High-Frequency Lead-Lag Effects and Cross-Asset Linkages: a Multi-Asset Lagged Adjustment Model*

Giuseppe Buccheri†, Fulvio Corsi‡, Stefano Peluso§

First version: March, 2017

This version: July, 2019

Abstract

Motivated by the empirical evidence of high-frequency lead-lag effects and cross-asset linkages, we introduce a multi-asset price formation model which generalizes standard univariate microstructure models of lagged price adjustment. Econometric inference on such model provides: (i) a unified statistical test for the presence of lead-lag correlations in the latent price process and for the existence of a multi-asset price formation mechanism; (ii) separate estimation of contemporaneous and lagged dependencies; (iii) an unbiased estimator of the integrated covariance of the efficient martingale price process that is robust to microstructure noise, asynchronous trading and lead-lag dependencies. Through an extensive simulation study, we compare the proposed estimator to alternative approaches and show its advantages in recovering the true lead-lag structure of the latent price process. Our application to a set of NYSE stocks provides empirical evidence for the existence of a multi-asset price formation mechanism and sheds light on its market microstructure determinants.

JEL codes: C32; C58; G14;

Keywords: price discovery; microstructure noise; asynchronicity; quadratic covariation; Granger causality

*This paper was previously circulated under the title "Hidden Leaders: Identifying High-Frequency Lead-Lag Structures in a Multivariate Price Formation Framework". We are particularly grateful for suggestions we have received from Tim Bollerslev, Giacomo Bormetti, Fabrizio Lillo, Loriana Pelizzon, Roberto Renò and participants to the 10th CFE conference in Seville and the 2017 SoFiE conference in New York.

†Buccheri: Scuola Normale Superiore, giuseppe.buccheri@sns.it (corresponding author)

‡Corsi: Department of Economics, University of Pisa and City University of London, fulvio.corsi@gmail.com

§Peluso: Università Cattolica del Sacro Cuore, stefano.peluso@unicatt.it

1 Introduction

The dynamics of high-frequency asset prices are known to be characterized by "lead-lag effects": some assets (the laggers) tend to follow the movements of other assets (the leaders). This phenomenon is of fundamental interest in market microstructure research and in high-frequency financial econometrics. However, while it has received some attention in the empirical finance literature (Chan 1992, de Jong and Nijman 1997, Chiao et al. 2004, Huth and Abergel 2014, Dobrev and Schaumburg 2017) and in the statistical literature (Hoffmann et al. 2013, Hayashi and Koike 2018, Hayashi and Koike 2017), there is still a lack of econometric approaches aimed to describe lead-lag effects from a market microstructure perspective. On the one hand, there is no well-established microstructure theory explaining the existence of lead-lag effects. On the other hand, compared to the case of low-frequency (e.g. daily) data, the estimation of both contemporaneous and lagged correlations among assets traded at high-frequency is a more complex task. This is mainly due to the presence of asynchronous trading, which prevents the use of traditional multivariate techniques.

Motivated by the empirical evidence of cross-asset trading (Hasbrouck and Seppi 2001, Bernhardt and Taub 2008, Pasquariello and Vega 2015), i.e. the fact that dealers may rely on the prices of other securities when setting their quotes, we introduce a multi-asset price formation mechanism which generalizes well-known univariate microstructure models of lagged price adjustment (see Hasbrouck, 1996 for a review on lagged price adjustment models). The latter were originally introduced in the market microstructure literature to account for the positive autocorrelation observed empirically on univariate time-series of high-frequency returns. Our multivariate generalization, that we name Multi-asset Lagged Adjustment (MLA), leads to a micro-founded model where cross-autocorrelations among high-frequency returns naturally arise as a result of the multivariate nature of the price formation process. Econometric inference on the MLA allows us to test for the presence of lead-lag correlations in the latent price process or, equivalently, for the existence of a non-trivial multi-asset price generation mechanism. Furthermore, by separating the estimation of lead-lag and contemporaneous dependencies, we obtain an estimate of the integrated covariance of the efficient martingale price process that is robust to microstructure noise, asynchronous trading and lead-lag effects. In contrast, the current econometric literature on realized covariance estimation has mainly considered microstructure noise and asynchronous trading, ignoring the existence of lead-lag dependencies (cf. e.g. Hayashi and Yoshida 2005, Corsi et al. 2015 and Shephard and Xiu 2017).

Univariate lagged price adjustment models, also known as partial price adjustment models,

were proposed, among others, by Hasbrouck and Ho (1987), Amihud and Mendelson (1987) and Damodaran (1993). The theoretical concept underlying these models is that prices do not instantaneously adjust when new information arrives. Instead, the adjustment process is delayed because of several market frictions, such as lagged dissemination of information and price smoothing by market dealers. We extend this idea to a multivariate framework by viewing the price formation mechanism as a multi-asset process where the information related to one asset can affect the price discovery of another asset, thus leading to a lag in their dynamics. By doing so, we establish a link between the market microstructure literature on lagged price adjustment and that on cross-asset trading. The concept of cross-asset trading (also known as cross-asset pricing or cross-asset learning) has been extensively exploited by researchers since the seminal work of Caballé and Krishnan (1994), who developed a model of insider trading based on the informational assumption that market makers can learn about one security from observing all order flows in the market. Based on cross-asset learning, Cespa and Focault (2011) developed a transmission mechanism of liquidity shocks among many stocks, the so-called "liquidity spillovers". Pasquariello and Vega (2015) described the relation between cross-price impact and informed multi-asset trading by assuming that dealers in one security can condition on prices of all other securities. Common factors in the price discovery process have been investigated by Hasbrouck and Seppi (2001), Harford and Kaul (2005), Andrade et al. (2008) and Tookes (2008).

Econometric inference on the MLA can be conveniently carried out by casting the model into a state-space representation. The transition equation is a VAR process for the returns of the "adjusted" price, while the observation equation incorporates microstructure effects as an additive noise term. Estimation is performed through a Kalman-EM algorithm, which easily handles missing observations. Asynchronous trading can thus be treated as a typical missing value problem, in a similar fashion to Corsi et al. (2015), Peluso et al. (2015) and Shephard and Xiu (2017). This approach allows to estimate the parameters using all the available prices and avoids the use of standard synchronization schemes. The latter may introduce spurious lead-lag correlations or destroy true short-term lead-lag effects (cf. e.g. Huth and Abergel 2014).

We also contribute to the literature on price discovery. The MLA encompasses a rather interesting case, the one of arbitrage-linked securities. In this particular framework, there are d observed prices pertaining the same security, rather than d different securities. For instance, one can observe the price of the same security in d different exchanges. To guarantee that the dynamics of these securities are cointegrated, the system matrices of the state-space representation assume a

particular structure. Under the cointegration restrictions, the MLA can be employed to investigate in which exchange price discovery occurs. One of the main differences with respect to traditional methods (e.g. the VECM approach of Hasbrouck 1995) is that the recovered lead-lag structure is not affected by differences in the level of trading activity among different exchanges.

In the MLA, there is a one-to-one correspondence between the VAR matrix of lead-lag coefficients and the speed of adjustment matrix in the lagged price adjustment process. The presence of statistically significant lead-lag correlations can thus be interpreted as an evidence of the existence of a multi-asset price formation mechanism. Furthermore, the assets which lead the dynamics of other assets can be regarded as the ones which are faster in incorporating the information. Due to the underlying VAR structure, the MLA can detect "latent" Granger causality, i.e. Granger causality relationships among data recorded asynchronously and with noise. In particular, it allows to disentangle lead-lag correlations arising from nonzero non-diagonal coefficients in the VAR matrix from trivial lead-lag dependencies due to combination of autocorrelation and contemporaneous correlation effects. The latter are not associated to cross-asset pricing in our framework.

The state-space approach to high-frequency covariance estimation was first developed by Corsi et al. (2015), Peluso et al. (2015) and Shephard and Xiu (2017). By modelling microstructure effects as an additive noise term and treating asynchronicity as a missing value problem, they provided a consistent estimator of the quadratic covariation of a Brownian semimartingale observed asynchronously and with noise. The MLA differs in that it introduces a mechanism of price generation that is more realistic than the simple random walk plus noise model. Indeed, while the latter describes the strong negative first-order autocorrelation that is observed in high-frequency returns, the MLA can additionally capture cross-autocorrelations at higher orders.

An extensive simulation study is performed to investigate the statistical properties of the MLA. Our state-space representation assumes that the instantaneous covariance of the efficient price process is constant over time. We show that the MLA is robust under a misspecified DGP with time-varying covariances. This result is similar to that of Shephard and Xiu (2017), who provided the asymptotic theory of the QMLE estimator of the integrated covariance matrix of a Brownian semimartingale observed with noise. We then show that, in contrast to alternative estimators, the MLA is not affected by "spurious" correlations, i.e. nonzero lead-lag correlations arising as a result of asynchronous trading and which are not associated to cross-asset trading. A similar result is found in the case of arbitrage-linked securities.

The MLA is tested on a cross-section of NYSE tick data. We provide empirical evidence for

the existence of a multi-asset price formation mechanism. This is done by recovering the lagged adjustment matrix from the estimated lead-lag correlations and showing that it contains statistical significant non-diagonal elements. The likelihood ratio test shows that the null hypothesis of a random walk plus noise model, along the lines of Corsi et al. (2015) and Shephard and Xiu (2017), is not able to capture some relevant features of the dynamics of high-frequency prices. In particular, deviations from the null assumption of a random walk are more pronounced in periods of large volatility. In a similar fashion to Dobrev and Schaumburg (2017), this empirical finding can be interpreted in light of high-frequency trading: volatility can create short-living cross-autocorrelations that are exploited by short-term, high-frequency trading strategies. A positive relation between volatility and high-frequency trading was found by Zhang (2010), amongst others. We examine in detail the cross-autocorrelation structure of the market and find that, in contrast to what is typically recovered by alternative, non-robust estimators, assets characterized by lower trading activities can be highly informative and can lead the dynamics of more liquid assets. The robustness of the these results is tested with respect to the choice of the assets and to the inclusion of an ETF which is heavily traded in the NYSE.

The rest of the paper is organized as follows: In Section (2) we introduce the model and discuss the estimation strategy based on the EM algorithm; in Section (3) we make a comparison with other estimators and perform extensive Monte-Carlo simulations under a misspecified DGP; in Section (4) we test the MLA on real transaction data and discuss the results; Section (5) concludes. We collect the details of the estimation of the model, the proofs and part of the robustness checks in an online appendix.

2 Theoretical framework

2.1 The Multi-asset Lagged Adjustment model

A significant portion of the empirical research on market microstructure has been devoted to understanding the autocorrelation structure of univariate and multivariate high-frequency returns. There is well-established evidence of three key empirical properties: (i) strong negative first-order autocorrelation, (ii) existence of positive autocorrelation at lags larger than one, (iii) existence of lead-lag correlations. Simple bid-ask models such as the model of Roll (1984) reproduce the negative first-order autocorrelation observed in return series. Univariate bid-ask models were later generalized to capture correlations at orders larger than one through the introduction of lagged

price adjustments (Hasbrouck, 1996). Here, we introduce a multi-asset version of the price adjustment model of Hasbrouck and Ho (1987) which generates lead-lag correlations among different assets.

We assume that the efficient log-price P_t is a d-dimensional vector that evolves as a Brownian semimartingale defined on some filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in[0,T]}, \mathbb{P})$:

$$P_t = \int_0^t \mu_s ds + \int_0^t \sigma dW_s, \quad \Sigma = \sigma\sigma' \tag{1}$$

where $t \in [0, T]$, μ_s is a vector of predictable locally bounded drifts, σ is a volatility matrix and W_s is a vector of independent Brownian motions. The interval $[0, T]$ can be thought of as representing the trading day.

Let $0 \leq t_1, \ldots, t_n \leq T$ denote n equally-spaced observation times. Opposed to P_t, we consider the d-dimensional observed log-price process Y_{t_i}. The difference $\tau = t_{i+1} - t_i$ between consecutive observation times is assumed to be a very short time interval (e.g. $\tau = 1$ second in our empirical application). Note that, because of asynchronous trading, only the components of Y_{t_i} corresponding to traded assets are observed at each time t_i, while the observations of the remaining assets are missing. For simplicity, we assume that the drift term in Eq. (1) is zero[1]. We can write:

$$P_{t_{i+1}} = P_{t_i} + u_{t_i}, \quad u_{t_i} \sim \text{NID}(0, \tau\Sigma) \tag{2}$$

These are the prices that, abstracting from microstructure effects, would be observed in a perfect market, i.e. one in which prices instantaneously react to new information. In real markets, dealers do not instantaneously adjust their quotes to new information. The adjustment process is gradual and reflects lagged dissemination of information and several market imperfections, such as trading costs, discreteness and price smoothing by market makers. In addition, due to cross-asset trading (Hasbrouck and Seppi 2001, Bernhardt and Taub 2008, Pasquariello and Vega 2015), dealers tend to look at more informative securities before setting their quotes. In order to capture lagged dissemination of information *across* stocks, we start from the simple univariate lagged adjustment mechanism proposed by Hasbrouck and Ho (1987) and adapt it to a multi-asset framework.

Let X_{t_i}, $i = 1, \ldots, n$ denote a d-dimensional vector of "adjusted" prices reflecting the imperfections of the trading process. We assume that X_{t_i} is related to the efficient log-price process P_{t_i} by:

$$X_{t_{i+1}} = X_{t_i} + \Psi(P_{t_{i+1}} - X_{t_i}) \tag{3}$$

[1]This assumption is not restrictive, since we are considering ultra-high-frequency returns for which drift effects are negligible.

where Ψ is a $d \times d$ matrix characterizing the speed of adjustment of X_{t_i} to the true efficient log-price P_{t_i}. If $\Psi = \mathbb{I}_d$, then $X_{t_i} = P_{t_i}$ and the adjustment process is instantaneous. Instead, if $\Psi \neq \mathbb{I}_d$ the adjustment process is gradual and, as a result, there is a delay between X_{t_i} and P_{t_i}. Note that the matrix Ψ may be non-diagonal. This implies that the adjustment process of one asset is affected by the adjustment process of other assets, and the strength of this effect is quantified by the non-diagonal elements of Ψ.

Due to the presence of market microstructure effects (e.g. bid-ask bounces), the observed log-price process Y_{t_i} deviates from the lagged price X_{t_i}. Therefore, we assume that X_{t_i} is observed under additive noise:

$$Y_{t_i} = X_{t_i} + \epsilon_{t_i}, \quad \epsilon_{t_i} \sim \text{NID}(0, H) \tag{4}$$

where ϵ_{t_i} is a normal white noise term summarizing microstructure effects. In line with Corsi et al. (2015) and Shephard and Xiu (2017), the noise covariance matrix H is assumed to be diagonal. Denoting by $\Delta X_{t_{i+1}} = X_{t_{i+1}} - X_{t_i}$ the log-returns of the lagged price, Eq. (2) and (3) imply:

$$\Delta X_{t_{i+1}} = (\mathbb{I}_d - \Psi)\Delta X_{t_i} + \Psi u_{t_i} \tag{5}$$

that is, a first order vector autoregressive VAR(1) process. If Ψ is non-diagonal, the knowledge at time t_i of the return of one asset is useful for forecasting the return of another asset at time t_{i+1}. In this multi-asset framework, lead-lag correlations naturally arise as a consequence of cross-asset trading and of the mutual influence between adjustment processes of different assets. In particular, if one asset leads the dynamics of the other assets, it can be regarded as the one which incorporates the information with highest speed.

Let us assume, without loss of generality, that $\tau = 1$ and re-write Eq. (4) and (5) as:

$$Y_t = X_t + \epsilon_t, \quad \epsilon_t \sim \text{NID}(0, H) \tag{6}$$

$$\Delta X_{t+1} = F\Delta X_t + \eta_t, \quad \eta_t \sim \text{NID}(0, Q) \tag{7}$$

where $F = \mathbb{I}_d - \Psi$ and $Q = \Psi\Sigma\Psi'$. Eq. (6) is a measurement equation expressing the fact that observations of latent prices are affected by noise. Eq. (7) is instead a transition equation describing the dynamics of latent returns. We name the model described by Eq. (6), (7) "Multi-Asset Lagged Adjustment" (MLA) model. As discussed in Section (2.2), the MLA has a linear-Gaussian state-space representation and can conveniently be estimated through a standard Kalman-EM algorithm.

