Hidden leaders: Identifying high-frequency lead-lag structures in a multivariate price formation framework

Giuseppe Buccheri, Fulvio Corsi, Stefano Peluso*

March, 2017

Abstract

We model the empirically observed lead-lag dependencies among high-frequency returns by introducing a multivariate price formation process which generalizes standard univariate microstructure models of lagged price adjustment. This theoretical framework for multivariate price dynamics provides: (i) a unified statistical test for the presence of lead-lag structures in latent price processes and for the existence of a multivariate price formation mechanism; (ii) separate estimation for both contemporaneous and lagged dependencies; (iii) an unbiased estimator of the integrated covariance matrix of the efficient martingale price process that is robust to microstructure noise, asyncronicity and lead-lag dependencies. Extensive simulation analyses show the accuracy of the estimator in recovering the true lead-lag structure of the latent price process and how neglecting lagged dependencies causes severe distortions in the estimation of contemporaneous covariances. Finally, the empirical application to equity data provides information on non-trivial latent lead-lag relationships among high-frequency returns and thus evidences the existence of a multivariate price formation process.

JEL codes: C10; C32; C58; G14; D81

Keywords: price formation; lagged adjustment; lead-lag effects; microstructure noise; asynchronicity; quadratic covariation; EM algorithm; Granger causality

*Buccheri: Scuola Normale Superiore, giuseppe.buccheri@sns.it. Corsi: Department of Economics, Ca' Foscari University of Venice and City University of London, fulvio.corsi@unive.it. Peluso: Università Cattolica del Sacro Cuore, stefano.peluso@unicatt.it. We are particularly grateful for suggestions we have received from Roberto Renò, Giacomo Bormetti, Fabrizio Lillo, Loriana Pelizzon and participants to the 10th CFE Conference in Seville and the XVIII Workshop in Quantitative Finance in Milan.

1 Introduction

The dynamics of high-frequency asset prices is 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 particular in cross-security and cross-market pricing (see e.g. Foucault and Cespa, 2011 and references therein). However, while it has received some attention in the empirical finance literature (see e.g. de Jong and Nijman, 1997, Chiao et al., 2004 and Huth and Abergel, 2014), there is a lack of theoretical approaches aiming to provide formal definition of lead-lag effects among high-frequency asset prices and to describe them from a microstructure point of view. Moreover, compared to the case of low-frequency (e.g. daily) data, the estimation of both contemporaneous and lagged dependencies among assets traded at high-frequency is a more complex task. This is mainly due to two reasons: the first is that prices are contaminated by microstructure effects and the second is that assets are traded asynchronously. Both types of effects prevent the use of traditional multivariate techniques which would result in biased estimates of both contemporaneous[1] and lead-lag correlations.

Motivated by cross-asset pricing, i.e. the fact that dealers can rely on the prices of other securities when setting their quotes, we introduce a multivariate 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). This leads to a micro-founded model where lead-lag correlations among high-frequency returns simply arise as a consequence of the multivariate nature of the price formation process. The statistical inference we develop for this model allows to test for the presence of lead-lag correlations in the latent price process or, equivalently, for the existence of a non-trivial multivariate price generation mechanism. Interestingly, by separating the estimation of lead-lag correlations from contemporaneous correlations, we are able to obtain an estimate of the integrated covariance matrix of the efficient martingale process that is robust to microstructure noise, asynchronous trading and lead-lag effects.

[1] The so called Epps effect Epps (1979), i.e. the bias towards zero of (contemporaneous) sample correlations as the sampling frequency increases.

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, due to lagged dissemination of information and price smoothing by market dealers, the adjustment process is delayed. Hasbrouck and Ho (1987) introduced a lagged adjustment price process that allows to describe return autocorrelations at orders greater than one, as observed in real transaction data. We extend this idea to a multivariate framework by viewing the price formation process as a genuine multivariate process where information related to other assets affects the price discovery process of a given asset.

By doing so, we merge the market microstructure literature on lagged price adjustment with that on cross-asset pricing. The latter is a concept that 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 every security from observing all order flows in the market. Based on cross-asset learning, Foucault and Cespa (2011) described 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. Finally, 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).

We show that the econometric inference on our multi-asset lagged adjustment model can be conveniently cunducted by casting the process into a state-space representation. The transition equation is a VAR(1) process for the returns of the "adjusted" price while the observation equation incorporates microstructure effects as an additive noise term. Using the Kalman filter, estimation is performed through an EM algorithm, easily handling missing observations. Thus, asynchronicity is treated as a typical missing value problem, in a similar fashion to Corsi et al. (2015) and Peluso et al. (2015). In this approach the model is estimated using all the available data, avoiding standard synchronization schemes that may introduce spurious lead-lag correlations or destroy true short-term lead-lag effects. As a result, the

proposed estimator is robust to observed price asynchronicity and to differences in the level of trading activities.

Since in our framework 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 be interpreted as an evidence for the existence of a multivariate price formation mechanism. In addition, due to the latent VAR structure, the model can be generically used to test for the presence of "latent" Granger causality in multivariate noisy and asynchronously observed data. Hence, it is able to disentangle true Granger causality, i.e. latent lead-lag correlations arising from nonzero non-diagonal coefficients in the VAR matrix, from trivial lead-lag dependencies due to the combined presence of autocorrelation and contemporaneous correlation effects which, instead, are not associated to cross-asset pricing. Finally, through the Kalman filter and smoothing recursions, our approach recovers the filtered and smoothed estimates of the unobserved efficient martingale price process.

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 (2016). By modelling microstructure effects as an additive noise term and treating asynchronicity as a missing value problem, they provided robust estimates of the quadratic covariation of a Brownian semimartignale process that is observed asynchronously and with noise. Our model differs from these approaches in that it introduces a mechanism of price generation that is more realistic than the simple semimartingale plus noise model. Indeed, while the latter describes the strong negative first-order autocorrelations that is observed in high-frequency returns, our model is also able to capture (cross) autocorrelations at higher orders.

