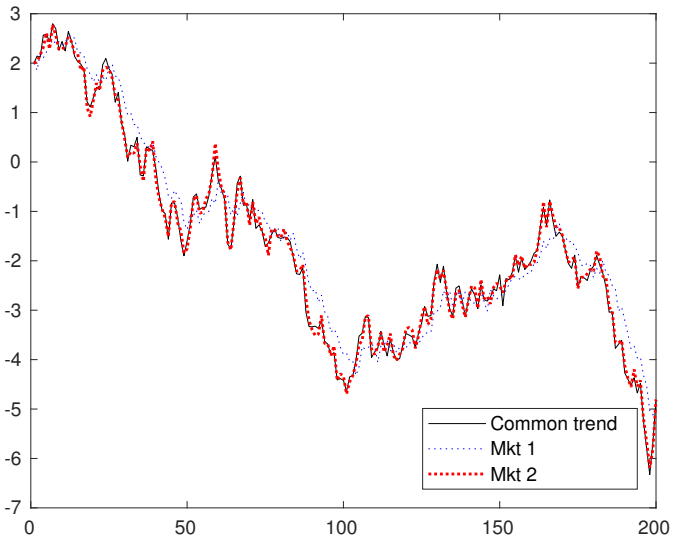

Figure 4: Simulated paths of a bivariate structural model with lagged adjustment. Measurement errors are small in both markets.

the price formation mechanism emerges as the resolution increases. However, the standard information share measure can suffer from potential limitations, especially when many lags are present in the underlying VECM model. We illustrated such limitations and proposed a structural model of price formation characterized by a lagged adjustment mechanism, meaning that markets can have a lag with respect to the efficient price process. This structural model includes microstructure effects and has a linear-Gaussian state-space representation. One of the advantages is that sparse data can be handled as missing values in the state-space representation. Finally, we introduced an information share measure that consistently accounts for both lagged adjustment and microstructure effects. The results obtained in simulations are promising and show the ability of the proposed measure to discover the most informative market.

References

Bandi, F. M., Pirino, D., Reno, R., 2017. EXcess Idle Time. Econometrica 85 (6), 1793–1846.

Buccheri, G., Corsi, F., Peluso, S., 2018a. High-Frequency Lead-Lag Effects and Cross-Asset Linkages: a Multi-Asset Lagged Adjustment Model. Working Paper. Available at https://ssrn.com/abstract=2938619.

Buccheri, G., Livieri, G., Pirino, D., Pollastri, A., 2018b. A closed-formula characterization of the Epps effect. Working Paper.

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

Corsi, F., Peluso, S., Audrino, F., 2015. Missing in Asynchronicity: A Kalman-EM Approach for Multivariate Realized Covariance Estimation. Journal of Applied Econometrics 30 (3), 377–397.

De Jong, F., Schotman, P. C., Winter 2010. Price Discovery in Fragmented Markets. Journal of Financial Econometrics 8 (1), 1–28.

Dias, G., Fernandes, M., Scherrer, C., 2018. Price discovery and market microstructure noise. Working paper, Sao Paulo School of Economics.

Durbin, J., Koopman, S., 2012. Time Series Analysis by State Space Methods: Second Edition. Oxford Statistical Science Series. OUP Oxford.

Epps, T. W., 1979. Comovements in stock prices in the very short run. Journal of the American Statistical Association 74 (366), 291–298.

Harvey, A., 1991. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press.

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. Working paper.

Hasbrouck, J., Ho, T. S. Y., 1987. Order arrival, quote behavior, and the return-generating process. Journal of Finance 42 (4), 1035–48.

Hayashi, T., Yoshida, N., 04 2005. On covariance estimation of non-synchronously observed diffusion processes. Bernoulli 11 (2), 359–379.

Kilian, L., Lütkepohl, H., 2017. Structural Vector Autoregressive Analysis. Cambridge University Press.

Muller, U. A., Dacorogna, M., Dave, R. D., Pictet, O. V., Olsen, R., Ward, J., 1993. Fractals and intrinsic time - a challenge to econometricians. Working Papers 1993-08-16, Olsen and Associates.

Nguenang, C., 2016. Evaluating market's contribution to price discovery in high-frequency for co-listed assets. Working Paper.