The assumption of a constant instantaneous covariance matrix Σ in the efficient log-price process may be regarded as too restrictive, since there is well-established evidence that both volatilities

and correlations exhibit strong intraday variation (cf. e.g Andersen and Bollerslev, 1997, Tsay, 2005, Bibinger et al., 2014, Buccheri et al., 2019a). However, by performing extensive Monte-Carlo simulations on a misspecified DGP with a time-varying covariance matrix, we show in Section (3.2) that two relevant properties hold. First, the maximum-likelihood estimate $\hat{F}$ of the VAR(1) matrix of lead-lag coefficients remains unbiased even in presence of time-varying covariances. Second, denoting by $\hat{\Psi}$ and $\hat{Q}$ the two maximum-likelihood estimates of Ψ and Q, the matrix $\hat{\Sigma} = \hat{\Psi}^{-1}\hat{Q}\hat{\Psi}'^{-1}$ is an unbiased estimator of $\frac{1}{T}QV$, the quadratic covariation of the efficient log-price process:

$$QV = \int_0^T \Sigma_s ds \tag{8}$$

This result is similar to that obtained by Shephard and Xiu (2017), who derived the asymptotic theory of the QMLE of the integrated covariance of a Brownian semimartingale process observed with noise. Thus, the MLA provides, as a byproduct, a robust estimator of the quadratic covariation of a Brownian semimartingale process in the presence of lead-lag dependencies.

We finally note that inference on the MLA can be regarded as testing for one-lag Granger causality in the latent return process ΔX_t. Testing for higher order lags would result in additional $d \times d$ lead-lag matrices to be estimated, thus increasing considerably the dimensionality of the parameter space. In practice, a less efficient but more feasible solution is to consider observations sampled at a smaller frequency, which avoids estimating complex higher order VAR(p) models.

A particularly interesting case is when all or some of the assets are linked by non-arbitrage relationships. For instance, one can observe the price of the same security in different exchanges, or the price of a stock and that of a call option on the same stock. These securities are driven by the same efficient martingale process, and are therefore cointegrated. The system matrices in the MLA can be constrained in such a way to guarantee cointegration among the arbitrage-linked securities. In particular, the covariance matrix Q turns out to have a rank equal to the number of efficient martingale processes. For instance, if there are d observed prices pertaining the same security, the rank of Q is one, regardless the value of d. The latter case was discussed by Buccheri et al. (2019b). In the online appendix we provide the details of the model and generalize to the case of k securities, each one observed in d_i markets, with $i = 1, \ldots, k$. One of the main advantages of using the MLA in this setting is that parameter estimates are not affected by differences in the level of trading activity among different markets. For instance, the MLA can detect that a market is more informative even if its trading activity is lower than in other markets. We illustrate several examples in the online appendix.

2.2 Estimation

The MLA can be conveniently estimated by writing Eq. (6), (7) in a linear Gaussian state-space representation. This is possible if one introduces the $2d$-dimensional state vector $Z_t = [X_t', X_{t-1}']'$ and re-writes the two equations as:

$$Y_t = MZ_t + \epsilon_t, \quad \epsilon_t \sim \text{NID}(0, H) \tag{9}$$

$$Z_t = \Phi Z_{t-1} + \xi_t, \quad \xi_t \sim \text{NID}(0, W) \tag{10}$$

where:

$$\Phi = \begin{pmatrix} \mathbb{I}_d + F & -F \\ \mathbb{I}_d & 0_d \end{pmatrix}, \quad W = \begin{pmatrix} Q & 0_d \\ 0_d & 0_d \end{pmatrix} \tag{11}$$

with $M = [\mathbb{I}_d, 0_d]$ being a matrix that selects the first d components of Z_t and 0_d denoting a $d \times d$ matrix of zeros.

Model (9), (10) is a linear Gaussian state-space representation for which the Kalman filter can be applied and the log-likelihood function can be written down in the form of the prediction error decomposition (cf. e.g. Durbin and Koopman 2012 and Shumway and Stoffer 2015). In particular, we use the EM algorithm of A. P. Dempster (1977) to maximize the log-likelihood. All the details of the estimation procedure are reported in the online appendix. Here, we only report the update formulas for the system matrices in the EM algorithm. Let:

$$A = \sum_{t=1}^{n} (P_{t-1}^n + Z_{t-1}^n Z_{t-1}^{n\prime}) \tag{12}$$

$$B = \sum_{t=1}^{n} (P_{t,t-1}^n + Z_t^n Z_{t-1}^{n\prime}) \tag{13}$$

$$C = \sum_{t=1}^{n} (P_t^n + Z_t^n Z_t^{n\prime}). \tag{14}$$

where Z_t^n, P_{t-1}^n, $P_{t,t-1}^n$ are smoothed conditional mean and covariances of the latent state. Let us write the matrices A and B in the following form:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \tag{15}$$

where A_{ij} and B_{ij}, $i = 1, 2$ are $d \times d$ submatrices of A and B. In the online appendix, we prove the following result:

Proposition 1. *At the r-th step of the EM algorithm, the update formula for F is given by:*

$$\hat{F}_r = \Gamma \Theta^{-1} \tag{16}$$

where $\Gamma = B_{11} - B_{12} - A_{11} + A_{12}$ *and* $\Theta = A_{11} + A_{22} - A_{12} - A_{21}$. *The update formula for* Q *and* H *are instead given by:*

$$\hat{Q}_r = \frac{\hat{\Upsilon}}{n}, \quad \hat{H}_r = \frac{\text{diag}(\Lambda)}{n} \tag{17}$$

where $\hat{\Upsilon} = M(C - B\hat{\Phi}_r' - \hat{\Phi}_r B' + \hat{\Phi}_r A \hat{\Phi}_r')M'$, $\Lambda = \sum_{t=1}^{n}[(Y_t - MZ_t^n)(Y_t - MZ_t^n)' + MP_t^n M']$ *and* $\hat{\Phi}_r$ *is constructed using* $\hat{F}_r$.

We remark that we implicitly assume the constancy of the elements of Ψ within the period analyzed. Therefore, the user of the proposed method should be aware of the potential source of distortion in case the coefficients of Ψ vary over time (e.g. due to changes in market volatility and liquidity conditions). In Section (C.3) in the online appendix we estimate the MLA in several sub-periods of the trading day and show that the estimates of Ψ in these sub-periods do not differ substantially from that obtained in the entire day. Thus, as long as the MLA is re-estimated on a daily basis, the assumption of constancy of Ψ is not restrictive.

The choice of the sample frequency has an impact on the memory and computational power requirements for the estimation of the MLA. This choice depends on the specific nature of the market under study, in particular by market liquidity and information delay. Very liquid markets suggest the adoption of a higher frequency at the cost of a heavier computation burden. This higher cost can be mitigated reducing the within-day window of transactions or quotes analyzed. On the other hand, more illiquid markets lead to lower frequencies, since there is no additional information coming from a higher resolution of prices. Also, increasing the frequency in illiquid markets causes an increase in the proportion of missing values: even if the MLA can deal with very high proportions of missing values, still this is a factor to be accounted for since it renders the estimation process more difficult and therefore expects a higher number of iterations before convergence. Finally, markets where information spread faster across assets need a higher frequency of observations, since all delays that are exhausted at the rate higher than the observed frequency cannot be captured by the model. For further details on the choice of the sampling frequency we refer the reader to Section (C.2) in the online appendix.

3 Simulation study

3.1 Comparison with other estimators

As underlined in the introduction, non-synchronous trading can have deep consequences when analyzing multivariate high-frequency tick-by-tick data. One example is given by the Epps effect, i.e. the bias towards zero of sample correlations observed when shrinking the sampling frequency to zero (cf. e.g. Epps, 1979 and Hayashi and Yoshida, 2005).

Non-synchronous trading is responsible for two main types of "spurious" lead-lag correlations when employing standard estimators. First, some assets seem to lead other assets because they are traded more frequently. This effect is due to differences in the level of trading activity and is not necessarily related to cross-asset pricing. For instance, if we set $\Psi = \mathbb{I}$ and we assume that one asset is traded more frequently than the others, based on standard sample correlations we would conclude that the more frequently traded asset is the most informative. In contrast, the MLA would indicate that the level of informativeness of all the assets is the same. Second, as we will show below, even in presence of similar levels of trading activities (and irregularly spaced observations), one can find spurious nonzero lead-lag correlations that are not related to true lead-lag dependencies. Another source of spurious lead-lag correlations can simply arise as a result of the combination of autocorrelation and contemporaneous correlations. Even these cross-autocorrelations are not related to nonzero non-diagonal elements in the matrix Ψ.

A detailed analysis of the impact of asynchronicity on lead-lag correlations was performed by Huth and Abergel (2014), who considered the standard previous-tick correlation estimator (Griffin and Oomen, 2011) and the estimator proposed by Hoffmann et al. (2013). The latter is computed on bivariate time-series by applying the Hayashi-Yoshida (HY) estimator (Hayashi and Yoshida, 2005) after shifting the timestamps of one of the two time-series. In their simulation study, Huth and Abergel generated two (contemporaneously) correlated Brownian motions with different timestamps. Compared to the previous-tick correlation estimator, the HY estimator is not affected by differences in the levels of trading activity, meaning that the cross-autocorrelation function remains symmetric even if the two processes are characterized by different average durations. However, as a consequence of asynchronicity, it has a bias at nonzero leads and lags that implies nonzero correlations in absence of true lead-lag dependencies.

The MLA is robust to all these effects. The reason is that asynchronicity is handled as resulting from missing observations, which are incorporated in the EM algorithm without jeopardizing the

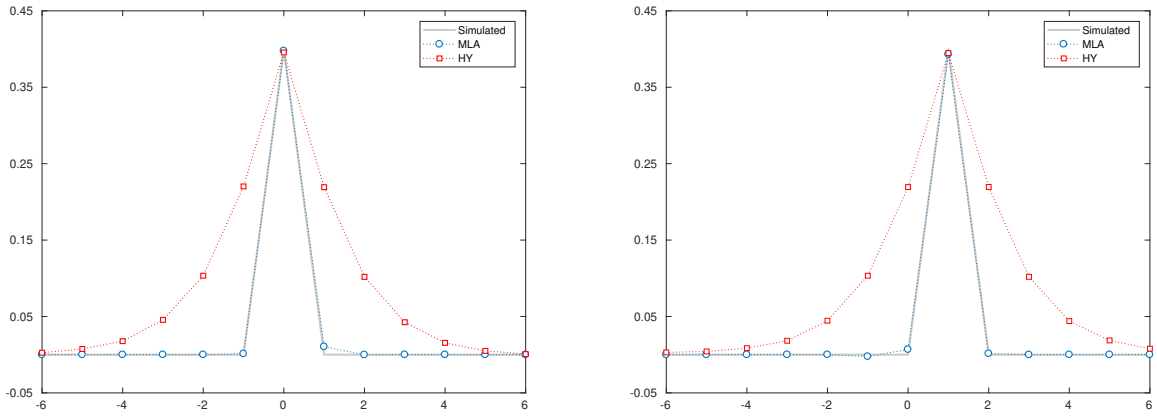


Figure 1: Left: cross-correlogram of two simulated Brownian motions with correlation $\rho = 0.4$ observed asynchronously over $T = 10000$ timestamps. The red line is obtained by averaging across HY estimates, while the blue line is obtained by averaging MLA estimates over $N = 250$ independent realizations. Right: as before but HY and MLA are estimated over a time-series obtained by shifting one of the two simulated Brownian motions.

inference. In order to show this property, we sample two Brownian motions over a time grid of $T = 10000$ equally spaced points. The correlation between the two Brownian motions is $\rho = 0.4$. Asynchronicity is reproduced by censoring the simulated observations through Poisson sampling. The probability of missing values is set equal to $\Lambda_1 = 0.3$ for the first time-series and $\Lambda_2 = 0.5$ for the second time-series. We repeat the experiment 250 times and, for each realization, we compute the lead-lag correlations using both the MLA and the HY estimator. Figure (1) shows on the left the cross-autocorrelation function obtained by averaging the lead-lag correlations over all the simulations. We note that both correlograms are symmetric, meaning that both estimators are not affected by differences in the level of trading activity. However, the HY estimator provides nonzero correlations at nonzero lags. As shown by Huth and Abergel (2014) (appendix B), this is due to observations not being synchronized. In contrast, the MLA is not affected by asynchronicity and correctly reproduces the cross-autocorrelation of the simulated data.

Our definition of lead-lag effects is formally different from that of Hoffmann et al. In the MLA lead-lag effects arise as a consequence of nonzero non-diagonal coefficients in the lagged adjustment matrix Ψ. In their work, Hoffman et al. considered a continuous-time bivariate process (X_t, Y_t) and focused on the estimation of the time shift θ such that the shifted process $(X_t, Y_{t+\theta})$ is a semi-martingale with respect to some filtration. In order to understand how the MLA behaves in

	avg×100	stDev×100	p-value	avg×100	stDev×100	p-value	avg×100	stDev×100	p-value
	$\delta = 0.5$			$\delta = 1$			$\delta = 2$		
	$\Lambda = 0$								
F_{11}	0.1832	2.6183	**0.0182**	0.0733	3.0965	**0.4578**	0.4132	3.3118	**0.5325**
F_{12}	−0.2249	2.5799	0.0041	−0.0844	3.2785	**0.4026**	−0.0618	3.4438	**0.5712**
F_{21}	0.0827	2.2626	**0.2592**	0.0156	2.8687	**0.8222**	0.0004	3.2345	**0.9989**
F_{22}	−0.1723	2.9102	**0.0602**	−0.1245	3.6656	**0.2887**	0.0080	3.8451	**0.9682**
Σ_{11}	−0.0022	0.0862	**0.0549**	−0.0010	0.0836	**0.4872**	−0.0037	0.0886	**0.1293**
Σ_{12}	−0.0039	0.0422	0.0011	−0.0038	0.0523	0.0052	−0.0041	0.0510	**0.0673**
Σ_{22}	0.0077	0.1375	**0.0527**	−0.0023	0.1429	**0.4230**	−0.0051	0.1311	**0.2840**
	$\Lambda = 0.5$								
F_{11}	−0.1693	3.3167	**0.0934**	0.2427	3.7529	**0.0451**	0.1361	3.2374	**0.1790**
F_{12}	0.0187	3.4766	**0.8621**	−0.3028	3.8709	**0.0155**	−0.2104	3.3821	**0.0434**
F_{21}	−0.9021	5.2328	0.0006	−0.4025	3.9742	0.0083	−0.2758	3.3078	**0.0125**
F_{22}	0.8625	5.7758	0.0002	0.1711	4.7587	**0.2589**	0.1922	3.9597	**0.1225**
Σ_{11}	0.0052	0.0968	**0.0246**	−0.0011	0.0938	**0.6278**	0.0014	0.0932	**0.6685**
Σ_{12}	0.0011	0.0659	**0.1822**	0.0009	0.0642	**0.5866**	−0.0004	0.0556	**0.8048**
Σ_{22}	−0.0052	0.1460	**0.1702**	0.0021	0.1556	**0.5512**	9.75e-6	0.1470	**0.9989**

Table 1: We report the sample averages and standard deviations of all the elements of the pivotal matrices $\hat{\theta}_F$, $\hat{\theta}_\Sigma$ and the p-value of the one-sample t-test in all the simulated scenarios. The cases in which the null hypothesis is not rejected at the 1% c.l. are denoted by bold numbers.

this different framework, we consider the bivariate time-series of the previous experiment and shift by a lag $\theta = 1$ all the timestamps of one of the two time-series. Figure (1) shows on the right the correlogram of the new bivariate time-series and those estimated by the HY and MLA. The HY estimator correctly estimates the lagged cross correlation but provides nonzero correlations at other leads and lags. The MLA correctly captures the true cross-correlogram. Indeed, the shifted time series can be written as a VAR(1) process with a nonzero non-diagonal element in F and uncorrelated disturbances. In the case $\theta > 1$, one can sample observations at a lower frequency and still use the MLA estimator.

As a final remark, we note that the MLA estimator is robust to measurement errors, since it allows observations of the underlying process X_t to be contaminated by noise. Another relevant advantage is that the proposed MLA estimator is not pairwise, i.e. it can be applied to a generic multivariate time-series of dimension $d \geq 2$.