Up to our knowledge, the only attempt to provide a formal definition of lead-lag effects among high-frequency prices has been made by Hoffmann et al. (2013). Given a two-dimensional process (X_t, Y_t), they defined the lead-lag parameter as the value of the time-shift θ for which the shifted system $(X_t, Y_{t+\theta})$ is a semimartingale under a given filtration. A consistent estimator of θ is found by maximizing lead-lag correlations computed using a modified version of the Hayashi-Yoshida estimator (Hayashi and Yoshida, 2005). Thus, Hoffmann et al. (2013) focuses on the estimation of the time-shift parameter rather

than the estimation of lead-lag correlations. Indeed, as pointed out by Huth and Abergel (2014), the Hayashi-Yoshida estimator used in computing cross autocorrelations is affected by spurious correlations, i.e. nonzero correlations arising as a result of asynchronicity even in absence of true lead-lag dependencies in the data generating process (DGP). Instead, as shown in Section (3.1), our estimator is robust to spurious correlations and correctly captures the true cross correlogram. A further advantage of our methodology is that, in contrast to the pairwise Hayashi-Yoshida estimator, it can be applied to a generic multivariate time series with cross-section dimension $d \geq 2$.

We test for the presence of lead-lag effects among high-frequency financial returns on a cross-section of NYSE tick data: in agreement with our proposed view of the multivariate nature of the price formation process, we find strong evidence of lead-lag effects among assets belonging to the same business sector or assets belonging to related sectors. For instance, we find that oil companies drive the dynamics of assets belonging to the energy and transport sectors while in the case of the banking sector it is possible to identify a leader that drives the dynamics of all the other assets.

The rest of the paper is organized as follows: In Section 2 we introduce the model and discuss the estimation problem using the EM algorithm; in Section 3 we make a comparison with other estimators and perform extensive Monte-Carlo simulations under misspecified DGP's; in Section 4 we apply the estimator to real transaction data and finally Section 5 concludes.

2 Theoretical framework

2.1 A multi-asset lagged adjustment model

A significant part of the empirical research on market microstructure has been devoted to understanding the autocorrelation structure of univariate and multivariate high-frequency return series. There is well-established evidence of three key empirical properties: strong negative first-order autocorrelation in the return series, existence of positive autocorrelation at lags greater than one and, finally, existence of lead-lag correlations. Simple bid-ask models

such as the model of Roll (1984) reproduce the negative first-order autocorrelation in the return series. Univariate bid-ask models were later generalized to capture correlations at orders greater than one through the introduction of lagged price adjustments (Hasbrouck, 1996). Here, we consider a multivariate version of a model with lagged price adjustment that is also able to keep into account lead-lag correlations.

We assume that the efficient log-price P_t is a d-dimensional vector that evolves as 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. $\tau = 1$ sec. in our empirical application). Note that, because of asynchronous trading, only the components of Y_{t_i} corresponding to traded assets are observed at time t_i, $i = 1, \ldots n$. This implies that there are missing values in the time-series of observed log-prices. In order to simplify the model, we assume that the drift term in eq. (1) is zero[2]. 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. Instead, 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 pricing, dealers tend to look at more informative securities before setting their quotes. In order to capture lagged dissemination of information

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

across stocks, we start from the simple univariate lagged adjustment mechanism proposed by Hasbrouck and Ho (1987) and adapt it to the multivariate 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:

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

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 (2016), the noise covariance matrix H is assumed to be diagonal. Denoting by $\Delta X_{t_{i+1}} \equiv X_{t_{i+1}} - X_{t_i}$ the log-return 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 to forecast the return of another asset at time t_{i+1}. Therefore, in this multivariate framework lead-lag effects naturally arise as a consequence of the mutual influence between adjustment processes of different assets.

Let us assume, without loss of generality, $\tau = 1$ 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 while eq. (7) is a transition equation describing the dynamics of latent returns at discrete times. As a consequence, model (7) cannot be estimated as a standard VAR model since X_t is not observed. We will show in Section (2.2) how to make inference on model (7) based on the observed transaction prices Y_t.

The assumption of a constant instantaneous 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 (see e.g Andersen and Bollerslev, 1997, Tsay, 2005, Bibinger et al., 2014, Buccheri et al., 2017). However, by performing extensive Monte-Carlo simulations and using a misspecified DGP with a time-varying covariance matrix Σ_t, we will show (see Section 3.2) two important properties. First, the estimate $\hat{F}$ of the matrix F of lead-lag coefficients remains unbiased even in presence of time-varying covariances. Second, $\hat{\Sigma} = \hat{\Psi}^{-1}\hat{Q}\hat{\Psi}'^{-1}$ is an unbiased estimator of $\frac{1}{T}QV$, where QV is the quadratic covariation of the efficient log-price process:

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

This result is similar to the one obtained by Shephard and Xiu (2016) who derived the asymptotic theory for the QML estimator of the integrated covariance of a Brownian semimartingale observed under noise and asynchronicity but neglecting lead-lag effects. Thus, our methodology provides an estimator of the quadratic covariation of a Brownian semimartingale that is robust to microstructure noise, asynchronicity and lead-lag effects.

Finally, note that making inference on model (7) can be regarded as testing for one-lag Granger causality in the latent process X_t, where the lag is $\tau = 1$ second. Testing for higher order lags results 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 method is to consider observations sampled at a smaller frequency. For instance, one can consider 2 second returns and apply model (7) in order to test for one-lag Granger causality with $\tau = 2$ seconds. Of course, this procedure implies a loss of efficiency since one is not using all the available information but avoids estimating more complex higher order

VAR(p) models.

2.2 Estimation

Statistical inference on model (6), (7) is conveniently performed by writing the two equations in a linear and Gaussian state-space representation. This is possible if one introduces the $2d-$dimensional state vector $\bar{X}_t = (X_t', X_{t-1}')'$ and re-writes the two equations as:

$$Y_t = M\bar{X}_t + \epsilon_t, \quad \epsilon_t \sim \text{NID}(0, H) \tag{9}$$

$$\bar{X}_t = \phi\bar{X}_{t-1} + \bar{\eta}_t, \quad \bar{\eta}_t \sim \text{NID}(0, \bar{Q}) \tag{10}$$

where:

$$\phi \equiv \begin{pmatrix} \mathbb{I}_d + F & -F \\ \mathbb{I}_d & \mathbf{0} \end{pmatrix}, \quad \bar{Q} \equiv \begin{pmatrix} Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{11}$$

with $M = (\mathbb{I}_d, \mathbf{0})$ being a known matrix that selects the first d components of $\bar{X}_t$ and $\mathbf{0}$ denoting a $d \times d$ matrix of zeros. We generically denote as Ω the set of parameters that we want to estimate, that is $\Omega = \{F, Q, H\}$.