Appendix

A State-space representation

Let us introduce the vector $X_t = [P_t', P_{t-1}']' \in \mathbb{R}^{2n}$ and the matrix $\mathbf{Z} = [\mathbf{I}_n, \mathbf{0}_n]'$. We can re-write model (16), (17) as:

$$Y_t = \mathbf{Z}X_t + \epsilon_t, \quad \text{Cov}[\epsilon_t] = \mathbf{\Sigma} \tag{A.1}$$

$$X_{t+1} = \mathbf{T}X_t + \mathbf{R}\eta_{t+1}, \quad \text{Cov}[\eta_t] = \mathbf{Q} \tag{A.2}$$

where

$$\mathbf{T} = \begin{pmatrix} 2\mathbf{I}_n - \mathbf{\Psi} & -\mathbf{I}_n + \mathbf{\Psi} \\ \mathbf{I}_n & \mathbf{0}_n \end{pmatrix}$$

and

$$\mathbf{R} = \begin{pmatrix} \mathbf{\Psi} & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{0}_n \end{pmatrix}, \quad Q = \begin{pmatrix} \sigma^2 \iota_n \iota_n' & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{0}_n \end{pmatrix}$$

Under the additional assumption that ϵ_t, η_t are normal, model (A.1), (A.2) is a linear Gaussian state-space representation and can be estimated as described by Harvey (1991) and Durbin and Koopman (2012).

B Proof of Lemma 1

We first prove that, for $i \geq 1$:

$$P_{t-i} - \iota_n p^*_{t-i} = (\mathbf{I}_n - \mathbf{\Psi})(P_{t-i-1} - \iota_n p^*_{t-i-1} - \iota_n \omega_{t-i}) \tag{A.3}$$

Indeed we have:

$$\begin{aligned} P_{t-i} - \iota_n p^*_{t-i} &= P_{t-i-1} + \mathbf{\Psi}(\iota_n p^*_{t-i} - P_{t-i-1}) - \iota_n p^*_{t-i-1} - \iota_n \omega_{t-i} \\ &= (\mathbf{I}_n - \mathbf{\Psi})P_{t-i-1} + \mathbf{\Psi}\iota_n p^*_{t-1} - \iota_n p^*_{t-i-1} - \iota_n \omega_{t-i} \\ &= (\mathbf{I}_n - \mathbf{\Psi})P_{t-i-1} + \mathbf{\Psi}\iota_n (p^*_{t-i-1} + \omega_{t-i}) - \iota_n p^*_{t-i-1} - \iota_n \omega_{t-i} \\ &= (\mathbf{I}_n - \mathbf{\Psi})(P_{t-i-1} - \iota_n p^*_{t-i-1} - \iota_n \omega_{t-i}) \end{aligned}$$

We now compute V_t:

$$\begin{aligned} V_t &= Y_t - \iota_n p^*_{t-1} \\ &= P_t + \epsilon_t - \iota_n p^*_{t-1} \\ &= P_{t-1} + \mathbf{\Psi}(\iota_n p^*_t - P_{t-1}) - \iota_n p^*_{t-1} + \epsilon_t \\ &= (\mathbf{I}_n - \mathbf{\Psi})P_{t-1} + \mathbf{\Psi}\iota_n (p^*_{t-1} + \omega_t) - \iota_n p^*_{t-1} + \epsilon_t \\ &= (\mathbf{I}_n - \mathbf{\Psi})(P_{t-1} - \iota_n p^*_{t-1}) + \mathbf{\Psi}\iota_n \omega_t + \epsilon_t \end{aligned}$$

Applying now the result in eq. (A.3), we have:

$$\begin{aligned} V_t &= (\mathbf{I}_n - \mathbf{\Psi})[(\mathbf{I}_n - \mathbf{\Psi})(P_{t-2} - \iota_n p^*_{t-2} - \iota_n \omega_{t-1}] + \mathbf{\Psi}\iota_n \omega_t + \epsilon_t \\ &= (\mathbf{I}_n - \mathbf{\Psi})^2 (P_{t-2} - \iota_n p^*_{t-2}) - (\mathbf{I}_n - \mathbf{\Psi})^2 \iota_n \omega_{t-1} + \mathbf{\Psi}\iota_n \omega_t + \epsilon_t \end{aligned}$$

which coincides with the expression in Lemma 1 for $k = 2$. Applying recursively eq. (A.3), one obtains:

$$V_t = (\mathbf{I}_n - \mathbf{\Psi})^k (P_{t-k} - \iota_n p^*_{t-k}) - \sum_{i=2}^{k} (\mathbf{I}_n - \mathbf{\Psi})^i \iota_n \omega_{t-i+1} + \mathbf{\Psi} \iota_n \omega_t + \epsilon_t$$

Provided that all the eigenvalues of $(\mathbf{I}_n - \mathbf{\Psi})$ lie inside the unit circle, the above sum converges in mean square error and in the limit $k \to \infty$ we can write:

$$V_t = -\sum_{i=2}^{\infty} (\mathbf{I}_n - \mathbf{\Psi})^i \iota_n \omega_{t-i+1} + \mathbf{\Psi} \iota_n \omega_t + \epsilon_t$$

Q.E.D.