3.2 Robustness to stochastic volatility

In Eq. (1) the covariance matrix Σ of the efficient log-price is assumed to be constant over time. This assumption is too restrictive, since real high-frequency data are characterized by significant changes in their covariance structure during the day. In order to assess the properties of $\hat{F}$ and $\hat{\Sigma}$ in a more realistic scenario, we simulate realizations from a misspecified DGP with a time-varying covariance matrix Σ_t. The latter is decomposed as:

$$\Sigma_t = D_t R D_t \tag{18}$$

where R is a constant correlation matrix and D_t is a diagonal matrix of time-varying standard deviations:

$$D_t = \begin{pmatrix} \sigma_{t,1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{t,d} \end{pmatrix} \tag{19}$$

The dynamic terms $\sigma_{t,i}$, $i = 1, \ldots, d$ are decomposed as:

$$\sigma_{t,i} = d_t g_{t,i} \tag{20}$$

where d_t is a common diurnal pattern given by:

$$d_t = C + Ae^{-at} + Be^{-b(1-t)} \tag{21}$$

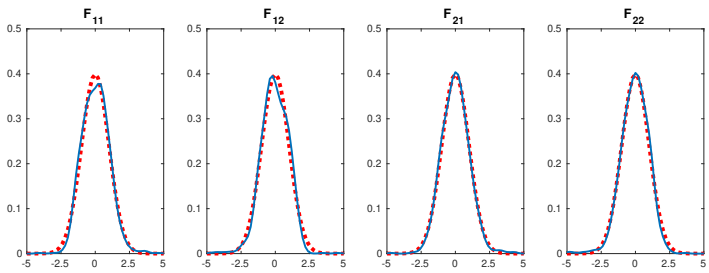
and $g_{t,i}$ evolve through the following stochastic volatility model:

$$dg_{t,i}^2 = k_i(v_i - g_{t,i}^2)dt + w_i g_{t,i} dB_{t,i} \quad i = 1, \ldots, d \tag{22}$$

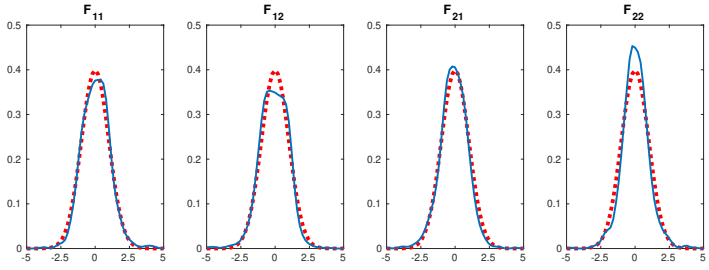
where $B_{t,i}$ are independent Wiener processes and $\mathbb{E}[dB_{t,i}dW_{t,j}] = \rho_{ij}dt$. Note that the diurnal pattern is the same as in Bollerslev et al. (2016) and the parameters $A = 0.75$, $B = 0.25$, $C = 0.88929198$ are set in such a way that $1/T(\int_0^T d_t^2 dt) = 1$.

We report the results obtained in the bivariate case ($d = 2$). The length of the trading day is assumed to be $T = 6.5$ hours and thus we consider $n = 23400$ timestamps of one second. The stochastic volatility process is discretized in the Euler scheme. The parameters in Eq. (22) are set as: $v_1 = 0.01$, $v_2 = 0.02$, $w_1 = w_2 = 0.1$, $k_1 = 10$, $k_2 = 7$. We adopt the following choices for the matrix F of lead-lag coefficients and the leverage matrix ρ:

$$F = \begin{pmatrix} 0.1 & 0.5 \\ 0.3 & 0.1 \end{pmatrix}, \quad \rho = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.1 \end{pmatrix} \tag{23}$$



(a)



(b)

Figure 2: Kernel density estimates of the elements of the matrix $\hat{\theta}_F$ standardized by their sample standard deviations in the two scenarios $\delta = 1, \Lambda = 0$ (a), $\delta = 1, \Lambda = 0.5$ (b) over $N = 1000$ independent realizations. The red line is the standard normal density.

The diagonal elements of the variance matrix H, that we denote as h_{ii}, $i = 1, 2$, are computed based on the average signal-to-noise ratio, that is defined as $\delta_i = v_i/h_{ii}$.

For simplicity, we assume $\delta_1 = \delta_2$. In order to mimic realistic noise scenarios, we set $\delta_i = 0.5, 1, 2$. Indeed, as shown in Table (2), these values are close to those estimated on real data. We consider both the case where observations are synchronized and the case where there are missing values. In the latter case, the simulated observations are censored using Poisson sampling. The probability of missing values is set equal to $\Lambda = 0.5$ for both series.

We estimate the MLA for $N = 1000$ independent realizations and consider the pivotal statistics $\hat{\theta}^i_F = \hat{F}^i - F$ and $\hat{\theta}^i_\Sigma = \hat{\Sigma}^i - \frac{1}{T}QV$, for $i = 1, \ldots, N$. The term QV is the quadratic covariation defined in Eq. (8). The three scenarios $\delta = 0.5, 1, 2$ are combined to the two scenarios $\Lambda = 0, 0.5$, obtaining a total of 6 scenarios. In Table (1) we show the sample mean and standard deviations of each entries of two matrices $\hat{\theta}^i_\Sigma$ and $\hat{\theta}^i_F$. A one sample t-test is performed in order to test the null assumption that the mean is zero. The p-values are reported in Table (1). The distribution of $\hat{\theta}_\Sigma$ and $\hat{\theta}_F$ is always centered in zero. This implies that, even in case of time-varying covariances, $\hat{F}$ is an unbiased estimate of the true matrix F of lead-lag coefficients, while $\hat{\Sigma}$ is an unbiased estimate

of the quadratic covariation of the efficient martingale process. This result is similar to that of Shephard and Xiu (2017), who proved the consistency of the QMLE estimator for the quadratic covariation of a Brownian semimartingale observed with noise.

In Figures (2), (3) the two scenarios $\delta = 1, \Lambda = 0$ and $\delta = 1, \Lambda = 0.5$ are considered. We plot kernel density estimates of each element of $\hat{\theta}_\Sigma$ and $\hat{\theta}_F$ after normalizing by their sample standard deviations. In the first scenario there are no missing values and the densities are perfectly compatible with a standard normal. In the second scenario the density is still centered in zero but, as a consequence of data reduction, we observe slight deviations from the normal.

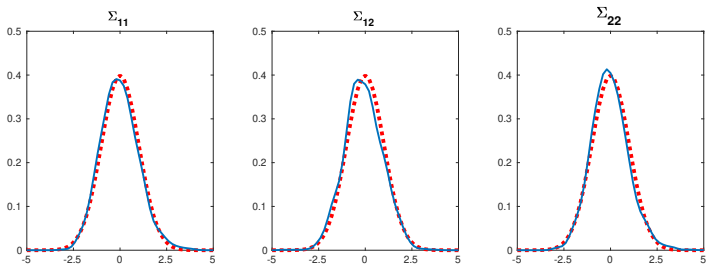
In order to assess the effect of lead-lag correlations on commonly used realized covariance estimators, we plot in Figure (4) the kernel density estimates obtained using the HY estimator. We set $\delta = 10^3$ to exclude that potential biases are due to the measurement noise. As can be seen, the Hayashi-Yoshida estimator is largely biased in case lead-lag correlations are present. A similar bias[2] is observed when using other estimators of the quadratic covariation, e.g. the pairwise estimator of Aït-Sahalia et al. (2010), the multivariate realized kernel of Barndorff-Nielsen et al. (2011), and the QML estimator of Shephard and Xiu (2017). Other properties of the MLA, including its behavior in presence of arbitrage-linked securities are examined in the online appendix.

4 Empirical evidence

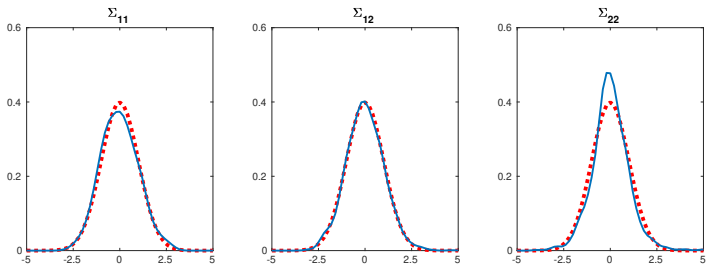
4.1 Dataset

In this empirical application we use two different datasets. The first dataset is provided by Thomson Reuters and includes intraday transaction data of 10 among the most frequently traded NYSE assets in the period between 03-01-2006 and 31-12-2014, a total of 2250 business days. The timestamp precision is the second. We use the procedure described by Barndorff-Nielsen et al. (2009) to clean the data. In particular: (i) we consider trades in the time window from 9:30 to 16:00; (ii) we aggregate high-frequency prices at the one second frequency by taking the median of trades within the same second. Table (2) shows the 10 assets together with several summary statistics related to their trading activity in the period 2006-2014. Note that the probability of missing values Λ ranges from 70% to 80%, indicating that even liquid assets are characterized by large levels of sparsity. We also report the market betas and the average signal-to-noise ratio estimated by the MLA. The latter is defined, for the i-th asset, as the ratio Q_{ii}/H_{ii}. Note that five of the selected assets, namely

[2]Results are available upon request to the authors.



(a)



(b)

Figure 3: Kernel density estimates of the elements of the matrix $\hat{\theta}_\Sigma$ standardized by their sample standard deviations in the two scenarios $\delta = 1, \Lambda = 0$ (a), $\delta = 1, \Lambda = 0.5$ (b) over $N = 1000$ independent realizations. The red line is the standard normal density.

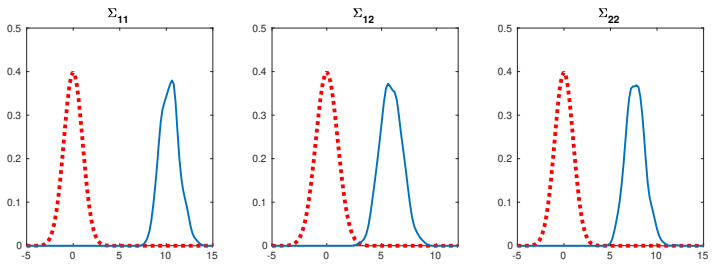


Figure 4: Kernel density estimates of the elements of the matrix $\hat{\theta}_\Sigma$ standardized by their sample standard deviations and computed using the HY estimator in the scenario $\delta = 10^3, \Lambda = 0$ over $N = 1000$ independent realizations. The red line is the standard normal density.

Stock	Symbol	Λ	$\overline{\Delta t}$	$\overline{n}$	$n_{\max}$	$n_{\min}$	$\overline{\delta}$	β	$\overline{RV}$
Exxon	XOM	0.704	3.378	6908	16896	1760	1.178	1.05	1.965
JPMorgan	JPM	0.737	3.802	6142	18248	1324	0.999	1.17	5.233
Citigroup	C	0.753	4.048	5764	17988	1148	1.246	1.84	9.636
Schlumberger	SLB	0.761	4.184	5590	15161	355	0.613	1.66	4.656
Chevron	CVX	0.773	4.405	5308	12782	1281	0.850	0.75	2.258
Bank of America	BAC	0.778	4.504	5193	18839	570	0.328	1.56	7.889
General Electric	GE	0.798	4.950	4698	17717	1028	0.641	0.92	3.287
ConocoPhillips	COP	0.799	4.975	4953	14560	935	0.494	0.66	2.836
Goldman Sachs	GS	0.803	5.076	4598	17120	551	0.630	1.22	4.945
Morgan Stanley	MS	0.808	5.208	4481	14965	611	0.741	1.15	11.106
S&P 500 ETF	SPY	0.275	1.382	16956	20949	10304	2.458	–	0.3759

Table 2: For each asset we show the probability of missing values Λ, the average duration $\overline{\Delta t}$ in seconds between consecutive observations, the average ($\bar{n}$), maximum ($n_{\max}$) and minimum ($n_{\min}$) number of observations per day. We also report the average signal-to-noise ratio $\overline{\delta}$ estimated by the MLA, the market betas and the average (percentage) realized variance. The summary statistics of the 10 NYSE assets are referred to the period from 03-01-2006 to 31-12-2014 whereas those of the SPY are refereed to the period from 03-01-2012 to 28-12-2012.

{C, JPM, BAC, MS, GS} belong to the banking sector, while the remaining five, namely {XOM, CVX, SLB, COP, GE} belong to the oil and energy sectors. We name the two sets of assets as "Group I" and "Group II", respectively.

The second dataset includes transaction data of SPDR S&P 500 ETF (SPY) from 03-01-2012 to 28-12-2012. The timestamp precision is the second and thus we employ the same cleaning procedure adopted for the NYSE dataset. The SPY is traded in the NYSE Arca but is much more liquid than individual stocks, as shown by the summary statistics in Table (2). We use SPY data in Section (4.3) to assess the effect of the inclusion of such heavily traded security in the MLA.

All the analyses in the following sections are performed by estimating the MLA at the sampling frequency of one second. As the resolution decreases, lead-lag cross-autocorrelations tend to die out, as it is shown in Section (C.2) of the online appendix. We thus exploit the largest information set available in our dataset to gain better insights on the nature of cross-asset trading.

4.2 Lead-lag effects and cross-asset trading

Our primary goal is to assess the statistical significance of the multi-asset lagged adjustment mechanism introduced in Section (2.1). To this end, we compare the MLA to a standard random walk plus noise model, also known as local level (LL) model (cf. e.g. Durbin and Koopman 2012). The LL is nested into the MLA since it is obtained by setting to zero the VAR matrix F in Eq. (7). The resulting model can be estimated through the EM algorithm with missing observations, as described by Corsi et al. (2015) and Shephard and Xiu (2017).

In order to compare the two models, for the 10 NYSE assets considered in Table (2) we estimate the LL and the MLA on each day of the sample and compute the likelihood ratio $\lambda = \frac{\mathcal{L}^{\text{LL}}}{\mathcal{L}^{\text{MLA}}}$. Figure (5) reports the test statistic $-2\log(\lambda)$, which is distributed according to a χ^2 with a number of degrees of freedom equal to d^2, coinciding with the difference between the number of parameters in the MLA and that in the LL model. The null hypothesis that the matrix of lead-lag coefficients F is zero is strongly rejected in most of the days of the sample. This indicates that the standard LL specification is not sufficient to capture some relevant features of the dynamics of high-frequency prices. In particular, the fact that F is nonzero implies that the lagged adjustment matrix $\Psi = \mathbb{I}_d - F$ differs from the identity and that the price adjustment process is delayed as a result of lagged dissemination of information across stocks.

The deviation of the MLA specification from the null assumption of a simple random walk process is larger in periods of high volatility. Figure (6) shows the logarithm of RV_{avg}. For each

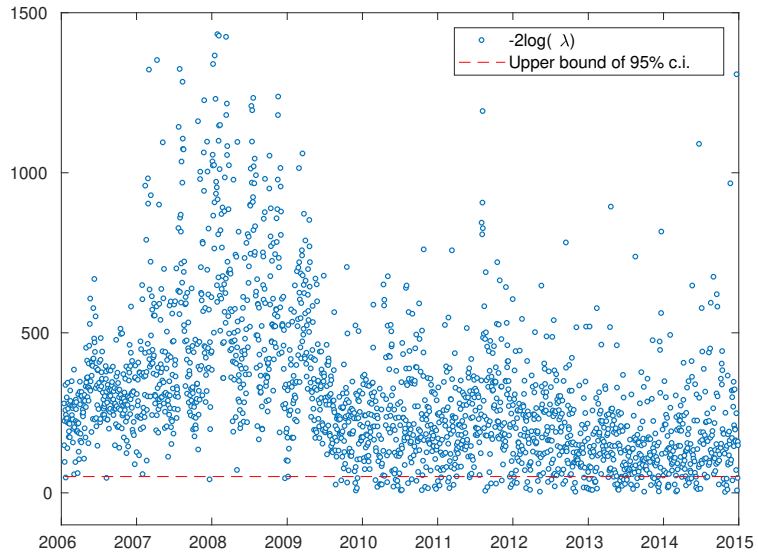


Figure 5: We show the test statistic $-2\log(\lambda)$ computed for each day of the sample and the upper bound of the 95% confidence interval evaluated using a χ^2 distribution with a number of degrees of freedom equal to d^2.

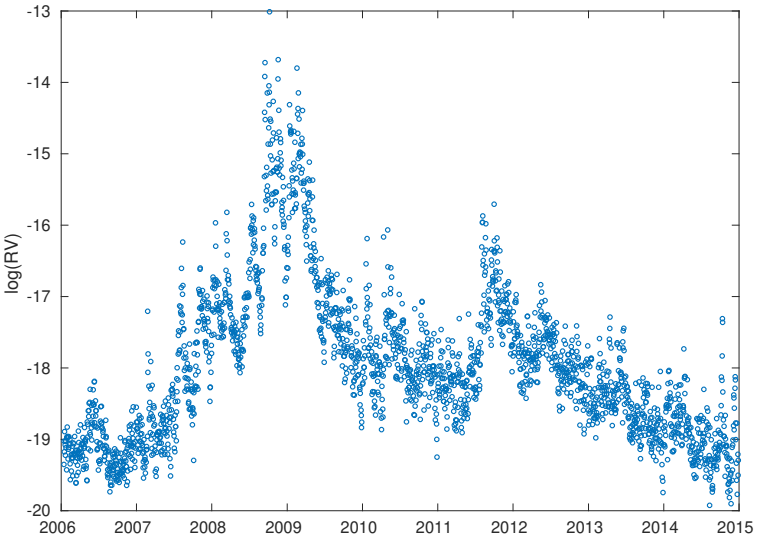


Figure 6: For each day of the sample, we show the logarithm of RV_{avg}. The latter is computed as the average realized variance of the 11 assets considered in Table (2).