Model (9), (10) is a linear and 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, see e.g. Durbin and Koopman (2012) and Shumway and Stoffer (2015). Instead of numerically optimizing the log-likelihood function using the Newton-Raphson method, we use the EM algorithm which does not require the computation of the inverse of the Hessian matrix. Indeed, the latter can be a quite large matrix in multivariate problems like the one we are considering here. For more details on the EM algorithm see e.g. A. P. Dempster (1977) and Shumway and Stoffer (1982).

As a first step, we assume there are no missing observations. We will show how to handle missing observations at the end of this section. We denote by $\mathcal{X}_n = \{\bar{X}_0, \ldots, \bar{X}_n\}$ 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 $\bar{X}_0 \sim \text{N}(\mu, \Sigma)$. Note that, since the knowledge of $\bar{X}_{t-1}$ completely determines the last d components of $\bar{X}_t$, the density function $f(\bar{X}_t|\bar{X}_{t-1})$ can be written as:

$$f(\bar{X}_t|\bar{X}_{t-1}) = f(M\bar{X}_t|\bar{X}_{t-1}) \tag{12}$$

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

$$\begin{aligned}\log L = \text{const} &- \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\bar{X}_0 - \mu)'\Sigma^{-1}(\bar{X}_0 - \mu) \\ &- \frac{n}{2}\log|Q| - \frac{1}{2}\sum_{t=1}^{n}(\bar{X}_t - \phi\bar{X}_{t-1})'M'Q^{-1}M(\bar{X}_t - \phi\bar{X}_{t-1}) \\ &- \frac{n}{2}\log|H| - \frac{1}{2}\sum_{t=1}^{n}(Y_t - M\bar{X}_t)'H^{-1}(Y_t - M\bar{X}_t)\end{aligned} \tag{13}$$

Of course, one cannot maximize the complete log-likelihood to obtain the MLE of Ω since $\mathcal{X}_n$ is not observed. 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 that can be recovered as an output of the Kalman filter and smoothing recursions (see Appendix (C)):

$$\bar{X}_t^s = \mathrm{E}[\bar{X}_t|\mathcal{Y}_s] \tag{14}$$

$$\bar{P}_t^s = \mathrm{Cov}[\bar{X}_t|\mathcal{Y}_s] \tag{15}$$

$$\bar{P}_{t,t-1}^s = \mathrm{Cov}[\bar{X}_t, \bar{X}_{t-1}|\mathcal{Y}_s] \tag{16}$$

With $s = t$, $s < t$ and $s > t$, the resulting conditional expectation is, respectively, a filter, a predictor and a smoother. The Kalman filter is initialized with diffuse initial conditions, i.e. we set $\mathbb{E}[\bar{X}_1|Y_1] = 0$ and $\mathrm{Cov}[\bar{X}_1|Y_1] = \kappa\mathbb{I}_d$ 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 Ω obtained at step $r-1$:

$$\begin{aligned}\mathrm{E}[\log L|\mathcal{Y}_n, \hat{\Omega}_{r-1}] = &-\frac{1}{2}\log|\Sigma| - \frac{1}{2}\mathrm{Tr}[\Sigma^{-1}[(\bar{X}_0^n - \mu)(\bar{X}_0^n - \mu)' + \bar{P}_0^n] \\ &- \frac{n}{2}\log|Q| - \frac{1}{2}\mathrm{Tr}[M'Q^{-1}M(C - B\phi' - \phi B' + \phi A\phi')] \\ &- \frac{n}{2}\log|H| - \frac{1}{2}\mathrm{Tr}[H^{-1}\sum_{t=1}^{n}[(Y_t - M\bar{X}_t^n)(Y_t - M\bar{X}_t^n)' + M\bar{P}_t^n M']\end{aligned} \tag{17}$$

where A, B and C are given by:

$$A = \sum_{t=1}^{n} (\bar{P}_{t-1}^{n} + \bar{X}_{t-1}^{n} \bar{X}_{t-1}^{n\prime}) \tag{18}$$

$$B = \sum_{t=1}^{n} (\bar{P}_{t,t-1}^{n} + \bar{X}_{t}^{n} \bar{X}_{t-1}^{n\prime}) \tag{19}$$

$$C = \sum_{t=1}^{n} (\bar{P}_{t}^{n} + \bar{X}_{t}^{n} \bar{X}_{t}^{n\prime}). \tag{20}$$

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:

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

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:

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

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

Proposition 1. *The solution of the matrix equation* $\nabla_F G_1(F,Q) = 0$ *is:*

$$\hat{F}_r = \Gamma\Theta^{-1} \tag{22}$$

where $\Gamma = B_{11} - B_{12} - A_{11} + A_{12}$ *and* $\Theta = A_{11} + A_{22} - A_{12} - A_{21}$.

The estimated value of F is then used to solve $\nabla_Q G_2(\hat{F}_r, Q) = 0$. In Appendix (B), (C) we prove the following proposition:

Proposition 2. *The solution of the two matrix equations* $\nabla_Q G_2(\hat{F}_r, Q) = 0$, $\nabla_H G_3(H) = 0$ *is:*

$$\hat{Q}_r = \frac{\hat{\Upsilon}}{n}, \quad \hat{H}_r = \frac{diag(\Lambda)}{n} \tag{23}$$

where $\hat{\Upsilon} = M(C - B\hat{\phi}_r' - \hat{\phi}_r B' + \hat{\phi}_r A \hat{\phi}_r')M'$, $\Lambda = \sum_{t=1}^{n}[(Y_t - M\bar{X}_t^n)(Y_t - M\bar{X}_t^n)' + M\bar{P}_t^n M']$ *and*

$$\hat{\phi}_r = \begin{pmatrix} \mathbb{I}_d + \hat{F}_r & -\hat{F}_r \\ \mathbb{I}_d & \mathbf{0} \end{pmatrix} \tag{24}$$

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: when it is lower than some fixed tolerance we stop the algorithm and take the last update of Ω as our MLE. The log-likelihood can be computed in the prediction error decomposition form:

$$\log L = \text{const} - \frac{1}{2}\sum_{t=1}^{n} \log |F_t| - \frac{1}{2}\sum_{t=1}^{n} v_t' F_t^{-1} v_t \tag{25}$$

where $v_t = Y_t - M\bar{X}_t^{t-1}$ is the prediction error and $F_t = M\bar{P}_t^{t-1}M' + H$ is its covariance matrix.