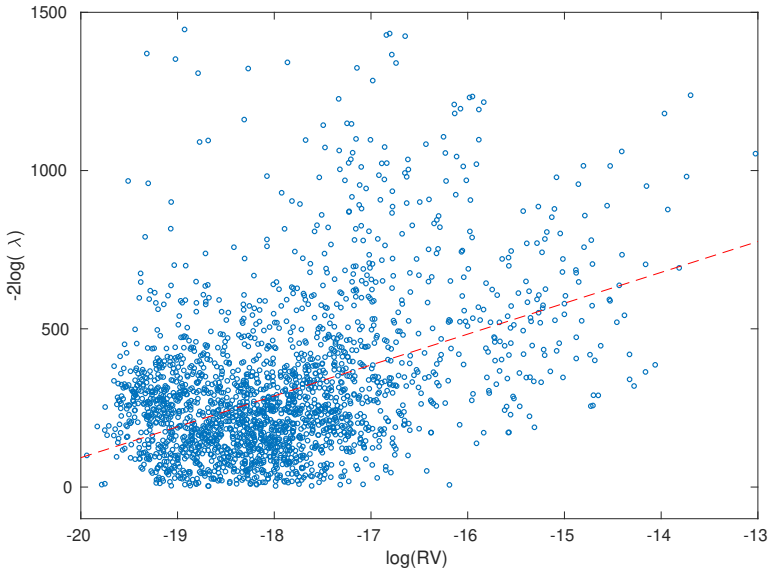


Figure 7: Scatter plot of $-2\log(\lambda)$ versus $\log(RV_{\text{avg}})$, where RV_{avg} is computed as the average realized variance of the assets. The correlation is $\rho = 0.4296$. We also show the result of the OLS regression of $-2\log(\lambda)$ on $\log(RV_{\text{avg}})$

day of the sample, the latter is computed as the average realized variance of the 10 assets considered in Table (2). Large values of $-2\log(\lambda)$ correspond to bursts in volatility. This is confirmed by Figure (7), which shows a scatter plot of the two quantities and the line obtained from the OLS regression. The correlation between $-2\log(\lambda)$ and $\log(RV_{\text{avg}})$ is 0.4296 ± 0.0871 and it turns out to be highly significant.

The relation between cross-asset effects and volatility can be ascribed to the impact of high-frequency trading (HFT). HFT is typically based on short-term statistical dependencies among assets. In periods of high uncertainty, volatility can create short-living cross-autocorrelations that are exploited by high-frequency traders. A positive relationship between HFT and volatility, especially in periods of large market uncertainty, was found by Zhang (2010). A similar result was recovered by Dobrev and Schaumburg (2017), who revealed a close relationship between volatility and lead-lag effects among different markets as a consequence of HFT surges in cross-market activity.

A more detailed description of the relation between market volatility and cross-asset effects can be obtained by looking at the estimated lead-lag correlations. Figure (8) shows the daily dynamics of one second lead-lag correlations between Bank of America and Citigroup. Lead-lag correlations

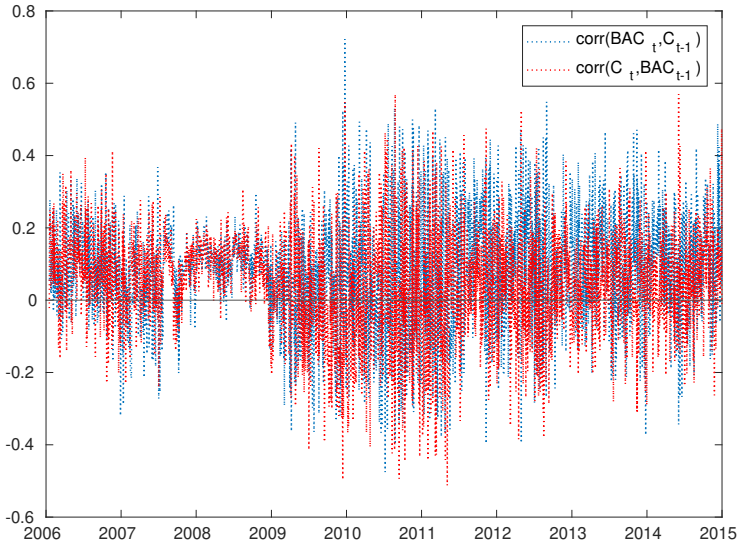


Figure 8: Dynamics of one second lead-lag correlations between Bank of America and Citigroup.

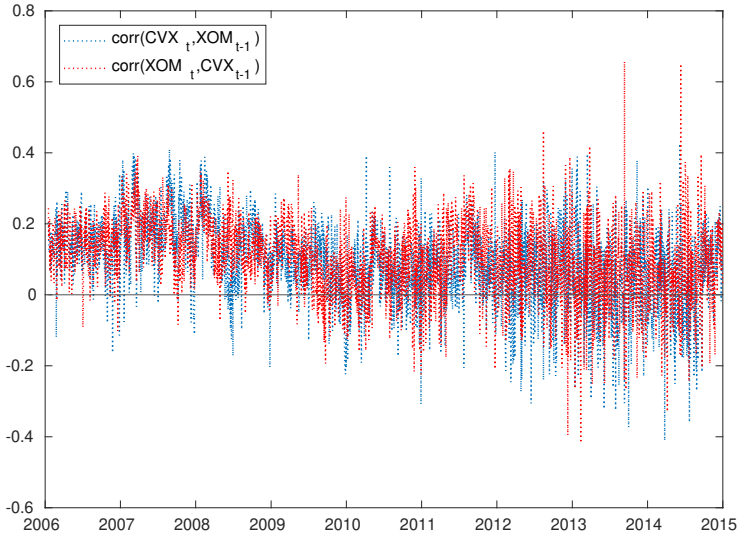


Figure 9: Dynamics of one second lead-lag correlations between Chevron and Exxon.

are computed from the MLA system matrices as described in the online appendix (Section A.4). Figure (9) shows the same result for Chevron and Exxon. In periods of high uncertainty the cross-autocorrelation structure of the market is more stable, with leag-lag correlations being typically above zero. This is particularly evident in the period of the 2007-2008 financial crisis, where lead-lag correlations are positive for both couples of assets.

In periods of low volatility leag-lag correlations are generally more erratic. We thus focus on subsamples in order to examine in more detail the cross-autocorrelation structure of the market in these periods. We consider here the last 250 days of the sample, i.e. the subsample coinciding with trades in 2014. In Figures (10), (11) we plot the average of the cross-autocorrelations of each couple of assets in Group I and Group II, respectively. Bars denote 95% confidence intervals. Nonzero correlations at positive lags imply that the second asset leads the first, while correlations at negative lags imply the opposite. As suggested by the previous analysis, we find strong evidences of cross-asset effects in both groups. For instance, in Group I Goldman Sachs appears to be the more informative asset since it leads all the other assets. In contrast, Bank of America is led by all the other assets. From Table (2) we note that Goldman Sachs is one of the least traded asset of Group I on average. Its lead role is therefore not due to differences in liquidities but to cross-asset effects that emerge as a consequence of nonzero non-diagonal coefficients in the lagged adjustment matrix Ψ. In Group II we note that oil companies like Exxon, ConocoPhillips, Chevron, Schlumberger lead the dynamics of energy companies like General Electric. We observe instead weaker lead-lag correlations among the leaders (e.g. between Exxon and Chevron).

There is statistical evidence of lead-lag correlations between assets belonging to different groups. These between-groups effects are in general weaker then within-group lead-lag effects. For instance, in Figure (12) we show the estimated correlogram of GS-GE (12a) and that of MS-CVX (12b). These are the two couples of assets belonging to different groups exhibiting the largest lead-lag correlations. In other cases we observe smaller or even non-significant correlations.

Lead-lag correlations can arise even if the non-diagonal elements of Ψ are all zero, as a consequence of combined autocorrelation and contemporaneous correlation effects. This is not the case here, as the estimated matrix $\hat{F} = \mathbb{I}_d - \hat{\Psi}$ has a lot of statistically significant non-diagonal elements. Tables (3), (4) show the average, over the subsample considered above, of the two sub-matrices of $\hat{F}$ corresponding to lead-lag coefficients of Group I and Group II, respectively. Non-diagonal elements are nonzero, with high significance. This implies that the recovered cross-asset structure arises as a direct consequence of the proposed multi-asset price formation mechanism. Such result can be

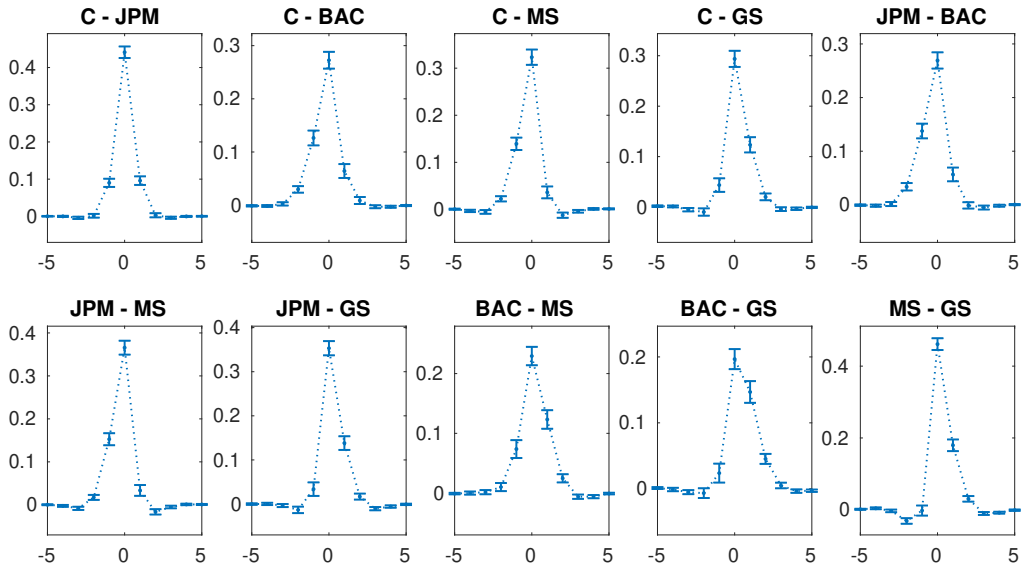


Figure 10: Average cross-autocorrelations of all the couples of assets in Group I. Averages are computed over all the business days of 2014. Error bars denote 95% confidence intervals. Correlations at positive lags imply that the second asset displayed in the title leads the first asset and the other way around for negative lags.

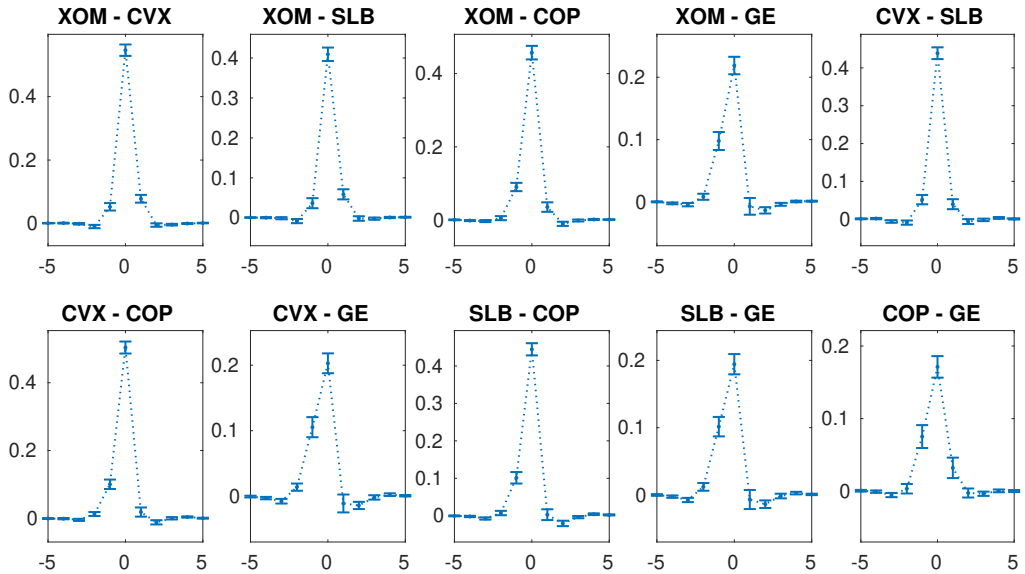


Figure 11: Average cross-autocorrelations of all the couples of assets in Group II. Averages are computed over all the business days of 2014. Error bars denote 95% confidence intervals. Correlations at positive lags imply that the second asset displayed in the title leads the first asset and the other way around for negative lags.

interpreted in light of cross-asset trading: dealers tend to rely on the prices of more informative securities in order to set their quotes and this translates into a lagged dissemination of information across assets, as captured by the non-diagonal elements of the matrix $\hat{\Psi}$.

	Group I				
	Average F_{ij}				
	C	JPM	BAC	MS	GS
C	0.0832****	0.0473****	0.0243*	-0.0637****	0.1212****
JPM	0.0311***	0.1029****	0.0102$^{(ns)}$	-0.0702****	0.1312****
BAC	0.0501****	0.0661****	0.0869****	0.0222$^{(ns)}$	0.1049****
MS	0.0731****	0.0969****	0.0111$^{(ns)}$	0.0199*	0.1551****
GS	0.0331**	0.0019$^{(ns)}$	-0.0101*	-0.0735****	0.1640****

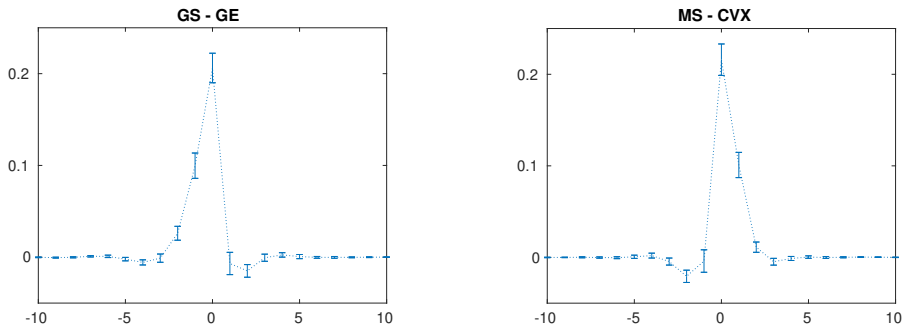
Table 3: We report the sample average, over the sub-sample of $N = 252$ days, of the elements of the estimated matrices $\hat{F}$ corresponding to assets belonging to Group I, together with significance levels obtained based on the p-value of the one-sample t-test: $^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$, $^{****}p \leq 0.0001$, $^{(ns)}p > 0.05$.

	Group II				
	Average F_{ij}				
	XOM	CVX	SLB	COP	GE
XOM	0.0769****	0.0434***	0.0266****	-0.0262$^{(ns)}$	-0.0418****
CVX	0.0231*	0.0973****	0.0158*	-0.0335*	-0.0329***
SLB	0.0141$^{(ns)}$	0.0660***	0.0934****	-0.0701****	-0.0346*
COP	0.0312**	0.0490***	0.0477****	0.0525****	-0.0043$^{(ns)}$
GE	0.0310*	0.0522***	0.0421****	-0.0013$^{(ns)}$	0.0533****

Table 4: We report the sample average, over the sub-sample of $N = 252$ days, of the elements of the estimated matrices $\hat{F}$ corresponding to assets belonging to Group II, together with significance levels obtained based on the p-value of the one-sample t-test: $^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$, $^{****}p \leq 0.0001$, $^{(ns)}p > 0.05$.