The consistency and asymptotic normality of the MLE of linear state-space models are studied under very general conditions by Douc et al. (2013). The two most important conditions needed for the MLE to be consistent and asymptotic normal are stability and identification. The first property holds whenever the eigenvalues of F are inside the unit circle while the second property is satisfied by model (9), (10) since M is fixed. Let us denote by $\hat{\Omega}_n$ the estimator obtained by maximizing the log-likelihood (25) and by Ω_0 the true parameter. Under the additional assumption that the model is Gaussian, theorem (6.4) in Shumway and Stoffer (2015) states that the MLE $\hat{\Omega}_n$ of Ω_0 is consistent and, as $n \to \infty$:

$$\sqrt{n}(\hat{\Omega}_n - \Omega_0) \xrightarrow{d} \mathrm{N}[0, \mathcal{I}(\Omega_0)^{-1}] \tag{26}$$

where $\mathcal{I}(\Omega_0)$ is the asymptotic Fisher information matrix. Note that this theorem holds in case the parameters in Ω are constant over time. As underlined above, in Section (3.2) we will show, using extensive Monte-Carlo simulations, that the MLE of F remains unbiased in case the covariance matrix Q evolves randomly over time. In practical applications, the Fisher information matrix can be approximated by numerically computing the Hessian of the likelihood function.

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

$$\hat{\Psi} = \mathbb{I}_d - \hat{F}, \quad \hat{\Sigma} = \hat{\Psi}^{-1}\hat{Q}\hat{\Psi}'^{-1} \tag{27}$$

Moreover, using the Kalman filter, one can recover filtered and smoothed estimates of the lagged price X_t and thus, using Eq. (3), one also obtains as a byproduct filtered and smoothed estimates of the latent efficient log-price process.

2.3 Missing value modification

The update formulas in the maximization step can be modified to keep 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 Appendix (C) and the prediction error decomposition form of the log-likelihood, eq. (25) are still valid, provided that one replaces Y_t, M and H with:

$$Y_{(t)} = \begin{pmatrix} Y_t^{(1)} \\ \mathbf{0} \end{pmatrix}, \quad M_{(t)} = \begin{pmatrix} M_t^{(1)} \\ \mathbf{0} \end{pmatrix}, \quad H_{(t)} = \begin{pmatrix} H_t^{(11)} & \mathbf{0} \\ \mathbf{0} & \mathbb{I}^{(22)} \end{pmatrix} \tag{28}$$

where $\mathbb{I}^{(22)}$ is the $d_2 \times d_2$ identity matrix and $\mathbf{0}$ 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 expectation in eq. (13) requires some modifications in case of missing observations. The second and the fourth term remain as in eq. (17), provided that one runs Kalman filter and smoothing recursions as described in (28). 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:

$$\mathbb{E}[(Y_t - M\bar{X}_t)(Y_t - M\bar{X}_t)'|\mathcal{Y}_n^{(1)}] = (Y_{(t)} - M_{(t)}\bar{X}_t^n)(Y_{(t)} - M_{(t)}\bar{X}_t^n)' + M_{(t)}\bar{P}_t^n M_{(t)}' + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \hat{H}_{22,t,r-1} \end{pmatrix} \tag{29}$$

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:

$$\hat{H} = \frac{\text{diag}(\Lambda^*)}{n} \tag{30}$$

where

$$\Lambda^* = \sum_{t=1}^{n} D_t \left[(Y_{(t)} - M\bar{X}_t^n)(Y_{(t)} - M\bar{X}_t^n)' + M_{(t)}\bar{P}_t^n M_{(t)}' + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \hat{H}_{22,t,r-1} \end{pmatrix} \right] D_t', \tag{31}$$

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

3 Simulation study

3.1 Comparison with other estimators

As underlined in the introduction, non-synchronous trading can have deep consequences when one wants to make inference on multivariate high-frequency tick-by-tick data. For instance, the Epps effect is certainly in part due to asynchronicity (see e.g. Epps, 1979 and Hayashi and Yoshida, 2005).

Non-synchronous trading is also responsible for the appearance of lead-lag cross correlations between stock returns characterized by different levels of trading activity. Indeed, some assets seem to lead other assets simply because, being traded more frequently, they are more likely to show the effect of new information arriving on the market before other assets which are traded less frequently. Thus, this effect is only due to differences in liquidity and not to true lead-lag dependencies. Even in presence of similar levels of trading activities, one can find spurious nonzero lead-lag correlation as a consequence of asynchronicity. Finally, even in synchronous time series lead-lag correlations which are not related to the arrival of new information can arise just because assets are autocorrelated and instantaneously correlated.

A detailed analysis of the impact of asynchronicity on lead-lag cross correlation estimation is performed by Huth and Abergel (2014). They 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 series by applying the Hayashi-Yoshida (HY) estimator (Hayashi and Yoshida, 2005) after shifting the timestamps of one of the two 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 lead-lag cross-correlation 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 even in absence of true lead-lag dependencies.

Our estimator is robust to both effects. The reason is that asynchronicity is handled as resulting from missing values in the observed time series and these can be easily incorporated into the EM algorithm without introducing any bias in the estimation of the parameters of the model.

In order to see this, we sample two Brownian motions over a time grid of $T = 10.000$ equally spaced points. The correlation between the two Brownian motions is $\rho = 0.4$. Asynchronicity is reproduced by censoring the simulated observations using Poisson sample. The probability of missing values is set equal to $\Lambda_1 = 0.3$ for the first series and $\Lambda_2 = 0.5$ for the second series. We repeat the experiment 250 times and for each realization we estimate lead-lag correlations by estimating model (6), (7) and using the HY estimator. Figure (1) shows the cross-correlogram obtained by averaging all the estimated correlations. 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 asynchronicity. Instead, the EM estimator is not affected by asynchronicity and correctly reproduces the correlogram of simulated data.