4.3 Inclusion of market portfolios

The 10 NYSE assets examined in Section (4.2) are homogeneous in terms of liquidity, as can be seen from the summary statistics in Table (2). As done in the simulation study, it is thus interesting



(a) **(b)**

Figure 12: Cross autocorrelations between stocks belonging to different groups. Averages are computed over all the business days of 2014. Error bars denote 95% confidence intervals. Correlations at positive lags imply that the second asset displayed in the title leads the first asset and the other way around for negative lags.

to assess the effect, on the estimated MLA parameters, of the inclusion of an asset with a very different level of trading activity. To this aim, we augment the previous sample with the SPY data. Compared to the 10 NYSE stocks, the SPY is a much more liquid security. The average duration between consecutive trades is slightly above one second, and therefore it is significantly smaller than that of the NYSE stocks, which is above three seconds.

The purpose of the following analysis is twofold. On the one hand, it serves as a robustness check to verify that the recovered MLA cross-autocorrelations are not altered by the inclusion of assets with different liquidities. On the other hand, it clarifies whether the well-known empirical finding that heavily traded market portfolios (e.g. ETF's and index futures) are "leaders" remains valid when using our MLA estimator which is robust to asynchronous trading.

Since SPY data are available from 03-01-2012 to 28-12-2012, we perform the analysis in this subsample. Specifically, we estimate the MLA on a daily basis in the dataset comprising the 10 NYSE stocks and the SPY. Figure (13) shows in red the average cross-autocorrelations of the assets belonging to Group I, estimated after including the SPY in the subsample. For comparison, we report in blue the average cross-autocorrelations estimated in absence of SPY data. We note that the inclusion of the SPY does not change significantly the estimated lead-lag correlations. In particular, our conclusions on which among two assets is the leader is the same in the two cases. A similar result is obtained by looking at Figure (14), which shows the average cross-autocorrelations of the assets belonging to Group II.

Figures (15a), (15b) show the average cross-autocorrelations between SPY and the 10 NYSE

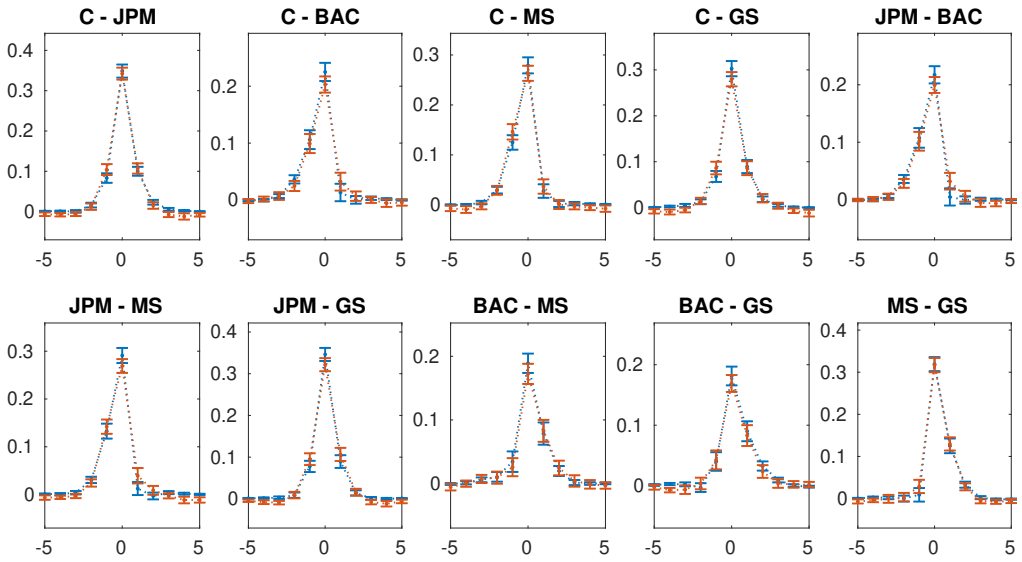


Figure 13: Average cross-autocorrelations of all the couples of assets in Group I. The cross-autocorrelations in red are obtained by the including the SPY in the sample. Averages are computed over all the business days of 2012. Error bars denote 95% confidence intervals. Correlations at positive lags imply that the second asset displayed in the title leads the first asset and the other way around for negative lags.

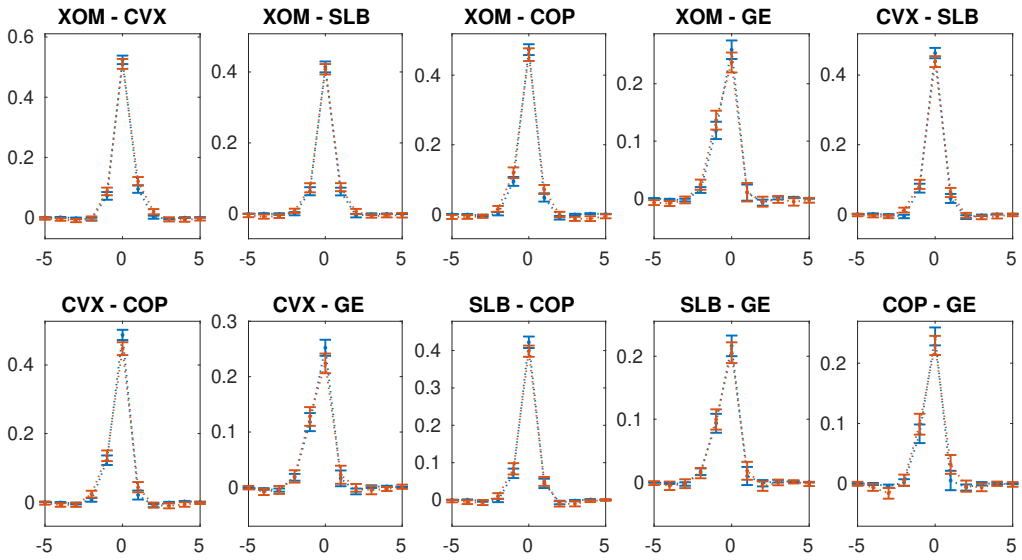
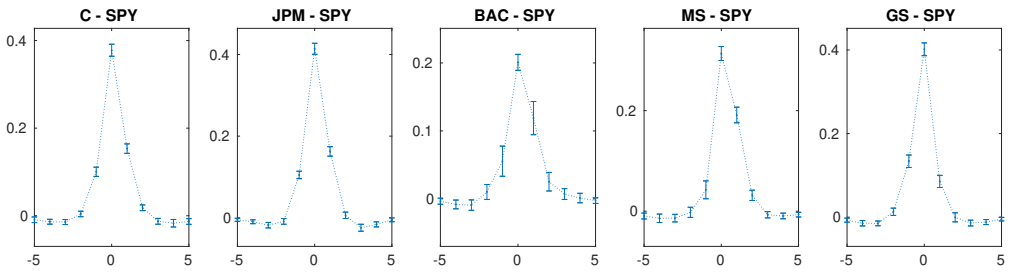
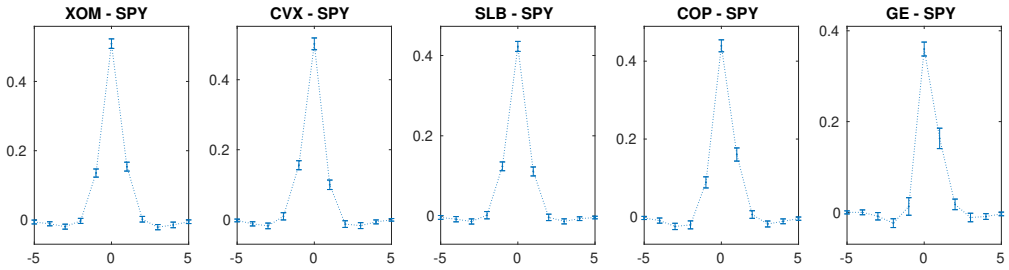


Figure 14: Average cross-autocorrelations of all the couples of assets in Group II. The cross-autocorrelations in red are obtained by including the SPY in the sample. Averages are computed over all the business days of 2012. Error bars denote 95% confidence intervals. Correlations at positive lags imply that the second asset displayed in the title leads the first asset and the other way around for negative lags.



(a)



(b)

Figure 15: Average cross-autocorrelations between SPY and the stocks of Group I (a) and between SPY and the stocks of Group II. Averages are computed over all the business days of 2012. Error bars denote 95% confidence intervals. Correlations at positive lags imply that the second asset displayed in the title leads the first asset and the other way around for negative lags

stocks. In six out of ten cases (C, JPM, BAC, MS, COP, GE), SPY has a larger lead correlation, in agreement with the common empirical finding that liquid market portfolios lead other stocks. In other two cases (XOM, SLB) there are no leaders, as the lead-lag correlations between SPY and the other assets are similar. Finally, Goldman Sachs and Chevron, which have been proven to be highly informative securities in the previous analysis, seem to lead SPY, although the differences between lead and lag correlations are small. As in Section (4.2), this result can be interpreted in light of the robustness of the MLA to asynchronous trading. In the MLA the presence of lead-lag correlations is not necessarily due to differences in the level of trading activities between assets. It is rather a consequence of cross-asset pricing: traders tend to look at the prices of the securities which are judged as the most representative of their economic sector before sending their orders, and this in turn generates lead-lag correlations.

5 Conclusions

In this paper we introduced the MLA, a multi-asset model of price generation that extends standard univariate microstructure models of lagged price adjustment. Lead-lag effects naturally arise in this framework as a consequence of nonzero non-diagonal coefficients in the lagged adjustment matrix Ψ. Inspired by the literature on cross-asset pricing, the latter captures lagged dissemination of information among different assets.

The MLA can be cast into a linear-Gaussian state-space representation where the transition equation is a VAR process for the returns of the adjusted price while the observation equation incorporates microstructure effects as an additive noise term. Asynchronicity is treated as a missing value problem, which is easily handled by the Kalman filter. The resulting estimator of lead-lag correlations is robust to asynchronous trading and microstructure noise. As a byproduct, we obtain an estimate of the integrated covariance of the efficient log-price process that takes into account asynchronous trading, microstructure noise and lead-lag dependencies. A particularly interesting case that is susceptible of treatment within the MLA is the one of assets linked by non-arbitrage constraints. In this case the MLA can be used for price discovery analysis and has the main advantage of not being affected by differences in trading activity among different markets.

Using extensive Monte-Carlo experiments, we showed that, as opposed to alternative estimators, the MLA is not affected by spurious correlations arising from asynchronous trading. Furthermore, we tested the performance of the estimator under different forms of misspecification for the underlying DGP. When covariances are time-varying, we found that the estimated VAR matrix $\hat{F}$ is unbiased and the covariance matrix $\hat{\Sigma}$ correctly estimates the integrated covariance of the efficient log-price process. In the online appendix we assessed the advantages of the MLA in price discovery.

The MLA was tested on a cross-section of NYSE stocks. The analysis provides empirical evidence for the existence of a multi-asset price formation mechanism. In particular, we observed strong deviations from the null assumption of a standard random walk plus noise process. The non-diagonal coefficients of the VAR matrix F are found to be significantly different from zero. As such, the speed of adjustment matrix Ψ includes non-diagonal elements that are responsible for the lagged dissemination of information across different assets. Cross-asset effects are more pronounced in periods of high volatility. Such empirical evidence is explained by the behavior of high-frequency trading strategies, which tend to exploit short living cross-autocorrelations that are likely to appear in periods of high uncertainty. We tested the robustness of the recovered results

with respect to the choice of the assets and the addition of market portfolios with much higher liquidity. In this respect, we found that the common empirical finding that highly liquid market portfolios are leaders is generally confirmed, though with some significant exceptions of highly informative securities leading the market portfolio.

These results could be further robustified through subsamples taken at the same frequency with different offsets (e.g. a first subsample starting at the first tenth of a second with a step of one second, a second subsample starting at the second tenth of a second with a step of one second, and so on). This could be particularly relevant for more recent periods and markets with multiple transactions within each second, dominated by high-speed algorithmic trading. Ideally, the inferred lead-lag effects across different subsamples should be in agreement and averaging the results across subsamples could further reduce the estimation error.

Finally, due to the latent VAR structure, our methodology can be viewed as a test for "latent" Granger causality, i.e. a Granger causality test on multivariate time-series of noisy and asynchronous observations. As such, it could be of potential interest for a broad spectrum of empirical applications.

References

A. P. Dempster, N. M. Laird, D.B.R., 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39, 1–38.

Aït-Sahalia, Y., Fan, J., Xiu, D., 2010. High-frequency covariance estimates with noisy and asynchronous financial data. J. Amer. Statist. Assoc. 105, 1504–1517.

Amihud, Y., Mendelson, H., 1987. Trading mechanisms and stock returns: An empirical investigation. The Journal of Finance 42, 533–553.

Andersen, T., Bollerslev, T., 1997. Intraday periodicity and volatility persistence in financial markets. Journal of Empirical Finance 4, 115–158.

Andrade, S., Chang, C., Seasholes, M., 2008. Trading imbalances, predictable reversals, and cross-stock price pressure. Journal of Financial Economics 88, 406–423.

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

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2011. Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. Journal of Econometrics 162, 149 – 169.

Barndorff-Nielsen, O.E., Hansen, P.R., Shephard, A.L.N., 2009. Realized kernels in practice: trades and quotes. The Econometrics Journal 12, C1–C32.

Bernhardt, D., Taub, B., 2008. Cross-asset speculation in stock markets. The Journal of Finance 63, 2385–2427.

Bibinger, M., Hautsch, N., Malec, P., Reiss, M., 2014. Estimating the spot covariation of asset prices: Statistical theory and empirical evidence. CFS Working Paper Series 477. Center for Financial Studies (CFS).

Bollerslev, T., Patton, A.J., Quaedvlieg, R., 2016. Exploiting the errors: A simple approach for improved volatility forecasting. Journal of Econometrics 192, 1 – 18.

Booth, G.G., So, R.W., Tse, Y., 1999. Price discovery in the german equity index derivatives markets. Journal of Futures Markets 19, 619–643.

Buccheri, G., Bormetti, G., Corsi, F., Lillo, F., 2019a. A score-driven conditional correlation model for noisy and asynchronous data: an application to high-frequency covariance dynamics. Working Paper, Available at https://ssrn.com/abstract=2912438.

Buccheri, G., Bormetti, G., Lillo, F., Corsi, F., 2019b. Comment on: Price Discovery in High Resolution. Journal of Financial Econometrics doi:10.1093/jjfinec/nbz008.

Caballé, J., Krishnan, M., 1994. Imperfect competition in a multi-security market with risk neutrality. Econometrica 62, 695–704.

Cespa, G., Focault, T., 2011. Learning from Prices, Liquidity Spillovers, and Market Segmentation. CSEF Working Papers 284. Centre for Studies in Economics and Finance (CSEF), University of Naples, Italy.

Chan, K., 1992. A further analysis of the lead–lag relationship between the cash market and stock index futures market. The Review of Financial Studies 5, 123–152.

Chiao, C., Hung, K., Lee, C.F., 2004. The price adjustment and lead-lag relations between stock returns: microstructure evidence from the taiwan stock market. Journal of Empirical Finance 11, 709 – 731.

Chu, Q.C., liang Gideon Hsieh, W., Tse, Y., 1999. Price discovery on the s&p 500 index markets: An analysis of spot index, index futures, and spdrs. International Review of Financial Analysis 8, 21 – 34.

Corsi, F., Peluso, S., Audrino, F., 2015. Missing in asynchronicity: A kalman-em approach for multivariate realized covariance estimation. Journal of Applied Econometrics 30, 377–397.

Damodaran, A., 1993. A simple measure of price adjustment coefficients. The Journal of Finance 48, 387–400.

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

Dobrev, D., Schaumburg, E., 2017. High-Frequency Cross-Market Trading: Model Free Measurement and Applications. Working Paper.

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, 291–298.

Glosten, L.R., Milgrom, P.R., 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. Journal of Financial Economics 14, 71 – 100.

Griffin, J.E., Oomen, R.C., 2011. Covariance measurement in the presence of non-synchronous trading and market microstructure noise. Journal of Econometrics 160, 58 – 68. Realized Volatility.

Hamilton, J., 1994. Time series analysis. Princeton Univ. Press, Princeton, NJ.

Harford, J., Kaul, A., 2005. Correlated order flow: Pervasiveness, sources, and pricing effects. The Journal of Financial and Quantitative Analysis 40, 29–55.

deB. Harris, F.H., McInish, T.H., Wood, R.A., 2002. Security price adjustment across exchanges: an investigation of common factor components for dow stocks. Journal of Financial Markets 5, 277 – 308. Price Discovery.

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

Hasbrouck, J., 1996. Modeling market microstructure time series. SSRN Electronic Journal NYU Working Paper No. FIN-95-024.

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

Hasbrouck, J., Seppi, D.J., 2001. Common factors in prices, order flows, and liquidity. Journal of Financial Economics 59, 383 – 411.