Our definition of lead-lag effects is formally different from the one of Hoffmann et al. In our microstructure framework, lead-lag effects arise as a consequence of nonzero non-diagonal

coefficients in the adjustment matrix Ψ. Instead, Hoffman et al. considered a continuous-time bivariate process (X_t, Y_t) and looked for some time shift θ such that the process $(X_t, Y_{t+\theta})$ is a semi-martingale with respect to some filtration. In order to understand how our estimator behaves in 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 series. Figure (2) shows the correlogram of the new bivariate time-series and those estimated by the HY and EM estimators. The HY estimator correctly estimates the lagged cross correlation but provides nonzero correlations at other leads and lags. Instead, the EM estimator correctly captures the true cross-correlogram. Indeed, the shifted time series can be written as a VAR(1) process with one nonzero non-diagonal element in F and uncorrelated disturbances. In case $\theta > 1$, one can sample observations at a lower frequency and still use the EM estimator.

Finally, there are other two noteworthy differences to be mentioned. First, the EM estimator is robust to microstructure noise in that it allows observations of the underlying process X_t to be affected by additive noise. Instead, the HY estimator ignores the presence of microstructure noise in the data. Second, while the HY estimator must be applied separately for each couple of time-series, model (6), (7) can be estimated for 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. 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:

$$\Sigma_t = D_t R D_t \tag{32}$$

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

$$D_t = \begin{pmatrix} \sigma_{t,1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{t,d} \end{pmatrix} \tag{33}$$

The diagonal entries in D_t evolve through a CIR stochastic volatility model:

$$d\sigma_{t,i}^2 = k(\theta_i - \sigma_{t,i}^2)dt + w_i \sigma_{t,i} dB_t, \quad i = 1, \ldots, d \tag{34}$$

with $\mathbb{E}[dW_t dB_t'] = \rho dt$. The assumption of a constant correlation matrix in not restrictive. Indeed, we also performed extensive simulations with time-varying correlations and verified that this does not affect the outcome of the experiment.

We report here the results obtained in the bivariate case ($d = 2$). The length of the trading day is assumed to be $T = 6.5$ hours and therefore we simulate 1 second time-series of $n = 23400$ timestamps. We discretize the CIR process in the Euler scheme and draw the first observation from a Gamma distribution $\Gamma(2k\theta_i/w_i^2, w_i^2/2k_i)$ centred in the mean variance. The parameters of the DGP are chosen in the following way: $\theta_1 = 0.01$, $\theta_2 = 0.02$, $w_1 = w_2 = 0.1$, $k_1 = 10$, $k_2 = 7$. The matrix F of lead-lag coefficients and the leverage matrix ρ are chosen as:

$$F = \begin{pmatrix} 0.1 & 0.5 \\ 0.3 & 0.1 \end{pmatrix}, \quad \rho = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.1 \end{pmatrix} \tag{35}$$

The variances h_i, $i = 1, 2$ of the noise are computed based on the average signal-to-noise ratio $\bar{\delta}_i = \theta_i / h_i$. In order to mimic realistic noise scenarios, we choose $\bar{\delta} = 0.5, 1, 2$. Indeed, these are the values that are typically observed in real markets (see Section 4). We consider both the cases where observations are synchronous and with missing values. In the latter case, simulated observations are censored using Poisson sampling. The probability of having a missing values is set equal to $\Lambda = 0.5$ for both series.

We estimate model (6), (7) for $N = 1000$ independent realizations and consider the pivotal statistics $\hat{\theta}_F^i = \hat{F}^i - F$ and $\hat{\theta}_\Sigma^i = \hat{\Sigma}^i - \frac{1}{T}QV$, for $i = 1, \ldots, N$ and where QV is the quadratic covariation defined in Eq. (8). The three scenarios $\bar{\delta} = 0.5, 1, 2$ are combined to each of 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 element of two matrices $\hat{\theta}_\Sigma$ and $\hat{\theta}_F$ and perform a one sample t-test whose p-value is also reported in the table. The distribution of $\hat{\theta}_\Sigma$ and $\hat{\theta}_F$ is always centred in 0. This implies that, in case of time-varying covariances, $\hat{F}$ correctly estimates the true matrix F of lead-lag coefficients while $\hat{\Sigma}$ correctly estimates the quadratic covariation of the efficient log-price process.

In figures (3), (4) and (5) the two scenarios $\Lambda = 0, \bar{\delta} = 1$ and $\Lambda = 0.5, \bar{\delta} = 1$ are considered. We plot histograms of each element of $\hat{\theta}_\Sigma$ and $\hat{\theta}_F$ after normalizing by their sample standard deviations. In the first scenario where the data are synchronous, the histograms are perfectly compatible with a standard normal distribution. Instead, in the second scenario we observe slight deviations from the normal. This is due to the fact that censoring halves, on average, the number of observations and thus the estimator is less efficient.

Finally, in figure (6) we plot the same histograms as in figure (5a), (5b), (5c) but now we use the Hayashi-Yoshida estimator in order to compute the quadratic covariation. As can be seen, the Hayashi-Yoshida estimator is largely biased in case lead-lag effects are present. We have verified[3] that a similar bias is observed also 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), the KEM estimator of Corsi et al. (2015) and the QMLE estimator of Shephard and Xiu (2016).

4 Empirical application

We assess the existence of lead-lag effects in high-frequency financial data by estimating model (6), (7) on a cross-section of NYSE stock prices. Our dataset is provided by Thomson Reuters and contains intraday transaction data of the 250 most liquid assets traded in the NYSE in 2014. Exchanges open at 9.30 and close at 16.00 local time, so that the number of seconds per day is $T = 23400$. We build a time-grid of T timestamps by taking the median of trades occurring at the same second, as suggested by O. E. Barndorff-Nielsen (2009). Outliers are removed using the non-parametric filter of the spot volatility described in Corsi et al. (2010) in appendix B. We identify as outliers those returns that are larger in absolute

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

value than some multiple of the spot volatility estimate provided by the filter.

Among the universe of 250 assets, we select the 11 assets listed in table (2) which also shows the probability of missing values Λ, the average duration $\overline{\Delta t}$ in seconds between observations, the average number of observations per day $\bar{n}$ and the average signal-to-noise ratio $\bar{\delta}$ as estimated in our model. Note that five of the selected assets, namely {C, JPM, BAC, MS, GS} belong to the banking sector while the remaining six, namely {XOM, CVX, SLB, GM, COP, GE} belong to the oil, energy and transport sectors. We name the two sets of assets as "Group I" and "Group II", respectively.

The analysis is carried by estimating the model on all the $N = 252$ business days of 2014. One can estimate the j-th order autocovariance matrix $S_j \equiv \mathbb{E}[\Delta X_t \Delta X'_{t-j}]$ from the estimated matrices $\hat{F}^i$ and $\hat{Q}^i$, $i, \ldots, N$ as:

$$\hat{S}^i_j = \hat{F}^i_j \hat{S}^i_{j-1}, \quad j = 1, 2, \ldots \tag{36}$$

where the covariance matrix $S_0 \equiv \mathbb{E}[\Delta X_t \Delta X'_t]$ is computed as:

$$\text{vec}(\hat{S}^i_0) = (\mathbb{I}_{d^2} - \hat{F}^i \otimes \hat{F}^i)^{-1} \text{vec}(\hat{Q}^i) \tag{37}$$

see e.g. Hamilton (1994). In order to compute cross autocorrelations, we normalize the autocovariances in $\hat{S}_j$ using the diagonal elements of $\hat{S}_0$.

In figures (7), (8) we plot the average, over all the days of the sample, of cross autocorrelations for every couple of assets in Group I and Group II, respectively. Bars denote 95% confidence intervals and are computed using the sample standard deviations. For a single estimate, errors can instead be evaluated by numerically computing the Hessian of the likelihood function, as underlined in Section (2.2). Correlations at positive lags imply that the second asset is leading the first while correlations at negative lags imply the opposite.

We find strong evidence of lead-lag effects in both groups. For instance, in Group I, Goldman Sachs appears to lead all other assets while Bank of America is lead by all the assets. Note that Goldman Sachs is the less traded asset of Group I. The fact that it leads the dynamics of the assets in its business sector means that this effect is not merely due to differences in liquidity but to true lead-lag correlations that emerge as a consequence of nonzero lead-lag coefficients in the matrix F.

In Group II we observe that oil companies like ExxonMobil, Chevron and Schlumberger lead the dynamics of energy and transport companies like General Electric and General Motors. Instead, we observe weaker lead-lag correlations among leaders (e.g. between XOM and CVX) and among laggers (e.g. between GE and GM).

We also find 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 (9) we show the estimated correlogram of GS-GM (9a) and the one of MS-CVX (9b). 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 zero correlations.

Note that lead-lag correlations can arise even if the non-diagonal elements of F are all zero, as a consequence of combined autocorrelation and contemporaneous correlation effects. This is not the case here, since the estimated matrix $\hat{F}$ has a lot of statistically significant non-diagonal elements, as shown in tables (3) and (4). This implies that the recovered lead-lag structures arise as a consequence of nonzero non-diagonal elements in the speed of adjustment matrix Ψ and therefore they can be viewed in our framework as an evidence of the existence of multivariate price formation mechanism. This result can be interpreted as a consequence of cross-asset pricing. 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 Ψ.

5 Conclusions

We have introduced a multivariate model of price generation that extends standard microstructure models of lagged price adjustment. Lead-lag effects are naturally incorporated in this framework through nonzero non-diagonal coefficients in the speed of adjustment matrix Ψ. The latter captures lagged dissemination of information across assets due to cross-asset pricing. The model can be cast into a state-space representation where the transition equation is a VAR(1) process for the returns of the adjusted price X_t, while the observation equation incorporates microstructure effects through an additive noise term. This state-space representation is conveniently estimated using the EM algorithm. Asynchronicity is treated

as a missing value problem and the latter is easily tackled by the Kalman filter. The resulting estimator of lead-lag correlations is thus robust to asynchronicity and microstructure noise. As a byproduct, we also obtain an estimate of the covariance matrix of the efficient log-price process that is robust to asyncronicity, microstructure noise and lead-lag dependencies.

Using extensive Monte-Carlo experiments, we have shown that, as opposed to other estimators of lead-lag correlations, our estimator is robust to spurious correlations arising from asynchronous trading. Also, we have tested the performance of the model in presence of misspecified DGP's and found that, when covariances are time-varying, $\hat{F}$ is still unbiased while $\hat{\Sigma}$ correctly estimates the integrated covariance of the efficient log-price process. The large sample distribution of both estimators is compatible with the Gaussian distribution. Finally, by estimating the model on real transaction data, we have found evidence of strong lead-lag effects among assets belonging to related industries, in support of our view of a multivariate price formation process.

Finally, due to the latent VAR(1) structure, our methodology can be viewed as a test for "Granger causality on latent processes", i.e. a Granger causality test on multivariate time-series of noisy and asynchronous observations. Thus, it could be of potential interest for a broad spectrum of empirical applications.

	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}	0.1950	2.6117	**0.0184**	0.0725	3.0954	**0.4589**	0.4159	3.3105	**0.5314**
F_{12}	−0.2393	2.5798	0.0034	−0.0857	3.2803	**0.4090**	−0.0616	3.4449	**0.5717**
F_{21}	0.0812	2.2679	**0.2580**	0.0191	2.8767	**0.8341**	0.0002	3.2539	**0.9983**
F_{22}	−0.1742	2.9171	**0.0592**	−0.1233	3.6635	**0.2876**	0.0067	3.8570	**0.9563**
Σ_{11}	−0.0053	0.0879	**0.0588**	−0.0019	0.0861	**0.4937**	−0.0045	0.0873	**0.1043**
Σ_{12}	−0.0043	0.0416	0.0011	−0.0057	0.0500	0.0040	−0.0063	0.0563	**0.0521**
Σ_{22}	0.0083	0.1348	**0.0510**	−0.0038	0.1462	**0.4159**	−0.0043	0.1285	**0.2870**
	$\Lambda = 0.5$								
F_{11}	−0.1778	3.3155	**0.0903**	0.2465	3.7535	**0.0381**	0.1386	3.2338	**0.1756**
F_{12}	0.0212	3.4834	**0.8476**	−0.3053	3.8869	**0.0132**	−0.2190	3.3824	**0.0409**
F_{21}	−0.9099	5.2234	0.0002	−0.4047	3.9738	0.0075	−0.2804	3.3068	**0.0105**
F_{22}	0.8688	5.7858	0.0002	0.1713	4.7535	**0.2548**	0.1942	3.9592	**0.1211**
Σ_{11}	0.0068	0.0935	**0.0209**	−0.0014	0.0923	**0.6268**	0.0013	0.0949	**0.6696**
Σ_{12}	0.0028	0.0671	**0.1831**	0.0011	0.0650	**0.5876**	−0.0004	0.0550	**0.8036**
Σ_{22}	−0.0064	0.1473	**0.1723**	0.0030	0.1584	**0.5510**	9.81e-6	0.1471	**0.9983**