Hayashi, T., Koike, Y., 2017. Multi-scale analysis of lead-lag relationships in high-frequency financial markets. Papers 1708.03992. arXiv.org.

Hayashi, T., Koike, Y., 2018. Wavelet-based methods for high-frequency lead-lag analysis. SIAM Journal on Financial Mathematics 9, 1208–1248.

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

Hoffmann, M., Rosenbaum, M., Yoshida, N., 2013. Estimation of the lead-lag parameter from non-synchronous data. Bernoulli 19, 426–461. doi:10.3150/11-BEJ407.

Huth, N., Abergel, F., 2014. High frequency lead/lag relationships — empirical facts. Journal of Empirical Finance 26, 41 – 58.

de Jong, F., Nijman, T., 1997. High frequency analysis of lead-lag relationships between financial markets. Journal of Empirical Finance 4, 259 – 277.

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

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

Pasquariello, P., Vega, C., 2015. Strategic cross-trading in the u.s. stock market*. Review of Finance 19, 229.

Peluso, S., Corsi, F., Mira, A., 2015. A bayesian high-frequency estimator of the multivariate covariance of noisy and asynchronous returns. Journal of Financial Econometrics 13, 665–697.

Roll, R., 1984. A simple implicit measure of the effective bid-ask spread in an efficient market. The Journal of Finance 39, 1127–1139.

Shephard, N., Xiu, D., 2017. Econometric analysis of multivariate realised qml: Estimation of the covariation of equity prices under asynchronous trading. Journal of Econometrics 201, 19 – 42.

Shumway, R.H., Stoffer, D.S., 1982. An approach to time series smoothing and forecasting using the em algorithm. Journal of Time Series Analysis 3, 253–264.

Shumway, R.H., Stoffer, D.S., 2015. Time series analysis and its applications : with R examples. Springer texts in statistics, Springer, New York.

Tookes, H.E., 2008. Information, trading, and product market interactions: Cross-sectional implications of informed trading. The Journal of Finance 63, 379–413.

Tsay, R.S., 2005. Analysis of financial time series. Wiley series in probability and statistics, Wiley-Interscience, Hoboken (N.J.).

Wu, C.F.J., 1983. On the convergence properties of the em algorithm. Ann. Statist. 11, 95–103.

Zhang, F., 2010. High-Frequency Trading, Stock Volatility, and Price Discovery. Working paper, available at https://ssrn.com/abstract=1691679.

Appendix

A EM algorithm

As a first step, we assume there are no missing observations. We will show how to handle missing observations in the next sub-section. We denote by $\mathcal{X}_n = \{Z_0, \ldots, Z_n\}$ the set of latent prices and by $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ the set of observed prices. Also, let us assume that $Z_0 \sim \mathrm{N}(\mu, \Sigma)$. Note that, since the knowledge of Z_{t-1} completely determines the last d components of Z_t, the density function $f(Z_t|Z_{t-1})$ can be written as:

$$f(Z_t|Z_{t-1}) = f(MZ_t|Z_{t-1}) \tag{24}$$

Therefore, denoting by $\log L = \log L(\mathcal{Y}_n, \mathcal{X}_n)$ the complete log-likelihood function, we have:

$$\begin{aligned} \log L = \text{const} &- \frac{1}{2}\log|\Sigma| - \frac{1}{2}(Z_0-\mu)'\Sigma^{-1}(Z_0-\mu) \\ &- \frac{n}{2}\log|Q| - \frac{1}{2}\sum_{t=1}^{n}(Z_t - \Phi Z_{t-1})'M'Q^{-1}M(Z_t - \Phi Z_{t-1}) \\ &- \frac{n}{2}\log|H| - \frac{1}{2}\sum_{t=1}^{n}(Y_t - MZ_t)'H^{-1}(Y_t - MZ_t) \end{aligned} \tag{25}$$

The EM algorithm provides an iterative method for finding the MLE by successively maximizing the conditional expectation of the complete log-likelihood function. The latter can be computed using the Kalman filter and smoothing recursions.

Let us introduce the following quantities which can be recovered as an output of the Kalman filter and smoothing recursions in Section (A.3):

$$Z_t^s = \mathrm{E}[Z_t|\mathcal{Y}_s] \tag{26}$$

$$P_t^s = \mathrm{Cov}[Z_t|\mathcal{Y}_s] \tag{27}$$

$$P_{t,t-1}^s = \mathrm{Cov}[Z_t, Z_{t-1}|\mathcal{Y}_s] \tag{28}$$

With $s = t$, $s < t$ and $s > t$, the resulting conditional expectation is, respectively, an update filter, a predictive filter and a smoother. The Kalman filter is initialized with diffuse initial conditions, i.e. we set $\mathbb{E}[Z_1|Y_1] = 0$ and $\mathrm{Cov}[Z_1|Y_1] = \kappa\mathbb{I}_d$ with $\kappa \to \infty$. At iteration r, the expectation step in the EM algorithm consists in taking the conditional expectation of the complete log-likelihood

given the observations $\mathcal{Y}_n$ and using the estimate of $\Omega = \{F, Q, H\}$ obtained at step $r-1$:

$$\begin{aligned}\mathrm{E}[\log L|\mathcal{Y}_n, \hat{\Omega}_{r-1}] = &-\frac{1}{2}\log|\Sigma| - \frac{1}{2}\mathrm{Tr}[\Sigma^{-1}[(Z_0^n - \mu)(Z_0^n - \mu)' + P_0^n] \\ &- \frac{n}{2}\log|Q| - \frac{1}{2}\mathrm{Tr}[M'Q^{-1}M(C - B\Phi' - \Phi B' + \Phi A\Phi')] \\ &- \frac{n}{2}\log|H| - \frac{1}{2}\mathrm{Tr}[H^{-1}\sum_{t=1}^{n}[(Y_t - MZ_t^n)(Y_t - MZ_t^n)' + MP_t^nM']\end{aligned} \tag{29}$$

where A, B and C are given by:

$$A = \sum_{t=1}^{n}(P_{t-1}^n + Z_{t-1}^n Z_{t-1}^{n\prime}) \tag{30}$$

$$B = \sum_{t=1}^{n}(P_{t,t-1}^n + Z_t^n Z_{t-1}^{n\prime}) \tag{31}$$

$$C = \sum_{t=1}^{n}(P_t^n + Z_t^n Z_t^{n\prime}). \tag{32}$$

In the maximization step, the function $Q(\Omega|\hat{\Omega}_{r-1}) = \mathrm{E}[\log L|\mathcal{Y}_n, \hat{\Omega}_{r-1}]$ is maximized with respect to Ω. Let us consider the following terms depending on F, Q and H:

$$\begin{aligned}G_1(F,Q) &= -\frac{1}{2}\mathrm{Tr}[M'Q^{-1}M(C - B\Phi' - \Phi B' + \Phi A\Phi')] \\ G_2(F,Q) &= -\frac{n}{2}\log|Q| + G_1(F,Q) \\ G_3(H) &= -\frac{n}{2}\log|H| - \frac{1}{2}\mathrm{Tr}[H^{-1}[(Y_t - PZ_t)(Y_t - PZ_t)' + MP_t^nM']\end{aligned}$$

We start by solving the first order condition $\nabla_F G_1(F,Q) = 0$. Let us write the matrices A and B in the following form:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22}\end{pmatrix}, \quad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22}\end{pmatrix} \tag{33}$$

where A_{ij} and B_{ij}, $i = 1, 2$ are $d \times d$ submatrices of A and B. In Section (A.2) we prove the following:

Proposition 2. *The solution of the matrix equation $\nabla_F G_1(F,Q) = 0$ is:*

$$\hat{F}_r = \Gamma\Theta^{-1} \tag{34}$$

where $\Gamma = B_{11} - B_{12} - A_{11} + A_{12}$ and $\Theta = A_{11} + A_{22} - A_{12} - A_{21}$. The solution of the two matrix equations $\nabla_Q G_2(\hat{F}_r, Q) = 0$, $\nabla_H G_3(H) = 0$ are:

$$\hat{Q}_r = \frac{\hat{\Upsilon}}{n}, \quad \hat{H}_r = \frac{\mathrm{diag}(\Lambda)}{n} \tag{35}$$

where $\hat{\Upsilon} = M(C - B\hat{\Phi}_r' - \hat{\Phi}_r B' + \hat{\Phi}_r A \hat{\Phi}_r')M'$, $\Lambda = \sum_{t=1}^n [(Y_t - MZ_t^n)(Y_t - MZ_t^n)' + MP_t^n M']$ *and*

$$\hat{\Phi}_r = \begin{pmatrix} \mathbb{I}_d + \hat{F}_r & -\hat{F}_r \\ \mathbb{I}_d & 0_d \end{pmatrix} \tag{36}$$

Conditions under which the EM algorithm converges to a local maximum of the incomplete log-likelihood function are studied by Wu (1983). We check convergence by looking at the relative increase of the log-likelihood and stop the algorithm when it is lower than some small threshold ($\mu = 10^{-6}$ in our simulation and empirical study). The log-likelihood can be computed in the prediction error decomposition form:

$$\log L = \text{const} - \frac{1}{2}\sum_{t=1}^n \log |F_t| - \frac{1}{2}\sum_{t=1}^n v_t' F_t^{-1} v_t \tag{37}$$

where $v_t = Y_t - MZ_t^{t-1}$ is the prediction error and $F_t = MP_t^{t-1}M' + H$.

Once $\hat{F}$, $\hat{Q}$ and $\hat{H}$ have been estimated, the matrix of price adjustment Ψ and the covariance matrix of the efficient log-price process Σ can be computed as:

$$\hat{\Psi} = \mathbb{I}_d - \hat{F}, \quad \hat{\Sigma} = \hat{\Psi}^{-1}\hat{Q}\hat{\Psi}'^{-1} \tag{38}$$

The Kalman filter and smoothing recursions in Section (A.3) provide filtered and smoothed estimates of the lagged price X_t. From these, using Eq. (3), one also obtains as a byproduct filtered and smoothed estimates of the martingale efficient log-price process.

A.1 Missing value modification

The update formulas in the maximization step can be modified to take into account missing values. Let us assume that, at time t, d_1 components in the vector Y_t are observed while the remaining d_2 are not observed. We consider the d_1-dimensional vector $Y_t^{(1)}$ of observed components and the $d_1 \times d$ matrix $M_t^{(1)}$ whose lines are the lines of M corresponding to $Y_t^{(1)}$. Also, we consider the $d_1 \times d_1$ covariance matrix $H_t^{(11)}$ of observed components disturbances. Following Shumway and Stoffer (2015), the Kalman filter and smoothing recursions in Section (A.3) and the prediction error decomposition form of the log-likelihood, Eq. (37) are still valid, provided that one replaces Y_t, M and H with:

$$Y_{(t)} = \begin{pmatrix} Y_t^{(1)} \\ \underline{0} \end{pmatrix}, \quad M_{(t)} = \begin{pmatrix} M_t^{(1)} \\ \underline{0} \end{pmatrix}, \quad H_{(t)} = \begin{pmatrix} H_t^{(11)} & \underline{0} \\ \underline{0} & \mathbb{I}^{(22)} \end{pmatrix} \tag{39}$$

where $\mathbb{I}^{(22)}$ is the $d_2 \times d_2$ identity matrix and $\underline{0}$ generically denotes zero arrays of appropriate dimension. Note that the time dependence in $M_{(t)}$ and $H_{(t)}$ is only due to missing observations, while the matrices M and H are constant over time.

Taking the conditional expectation in Eq. (25) requires some modifications in case of missing observations. The second and the fourth term remain as in Eq. (29), provided that one runs Kalman filter and smoothing recursions as described in (39). The last term changes because one needs to evaluate expectations of Y_t conditioning to the incomplete data $\mathcal{Y}_n^{(1)} = \{Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_n^{(1)}\}$. If H is diagonal, as we are assuming here, Shumway and Stoffer (1982) showed that:

$$\begin{aligned}\mathbb{E}[(Y_t - MZ_t)(Y_t - MZ_t)'|\mathcal{Y}_n^{(1)}] &= (Y_{(t)} - M_{(t)}Z_t^n)(Y_{(t)} - M_{(t)}Z_t^n)' \\ &+ M_{(t)}P_t^n M_{(t)}' + \begin{pmatrix} \underline{0} & \underline{0} \\ \underline{0} & \hat{H}_{22,t,r-1} \end{pmatrix}\end{aligned} \tag{40}$$

where $\hat{H}_{22,r-1}$ is the $d_2 \times d_2$ covariance matrix of unobserved components disturbances at time t obtained using the estimate at step $r-1$ of the matrix H. Therefore, the update equation for H becomes:

$$\hat{H} = \frac{\text{diag}(\Lambda^*)}{n} \tag{41}$$

where

$$\Lambda^* = \sum_{t=1}^{n} D_t \left[(Y_{(t)} - MZ_t^n)(Y_{(t)} - MZ_t^n)' + M_{(t)}P_t^n M_{(t)}' + \begin{pmatrix} \underline{0} & \underline{0} \\ \underline{0} & \hat{H}_{22,t,r-1} \end{pmatrix} \right] D_t', \tag{42}$$

D_t being a permutation matrix that rearranges the components of Y_t in their original order.

A.2 Proof of Proposition 2

We will use the following matrix differentiation rules:

$$\nabla_A \text{tr}(AB) = B' \tag{43}$$

$$\nabla_A \text{tr}(ABA'C) = CAB + C'AB' \tag{44}$$

$$\nabla_A |A| = |A|(A^{-1})' \tag{45}$$

where A, B and C are matrices of appropriate dimensions.

Let us re-write $G_1(F, Q)$ as:

$$G_1(F, Q) = -\frac{1}{2}\text{Tr}[Q^{-1}(MCM' - \tilde{B}\tilde{\Phi}' - \tilde{\Phi}\tilde{B}' + \tilde{\Phi}A\tilde{\Phi}')] \tag{46}$$

where we have defined $\tilde{B} = MB$ and $\tilde{\Phi} = M\Phi$. Let us compute explicitly the terms in $G_1(F, Q)$ depending on F:

$$\begin{aligned}
\tilde{B}\tilde{\Phi}' &= B_{11}(\mathbb{I} + F') - B_{12}F' \\
\tilde{\Phi}\tilde{B}' &= (\mathbb{I} + F)B_{11}' - FB_{12}' \\
\tilde{\Phi}A\tilde{\Phi}' &= (\mathbb{I} + F)A_{11}(\mathbb{I} + F') - FA_{21}(\mathbb{I} + F') \\
&\quad - (\mathbb{I} + F)A_{12}F' + FA_{22}F'
\end{aligned}$$

Therefore, we need to solve $\nabla_F \overline{G}_1(F) = 0$, where:

$$\begin{aligned}
\overline{G}_1(F) = \mathrm{Tr}[Q^{-1}(&-B_{11}(\mathbb{I} + F') + B_{12}F' - (\mathbb{I} + F)B_{11}' + FB_{12}' \\
&+ (\mathbb{I} + F)A_{11}(\mathbb{I} + F') - FA_{21}(\mathbb{I} + F') - (\mathbb{I} + F)A_{12}F' + FA_{22}F')]
\end{aligned}$$

This can be done using Eq. (43) and (44). One obtains:

$$\begin{aligned}
\nabla_F \overline{G}_1(F) = Q^{-1}[&-2(B_{11} - B_{12} - A_{11} + A_{12}) \\
&+ 2F(A_{11} + A_{22} - A_{21} - A_{12})]
\end{aligned} \tag{47}$$

and therefore:

$$\hat{F} = \Gamma\Theta^{-1} \tag{48}$$

We now solve $\nabla_{Q^{-1}} G_2(\hat{F}_r, Q) = 0$. We obtain:

$$\begin{aligned}
&\nabla_{Q^{-1}} G_2(\hat{F}_r, Q) = \\
&= \nabla_{Q^{-1}} \left[-\frac{n}{2} \log |Q| - \frac{1}{2} \mathrm{Tr}(Q^{-1}\hat{\Upsilon}) \right] \\
&= \frac{n}{2} Q - \frac{1}{2}\hat{\Upsilon}'
\end{aligned} \tag{49}$$

and therefore, since $\hat{\Upsilon}' = \hat{\Upsilon}$:

$$\hat{Q} = \frac{\hat{\Upsilon}}{n} \tag{50}$$

Finally, now solve $\nabla_H G_3(\hat{F}_r, Q) = 0$. Note that, since H is diagonal, we can write:

$$\begin{aligned}
&\nabla_H G_3(H) = \\
&= \nabla_H \left[-\frac{n}{2} \log |H| - \frac{1}{2} \mathrm{Tr}(H^{-1}\mathrm{diag}(\Lambda)) \right] \\
&= \frac{n}{2} H - \frac{1}{2}\Lambda
\end{aligned} \tag{51}$$

and therefore:

$$\hat{H} = \frac{\mathrm{diag}(\Lambda)}{n} \tag{52}$$

A.3 Kalman filter and smoothing recursions

The set of Kalman filter recursions for the state-space model (9), (10) are given by:

$$Z_t^{t-1} = \Phi Z_{t-1}^{t-1} \tag{53}$$

$$P_t^{t-1} = \Phi P_{t-1}^{t-1}\Phi' + Q \tag{54}$$

$$K_t = P_t^{t-1}M'(MP_t^{t-1}M' + H)^{-1} \tag{55}$$

$$Z_t^t = Z_t^{t-1} + K_t(Y_t - MZ_t^{t-1}) \tag{56}$$

$$P_t^t = P_t^{t-1} - K_t H P_t^{t-1} \tag{57}$$

for $t = 1, \ldots, n$. The set of backward smoothing recursions are given by:

$$J_{t-1} = P_{t-1}^{t-1}\Phi'(P_t^{t-1})^{-1} \tag{58}$$

$$Z_{t-1}^n = Z_{t-1}^{t-1} + J_{t-1}(X_t^n - \Phi Z_{t-1}^{t-1}) \tag{59}$$

$$P_{t-1}^n = P_{t-1}^{t-1} + J_{t-1}(P_t^n - P_t^{t-1})J_{t-1}' \tag{60}$$

for $t = n, \ldots, 1$. The covariance $P_{t,t-1}^n$ in Eq. (31) can be computed using the following backward recursion:

$$P_{t-1,t-2}^n = P_{t-1}^{t-1}J_{t-2}' + J_{t-1}(P_{t,t-1}^n - \Phi P_{t-1}^{t-1})J_{t-2}' \tag{61}$$

where $t = n, \ldots 2$ and $P_{n,n-1}^n = (I - K_n M)\Phi P_{n-1}^{n-1}$.