Table 1: *We report the sample average and standard deviation 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.*

Symbol	Λ	$\overline{\Delta t}$	$\overline{n}$	$\overline{\delta}$
XOM	0.816	5.434	4304	1.178
C	0.837	6.135	3832	1.246
JPM	0.840	6.250	3743	0.999
CVX	0.848	6.578	3553	0.850
SLB	0.853	6.802	3454	0.613
GM	0.866	7.462	3135	0.888
BAC	0.869	7.633	3079	0.328
COP	0.880	8.333	2828	0.494
GE	0.892	9.259	2543	0.641
MS	0.897	9.708	2416	0.741
GS	0.920	12.500	1873	0.630

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 number of observations per day $\bar{n}$ and the average signal-to-noise ratio as estimated in our model. The averages are computed over all the business days of 2014.*

	Group I				
	avg F_{ij}				
	C	JPM	BAC	MS	GS
C	0.0886****	0.0472****	0.0220*	-0.0635****	0.1276****
JPM	0.0318***	0.1023****	0.0065$^{(\mathrm{ns})}$	-0.0709****	0.1358****
BAC	0.0518****	0.0657****	0.0863****	0.0201$^{(\mathrm{ns})}$	0.1040****
MS	0.0752****	0.0973****	0.0107$^{(\mathrm{ns})}$	0.0193*	0.1542****
GS	0.0334**	0.0011$^{(\mathrm{ns})}$	-0.0031$^{(\mathrm{ns})}$	-0.0743****	0.1647****

Table 3: *We report the sample average, over the whole 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					
	avg F_{ij}					
	XOM	CVX	SLB	GM	COP	GE
XOM	0.0776****	0.0428***	0.0273****	-0.0143**	-0.0263$^{(\mathrm{ns})}$	-0.0408****
CVX	0.0232*	0.0981****	0.0155*	-0.0137$^{(\mathrm{ns})}$	-0.0327*	-0.0322***
SLB	0.0135$^{(\mathrm{ns})}$	0.0665***	0.0946****	-0.0251*	-0.0709****	-0.0355*
GM	0.0809****	0.0657**	0.0733****	0.0686****	-0.0035$^{(\mathrm{ns})}$	0.0182$^{(\mathrm{ns})}$
COP	0.0318**	0.0502***	0.0485****	-0.0114$^{(\mathrm{ns})}$	0.0531****	-0.0045$^{(\mathrm{ns})}$
GE	0.0321*	0.0542***	0.0424****	0.0137$^{(\mathrm{ns})}$	-0.0015$^{(\mathrm{ns})}$	0.0585****

Table 4: *We report the sample average, over the whole 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:* $^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$, $^{****}p \leq 0.0001$, $^{(ns)}p > 0.05$.

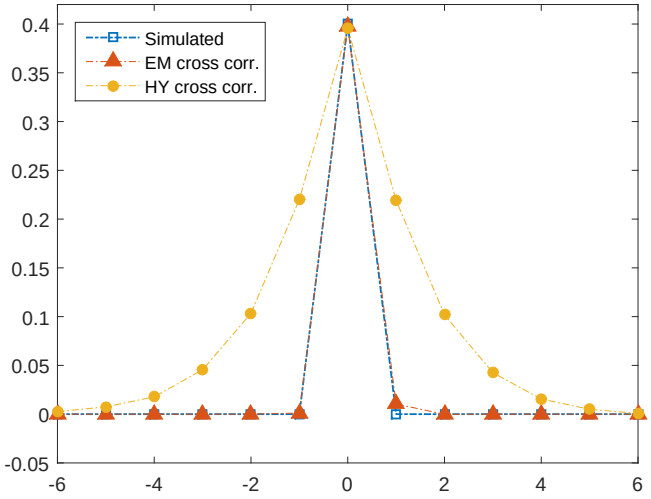


Figure 1: *Cross-correlogram of two simulated Brownian motions with correlation* $\rho = 0.4$ *observed asynchronously over* $T = 10.000$ *timestamps. The yellow line is the average estimates provided by the HY estimator while the red line is the one provided by the EM estimator over* $N = 250$ *independent realizations.*

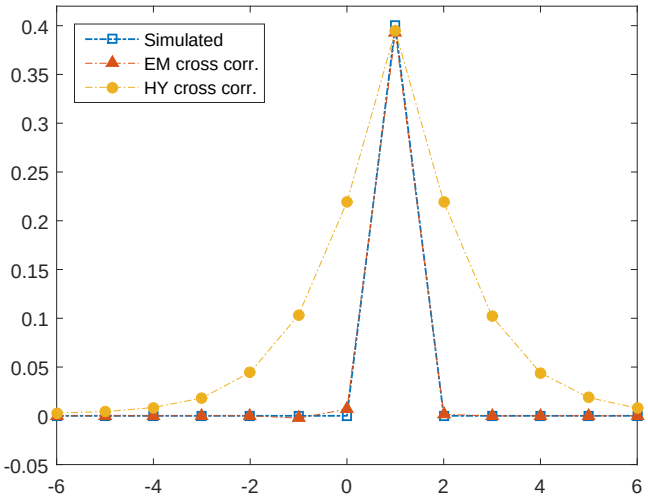


Figure 2: *As in Figure (1) but HY and EM are estimated over a time series obtained by shifting one of the two simulated Brownian motions. We show the average correlations provided by the HY (yellow line) and EM (red line) estimators and the cross-correlogram of the simulated time-series (blue line).*