A.4 Computation of lead-lag correlations

In order to compute lead-lag correlations, we first compute the j-th order autocovariance matrix, which is defined as:

$$S_j = \mathbb{E}[\Delta X_t \Delta X_{t-j}'] \tag{62}$$

It can be evaluated from the estimated matrices $\hat{F}$ and $\hat{Q}$ as:

$$\hat{S}_j = \hat{F}\hat{S}_{j-1}, \quad j = 1, 2, \ldots \tag{63}$$

where the covariance matrix $S_0 = \mathbb{E}[\Delta X_t \Delta X_t']$ is estimated as:

$$\mathrm{vec}(\hat{S}_0) = (\mathbb{I}_{d^2} - \hat{F} \otimes \hat{F})^{-1}\mathrm{vec}(\hat{Q}) \tag{64}$$

see e.g. Hamilton (1994). Lead-lag correlations are finally obtained by normalizing the autocovariances with the diagonal elements of $\hat{S}_0$.

B Arbitrage-linked securities

B.1 The MLA with cointegrated dynamics

In this section we discuss the case in which all or some of the assets are linked by non-arbitrage constraints. A paradigmatic example is given by a security traded in different exchanges. Due to non-arbitrage, the prices observed in the these exchanges cannot move "too far" from each others. In his pioneering work, Hasbrouck (1995) proposed a vector error correction model (VECM) approach and introduced the well-known information shares (IS), which quantify the fraction of the total variance of the (unique) efficient price process explained by individual exchanges. Alternative strategies have been proposed over the years by several researchers (including, amongst others, Booth et al. 1999, Chu et al. 1999, deB. Harris et al. 2002, De Jong and Schotman 2010).

The MLA with cointegration restrictions can be employed to determining the contribution of individual exchanges to the price discovery of arbitrage-linked securities. Let us first consider a simple case with two distinct securities. The efficient log-prices of the two securities evolve as a random walk:

$$P_{t+1}^{(i)} = P_t^{(i)} + u_{t+1}^{(i)}, \quad i = 1, 2 \tag{65}$$

where $\text{Var}[u_{t+1}^{(i)}] = q_i$ and $\text{Cov}[u_{t+1}^{(1)}, u_{t+1}^{(2)}] = c$. Let us assume that the first security is traded in $d_1 \geq 1$ different markets and the second security is traded in $d_2 \geq 1$ different markets. We denote by $Y_t^{(1)} \in \mathbb{R}^{d_1}$ the vector of observations of the first asset log-price in the d_1 markets and, similarly, we denote by $Y_t^{(2)} \in \mathbb{R}^{d_2}$ the vector of observations of the second asset log-price in the d_2 markets. We define $d = d_1 + d_2$ and write $Y_t = [Y_t^{(1)\prime}, Y_t^{(2)\prime}]' \in \mathbb{R}^d$ as:

$$Y_t = X_t + \epsilon_t \tag{66}$$

where $\text{Cov}[\epsilon_t] = H$ and X_t is the lagged log-price process. The latter is decomposed as $X_t = [X_t^{(1)\prime}, X_t^{(2)\prime}]'$, where $X_t^{(1)} \in \mathbb{R}^{d_1}$ and $X_t^{(2)} \in \mathbb{R}^{d_2}$ are given by:

$$X_{t+1}^{(i)} = X_t^{(i)} + \Psi^{(i)}(\iota_{d_i} P_{t+1}^{(i)} - X_t^{(i)}), \quad i = 1, 2 \tag{67}$$

with $\Psi^{(i)} \in \mathbb{R}^{d_i \times d_i}$ and $\iota_{d_i} \in \mathbb{R}^{d_i}$ is a vector of ones. The difference with respect to Eq. (3) in the paper is that $X_t^{(i)} \in \mathbb{R}^{d_i}$ pertain the same security and are therefore driven by the same scalar log-price $P_{t+1}^{(i)}$.

It is immediate to see that the log-return process $\Delta X_t^{(i)} = X_t^{(i)} - X_{t-1}^{(i)}$ follows the VAR(1) process:

$$\Delta X_{t+1}^{(i)} = (\mathbb{I}_{d_i} - \Psi^{(i)})\Delta X_t^{(i)} + \Psi^{(i)}\iota_{d_i}\omega_{t+1}^{(i)}, \quad i = 1, 2 \tag{68}$$

As before, the difference with respect to Eq. (5) in the paper is that $\Delta X_{t+1}^{(i)} \in \mathbb{R}^{d_i}$ are now driven by the same scalar innovation $\omega_t^{(i)}$. We then consider the whole vector of log-returns $\Delta X_t = [\Delta X_t^{(1)\prime}, \Delta X_t^{(2)\prime}]' \in \mathbb{R}^d$, which follows:

$$\Delta X_{t+1} = (\mathbb{I}_d - \Psi)\Delta X_t + \Psi\omega_{t+1} \tag{69}$$

where:

$$\Psi = \begin{pmatrix} \Psi^{(1)} & 0_{d_1\times d_2} \\ 0_{d_2\times d_1} & \Psi^{(2)} \end{pmatrix}, \quad \omega_{t+1} = \begin{pmatrix} \iota_{d_1}\omega_{t+1}^{(1)} \\ \iota_{d_2}\omega_{t+1}^{(2)} \end{pmatrix} \tag{70}$$

Note that the covariance matrix $Q = \mathrm{Cov}[\omega_t]$ has rank equal to two.

It is not difficult to generalize the previous equations to the case in which there are k distinct securities, with the i-th efficient log-price observed in $d_i \geq 1$ different markets, $i = 1, \ldots, k$ and $\mathrm{Cov}[u_{t+1}^{(i)}, u_{t+1}^{(j)}] = c_{ij}$. We recover the standard MLA considered in the paper when $d_i = 1$ for each i. If $d_i > 1$ for at least one i, some of the log-prices are cointegrated. Defining $d = \sum_{i=1}^k d_i$, the vector of log-returns $\Delta X_t = [\Delta X_t^{(1)\prime}, \ldots, \Delta X_t^{(k)\prime}]' \in \mathbb{R}^d$ follows:

$$\Delta X_{t+1} = (\mathbb{I}_d - \Psi)\Delta X_t + \Psi\omega_{t+1} \tag{71}$$

where:

$$\Psi = \begin{pmatrix} \Psi^{(1)} & \ldots & 0_{n_1\times n_k} \\ \vdots & \ddots & \vdots \\ 0_{n_k\times n_1} & \ldots & \Psi^{(k)} \end{pmatrix}, \quad \omega_{t+1} = \begin{pmatrix} \iota_{n_1}\omega_{t+1}^{(1)} \\ \vdots \\ \iota_{n_k}\omega_{t+1}^{(k)} \end{pmatrix} \tag{72}$$

The rank of the covariance matrix $Q = \mathrm{Cov}[\omega_{t+1}]$ is always equal to k.

Similarly to what we have done in the standard MLA, let us introduce the augmented state vector $Z_t = [X_t', X_{t-1}']' \in \mathbb{R}^{2d}$. We can re-write Eq. (66), (69) as:

$$Y_t = MZ_t + \epsilon_t, \quad \mathrm{Cov}[\epsilon_t] = H \tag{73}$$

$$Z_{t+1} = \Phi Z_t + R\xi_{t+1}, \quad \mathrm{Cov}[\xi_t] = W \tag{74}$$

where:

$$\Phi = \begin{pmatrix} 2\mathbb{I}_d - \Psi & -\mathbb{I}_d + \Psi \\ \mathbb{I}_d & 0_d \end{pmatrix}, \quad R = \begin{pmatrix} \Psi & 0_d \\ 0_d & 0_d \end{pmatrix}, \quad W = \begin{pmatrix} Q & 0_d \\ 0_d & 0_d \end{pmatrix} \tag{75}$$

and $M = [\mathbb{I}_d, 0_d]$. This is a linear-Gaussian state-space representation that is susceptible of treatment through the Kalman filter. Due to the singularity of Q, the complete log-likelihood in Eq. (25) does not exist. The complete log-likelihood exists in a k-dimensional subspace of $\mathbb{R}^d$ generated by linear combinations of the components of X_t. We thus estimate the model by standard maximum-likelihood, i.e. by numerically optimizing the log-likelihood in Eq. (37). Compared to the standard MLA described in the paper, numerical optimization is feasible here, because the number of parameters of Q is $\mathcal{O}(k^2)$ rather than $\mathcal{O}(d^2)$. For instance, if there is only one security traded in d exchanges, we only need to estimate the variance of the unique innovation driving the dynamics in the d exchanges, regardless the value of d.

The diagonal elements of $\Psi^{(i)}$ induce a delay between the d_i exchanges and the i-th efficient log-price. For this reason they can be regarded as a measure of the informativeness of each exchange. For instance, if we find that one market anticipates another market, we conclude that the former is more informative. In principle one can study "cross-market" effects among different exchanges generated by non-diagonal coefficients in Ψ_i. This is computationally feasible when d_i is not too large, since the number of parameters of a non-diagonal Ψ_i scales as $\mathcal{O}(d_i^2)$. Similarly, to include cross-asset effects among prices corresponding to different securities, we need non-diagonal coefficients in Ψ. Even in this case one should pay attention that the total number of coefficients in Ψ does not grow too fast with d.

B.2 Comparison with other methodologies

As we have seen in the Monte-Carlo analysis of Section (3) in the paper, one of the main advantages of the MLA is that parameter estimates are not affected by differences in the level of trading activity among different assets. Similarly, in the case of arbitrage-linked securities, parameter estimates are not affected by differences in the level of trading activity among different markets. From an empirical point of view, this circumstance is of particular relevance in presence of informed traders. As predicted by classical models of price formation (Glosten and Milgrom 1985, Kyle 1985), the informed traders buy or sell securities if their trade guarantees a profit net of transaction costs, i.e. when their size relative to fundamental values is large. According to this logic, the informed traders tend to buy or sell in the market with larger mispricing between the efficient log-price and the mid-quote price.

To illustrate this concept, we consider the price formation model in Bandi et al. (2017) and

adapt it to our framework with one security traded in several exchanges. In this model of price formation we have three components: an efficient log-price process, the midquote adjustments and the observed log-prices. The efficient log-price process follows a random walk:

$$P_{t+1} = P_t + u_{t+1} \tag{76}$$

where $\mathrm{Var}[u_t] = q$. We assume that the security described by P_t is traded in two exchanges. The midquote log-price is related to the efficient log-price by a lagged adjustment process:

$$X_{t+1}^{(i)} = X_t^{(i)} + \delta^{(i)}(P_{t+1} - X_t^{(i)}), \quad i = 1, 2 \tag{77}$$

The observed log-price depends on the trader type. Let us denote by $\mathcal{I}$ the probability of arrival of informed traders. For simplicity, we assume that $\mathcal{I}$ is the same in both markets. The informed trader knows the value of the efficient log-prices P_t and decides whether to trade or not by comparing the mispricing $|X_t^{(i)} - P_t|$ to the transaction cost $c^{(i)}$. More specifically, if $|X_t^{(i)} - P_t| > c^{(i)}$ the informed trader decides to trade and the observed log-price is:

$$Y_t^{(i)} = X_t^{(i)} + c \cdot 1_{\{P_t - X_t^{(i)} > c\}} - c \cdot 1_{\{P_t - X_t^{(i)} < -c\}}, \quad i = 1, 2 \tag{78}$$

In contrast, if $|X_t^{(i)} - P_t| \leq c^{(i)}$, the informed trader decides not to trade. Noise traders behave randomly. They simply toss a coin and decide whether to trade or not. When a noise trader arrives on the market, the observed log-price is:

$$Y_t^{(i)} = X_t^{(i)} + \nu_t^{(i)} c^{(i)} \tag{79}$$

where $\nu_t^{(i)}$ is a sequence of independent Bernoulli variables taking the values ± 1 with likelihood 50%.

If there are informed traders ($\mathcal{I} > 0$), the level of trading activity depends on the transaction cost $c^{(i)}$ and on the speed $\delta^{(i)}$ of adjustment to the efficient price. In particular, if $c^{(i)}$ is large, informed traders cannot reward themselves net of transaction costs, and decide not to trade. Similarly, if the market rapidly adjusts to the efficient price ($\delta^{(i)} \approx 1$), the mispricing $|X_t^{(i)} - P_t|$ is small and the prices are not updated. If $\mathcal{I}$ is sufficiently large, the latter case leads to a highly informative market (i.e. one with a high speed of adjustment to the efficient price) being less traded than an inefficient market (i.e. one which slowly adapts to the efficient price). In reality this situation might represent a transition to an equilibrium where the price in the less efficient market gradually reverts to the martingale process. This simple microstructure model can be extended into several directions, e.g.

	MLA		IS bounds	
	$\hat{\delta}^{(1)}$	$\hat{\delta}^{(2)}$	1^{st} market	2^{nd} market
$\mathcal{I}=0$	0.8456 (0.051)	0.2108 (0.027)	$(0.6187, 0.9969)$	$(0.0031, 0.3813)$
$\mathcal{I}=0.3$	0.9175 (0.054)	0.1975 (0.022)	$(0.3777, 0.9136)$	$(0.0863, 0.6222)$
$\mathcal{I}=0.5$	0.9181 (0.075)	0.1850 (0.042)	$(0.2717, 0.7692)$	$(0.2307, 0.7282)$
$\mathcal{I}=0.7$	0.8713 (0.078)	0.2234 (0.051)	$(0.0835, 0.3679)$	$(0.6320, 0.9164)$

Table 5: MLA estimates of parameters $\delta^{(i)}$ in Eq. (77) with standard errors reported in parenthesis and IS bounds.

introducing a time-varying $\delta_t^{(i)}$ or allowing noise traders to take into account transaction costs (see Bandi et al. 2017). For simplicity we examine here the basic specification, however nothing prevents adding other features to simulate more and more realistic scenarios.

The MLA, though misspecified in this setting (note that the measurement noise in Eq. (79) is not normal), has a clear advantage in being robust to differences in the level of trading activity. To see this, let us assume that we have two markets, one which rapidly adjusts to new information ($\delta^{(1)}=0.9$), and one which adapts slowly ($\delta^{(2)}=0.2$). For simplicity, transaction costs are assumed to be the same in the two markets. We set the remaining parameters as $c^{(1)}=c^{(2)}=0.2\,\$$, $q=0.5$. As a time-horizon we consider a trading day of 6.5 hours, from 9:30 to 16:00. We thus simulate 23400 one-second realizations of the efficient log-price P_t in Eq. (76) and of the two lagged adjustment log-prices $X_t^{(1)}$ and $X_t^{(2)}$ in Eq. (77).

We start by assuming that there are no informed traders ($\mathcal{I}=0$). The observed prices are determined only by noise traders who do not know the value of the true efficient log-price P_t. The level of trading activity in the two markets is thus the same. Table (5) shows the parameters $\hat{\delta}^{(i)}$, $i=1,2$ estimated by the MLA and corresponding to the diagonal elements of the matrix Ψ in Eq. (67). We also report the IS measure of Hasbrouck (1995). The IS is generally not uniquely defined since it depends on the order of the assets in the underlying VECM. For this reason, we report for each market the two bounds obtained by reversing the order of the two time-series. The MLA estimates are very close to the true parameters, and thus we conclude that the first market is the one where price discovery occurs. Using the IS we get to the same conclusion since the first market has larger information share and the bounds are relatively narrow.

As $\mathcal{I}$ increases, informed traders arrive on the market. They can decide not to trade should

the mispricing $|X_t^{(i)} - P_t|$ be smaller than the transaction costs. The absence of trading leads to missing values in the two time-series. The MLA estimates remain close to the true parameters, and we still conclude that the first market is the one where price discovery occurs. In order to estimate the VECM, we fill the missing values by previous-tick interpolation, as commonly done in the financial econometric literature. Such procedure leads to erroneous conclusions on the degree of informativeness of the two markets. As $\mathcal{I}$ increases, the IS bounds widen. When $\mathcal{I} = 0.5$, the two markets have very similar bounds, and it is impossible to discern where price discovery occurs. For $\mathcal{I} = 0.7$, the bounds narrow but we wrongly conclude that the second market is more informative than the first. This is due to the first market being less traded than the second because of the large amount of informed traders who exploit arbitrage opportunities in the less efficient market.

More generally, asynchronous trading leads to missing values in the observed time-series which are typically filled by previous-tick interpolation. This leads, in turn, to a large number of "artificial" zero-returns which are misspecified under the common semimartingale assumption (in continuous-time) or under the VAR/VECM assumption (in discrete-time). The most known distortions of zero-returns is the Epps effect (cf. e.g. Hayashi and Yoshida 2005 and references therein). Thus, the fact that we find misleading results when applying the VECM to asynchronous data is not surprising. The main advantage of the MLA is that it can handle missing observations without introducing artificial zero returns.

There are other differences between IS and the MLA with cointegrated dynamics. The IS of the i-th exchange is defined as the fraction of the long-run variance of the common trend imputable to that exchange. The market contributing with the largest fraction of variance is the most informative, i.e. the one where price discovery occurs. In the MLA with cointegrated dynamics, the most informative market is the one with highest speed of adjustment to the efficient log-price. In other words the market that leads all the other markets is the one where price discovery occurs.

Using these two approaches one may achieve a different conclusion on which market is the most informative. The main reason is that IS does not depend on the matrices of autoregressive coefficients of the underlying VECM, as formally shown by Buccheri et al. (2019b). This implies that,in some circumstances, a market with a substantial delay from the common trend may be judged as equally informative or even more informative than a faster market (see the examples in Buccheri et al. 2019b). In the MLA this cannot happen since the most informative market is, by definition, the one which adapts with highest speed to the common trend represented by the efficient log-price process. We refer to Buccheri et al. (2019b) for further discussions on these

aspects.

Finally, we point out that, if quote data are available, X_t becomes observable and thus IS is not affected by asynchronous trading. However, compared to the MLA, it is still true that it may lead to misleading results in the presence of significant lags between the observed prices and the underlying martingale process (cf. Buccheri et al. 2019b).

C Robustness checks

C.1 Model invariance to the choice of the assets

In Section (4.2) in the paper we estimated the MLA on a cross-section of 10 NYSE stocks and showed the average cross-autocorrelations in Figures (10), (11). The question naturally arises whether these lead-lag correlations depend on the specific choice of the dataset and whether by selecting a subset of these assets one would obtain the same result. This is a standard issue in the specification of VAR models (see e.g. Kilian and Lütkepohl 2017).

In order to investigate if the recovered lead-lag correlations are robust with respect to the choice of the assets, we perform the same analysis of Section (4.2) in the paper but here we estimate a different MLA for each couple of assets. Specifically, in the first case, the lead-lag correlations of two assets are computed using the log-prices of all the 10 NYSE assets, while in the second case they are computed based only on the log-prices of the two assets. This comparison is interesting because we are considering the scenario in which the discrepancy between the two kinds of lead-lag correlations is largest: estimating the MLA on cross-sections of growing dimensions provides lead-lag correlations which become more and more similar to those obtained using the entire cross-section.

The results are reported in Figures (16), (17), where we show in blue the correlogram obtained in Section (4.2) in the paper and in red the new lead-lag correlations. The lead-lag correlations recovered using the whole dataset of 10 assets are very similar to those obtained by estimating the MLA pairwisely. We only observe that, due to data reduction, pairwise correlations are slightly lower in some cases. However, our conclusions on which among two assets is the leader is invariant with respect to the choice of the dataset.

In Section (4.3) in the paper we assess the effect of the inclusion of the SPY in the sample of assets used for the estimation. The SPY differs significantly from the 10 NYSE assets in terms of liquidity, as it features a larger number of trades per day. Even in that case we find that the

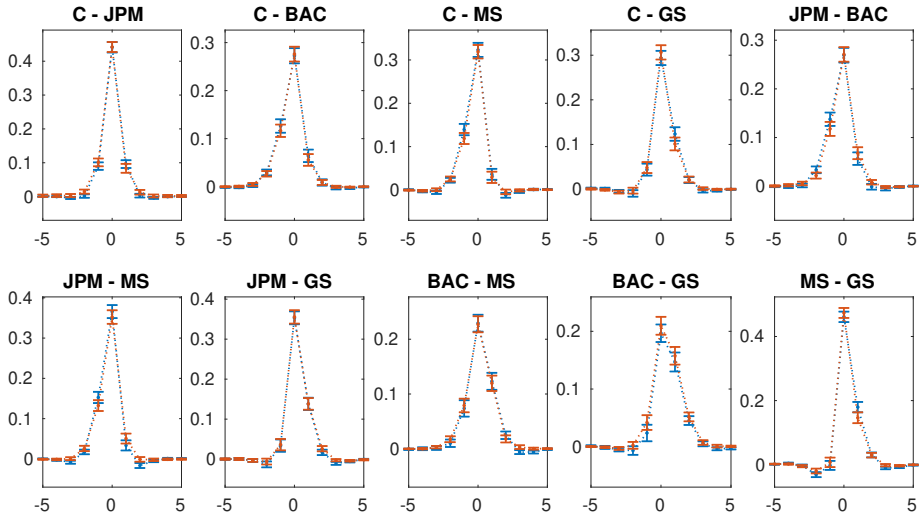


Figure 16: Average cross-autocorrelations of all the couples of assets in Group I. Averages are computed over all the business days of 2014. We show in blue the lead-lag correlations computed as in Section (4.2) in the paper and in red the lead-lag correlations computed pairwisely. Error bars denote 95% confidence intervals. Correlations at positive lags imply that the second asset displayed in the title leads the first asset and the other way around for negative lags.

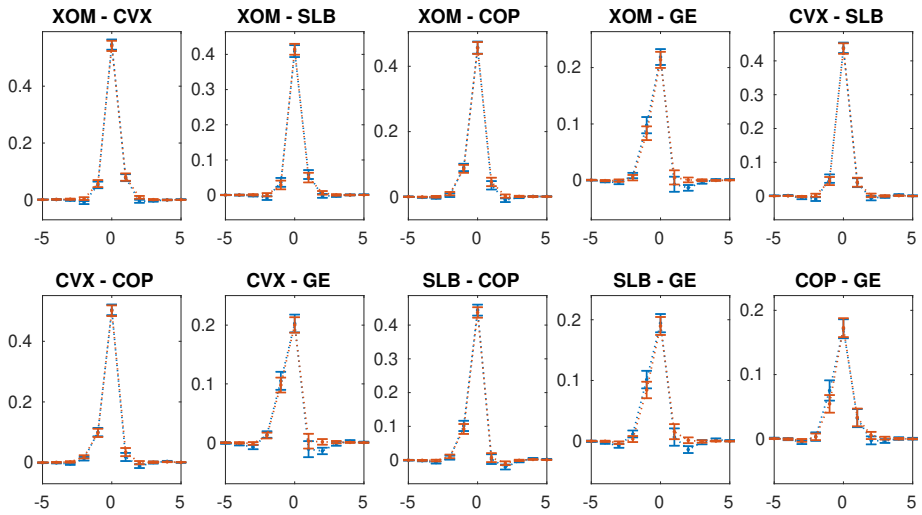


Figure 17: Average cross-autocorrelations of all the couples of assets in Group II. Averages are computed over all the business days of 2014. We show in blue the lead-lag correlations computed as in Section (4.2) in the paper and in red the lead-lag correlations computed pairwisely. Error bars denote 95% confidence intervals. Correlations at positive lags imply that the second asset displayed in the title leads the first asset and the other way around for negative lags.

recovered cross-autocorrelation structure is not altered significantly.

C.2 Statistical significance at sparser modelling frequencies

All the empirical results reported in Section (4) have been recovered at the sampling frequency of one second, which is the highest frequency achievable in our dataset. It is interesting to investigate how these results would change at sparser frequencies. As a result of data reduction, the use of sparser sampling frequencies naturally leads to a lower statistical efficiency. However, the loss of efficiency is not the only source of concern when subsampling prices at sparser frequencies. Another aspect that must be taken into account is the fact that, being generated by high-frequency trading strategies, lead-lad dependencies exist at small time-scales and tend to decay at longer time-scales.

To illustrate this phenomenon, we compare in Figure (18) the average cross-autocorrelations of the assets of Group I computed at the sampling frequency of one second (in red) and the cross-autocorrelations of the same assets computed at the sampling frequency of 10 seconds (in blue). We use the subsample of Section (4.3) which includes SPY data. Figure (19) shows a similar comparison for the lead-lag correlations between the SPY and the five assets of Group I. We immediately note that, as a consequence of modelling prices at sparser sampling frequencies, the contemporaneous correlations increase and the lead-lag correlations decrease. In other words, the lead-lag correlations detected at higher resolutions are "averaged out" and are eventually seen as contemporaneous correlations when observing the market at longer time-scales. The cross-autocorrelation structure of the market thus emerges at small time-scales, where algorithmic trading strategies are more likely to operate. We obtain the same result when looking at the cross-autocorrelations of the assets of Group II.

The result of this analysis serves as a guideline for the empirical implementation of the MLA. In order to exploit more and more information sets, it is preferable to choose the highest available sampling frequency. In certain circumstances (e.g. when data are available at the millisecond precision) the computational power required for the estimation might be onerous, depending on the cross-section dimension. As underlined in Section (2.2) in the paper, a feasible solution is to reduce the intraday estimation window. Given the large amount of high-frequency data available even on sub-periods of the trading day, the choice of a shorter window does not affect significantly the quality of the inference.

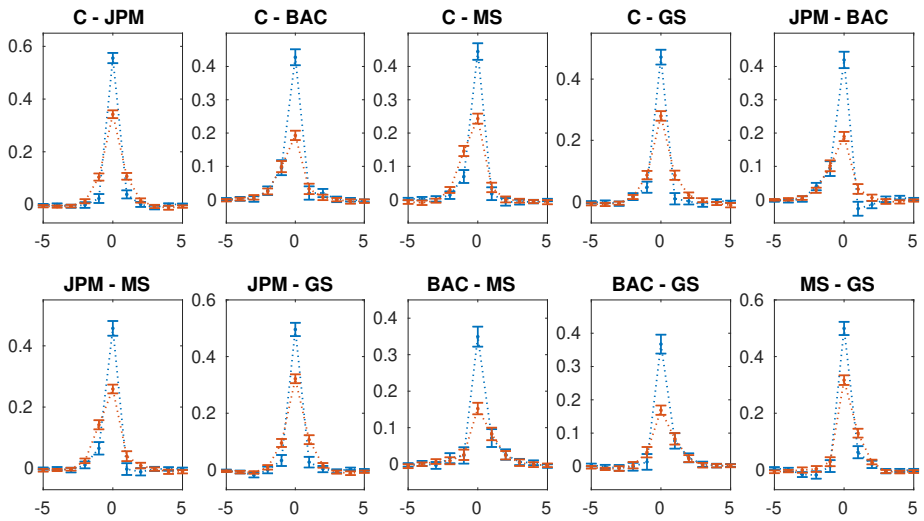


Figure 18: Average cross-autocorrelations of all the couples of assets in Group I. Averages are computed over all the business days of 2012. We show in blue the lead-lag correlations computed at the sampling frequency of 10 seconds and in red the lead-lag correlations computed at the sampling frequency of one second. Error bars denote 95% confidence intervals. Correlations at positive lags imply that the second asset displayed in the title leads the first asset and the other way around for negative lags.

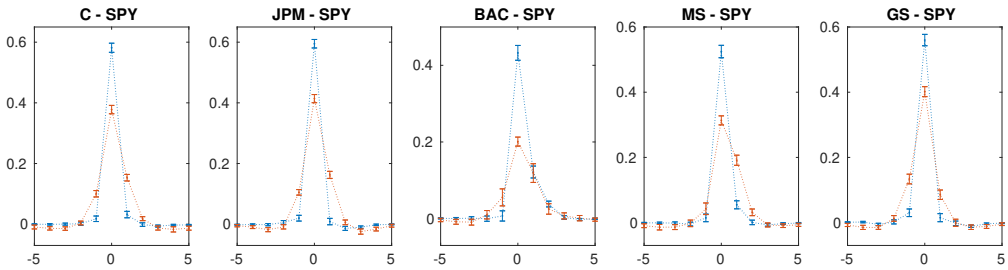


Figure 19: Average cross-autocorrelations between SPY and the stocks of Group I. Averages are computed over all the business days of 2012. We show in blue the lead-lag correlations computed at the sampling frequency of 10 seconds and in red the lead-lag correlations computed at the sampling frequency of 1 second. Error bars denote 95% confidence intervals. Correlations at positive lags imply that the second asset displayed in the title leads the first asset and the other way around for negative lags.

C.3 Potential diurnal effects in cross-asset trading

Intraday covariances are characterized by well-known diurnal effects (cf. e.g. Andersen and Bollerslev, 1997, Tsay, 2005, Bibinger et al., 2014, Buccheri et al., 2019a). For instance, volatilities are larger at the beginning and at the end of the trading day while correlations tend to increase throughout the day. We thus wonder whether cross-asset trading exhibits similar intraday patterns. Of course, lead-lag effects are naturally influenced by the intraday pattern of covariances. We are instead interested in potential diurnal effects induced by the lagged matrix Ψ of lagged price adjustment.

In order to investigate the behavior of Ψ at the intraday level, we consider the same subsample used in the previous analysis and divide the trading day into three sub-periods, the first from 9:30 to 11:00, the second from 11:01 to 14:30 and the third from 14:31 to 16:00. In each of these sub-periods we estimate the MLA and recover three lagged adjustment matrices, $\Psi_t^{(1)}$, $\Psi_t^{(2)}$, $\Psi_t^{(3)}$, where t is a daily index going from 03-01-2012 to 28-12-2012. We compare each of these matrices with the matrix Ψ_t estimated in the entire day. Specifically, for $i = 1, 2, 3$, we consider the differences $\theta_t^{(i)} = \text{vec}(\Psi_t^{(i)} - \Psi_t)$ and test the null hypothesis H_0 that the $d^2 = 121$ elements of $\theta_t^{(i)}$ have mean equal to zero.

Table (6) shows the results of the one-sample t-test performed for each of the 121 time-series corresponding to the coefficients of $\theta_t^{(i)}$, for $i = 1, 2, 3$. In the first line we report the number of coefficients for which the null hypothesis H_0 is not rejected at the 5% confidence level. In the second line we report the average p-value of the 121 t-tests. It is immediate to note that the vast majority of the coefficients of $\Psi_t^{(i)}$ estimated in the three sub-periods of the trading day are statistically indistinguishable from the coefficients of Ψ_t estimated in the entire period. The coefficients for which H_0 is rejected have p-values that are slightly smaller than 5% and always larger than 1%. This result corroborates the assumption of a constant lagged adjustment matrix Ψ in Section (2) in the paper and shows that cross-asset trading does not change significantly during the trading day.

	9:30−11:00	11:01−14:30	14:31−16:00
N. of coefficients of $\theta_t^{(i)}$ for which H_0 is not rejected (out of 121)	108	114	98
Avg. p-value	0.85	0.87	0.84

Table 6: We report for each of the three sub-periods of the trading day the number of coefficients of $\theta_t^{(i)}$ for which H_0 is not rejected at the 5% confidence level. We also show the average p-value of the 121 t-tests performed on each time series.