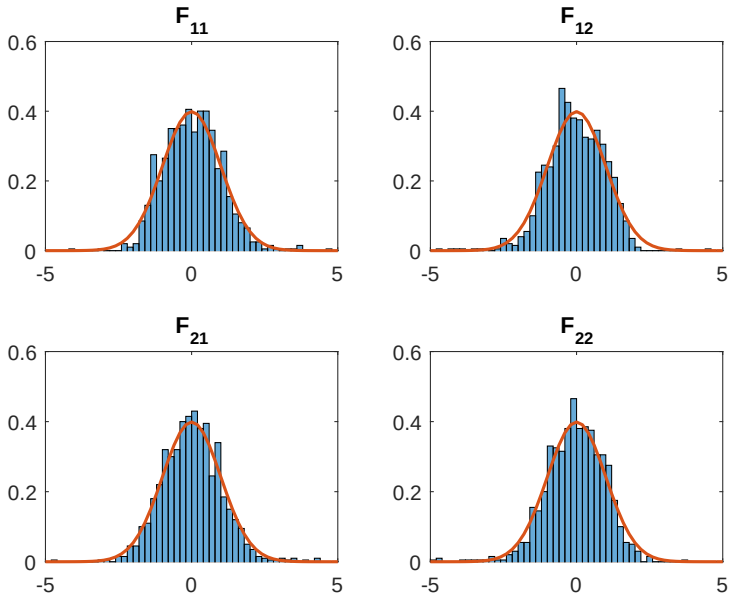


Figure 3: *Histograms of the elements of the matrix* $\hat{\theta}_F$ *standardized by their sample standard deviations in the scenario* $\bar{\delta} = 1, \Lambda = 0$ *over* $N = 1000$ *independent realizations. The red line is the standard normal distribution.*

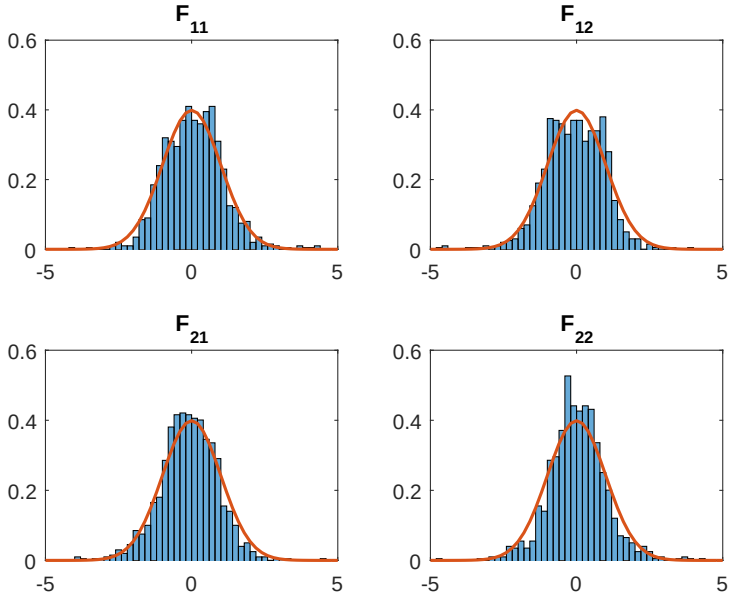


Figure 4: *Histograms of the elements of the matrix* $\hat{\theta}_F$ *standardized by their sample standard deviations in the scenario* $\bar{\delta} = 1, \Lambda = 0.5$ *over* $N = 1000$ *independent realizations. The red line is the standard normal distribution.*

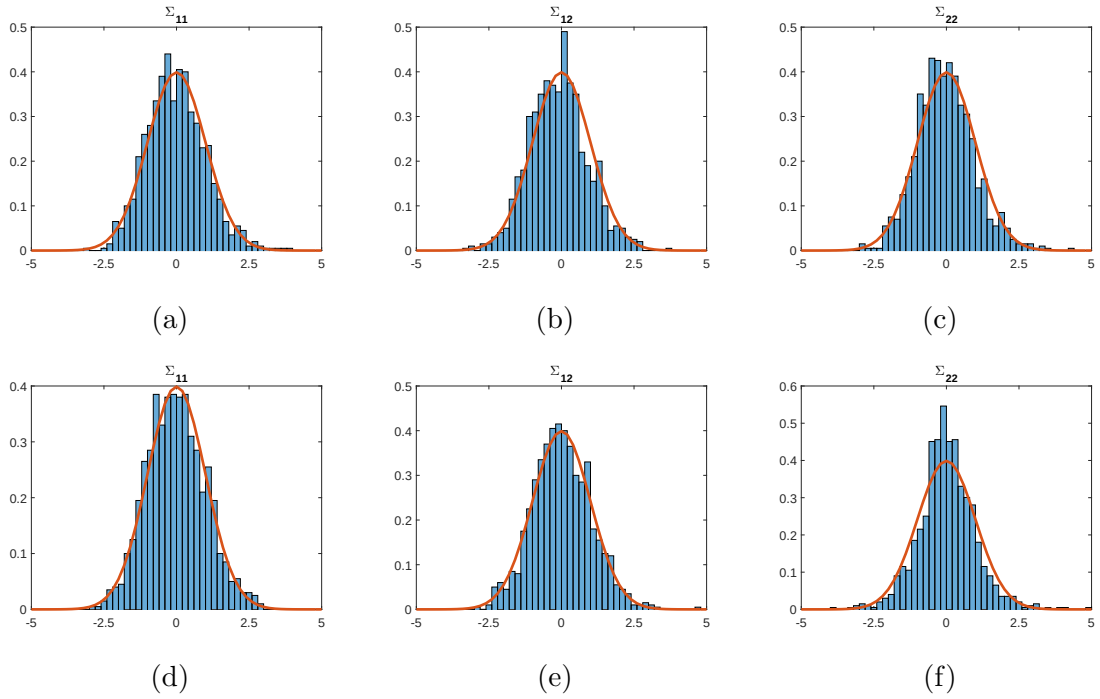


Figure 5: *Histograms of the elements of the matrix* $\hat{\theta}_Q$ *standardized by their sample standard deviations in the scenario* $\bar{\delta} = 1, \Lambda = 0$ *(a), (b), (c) and in the scenario* $\bar{\delta} = 1, \Lambda = 0.5$ *(d), (e), (f) over* $N = 1000$ *independent realizations. The red line is the standard normal distribution.*

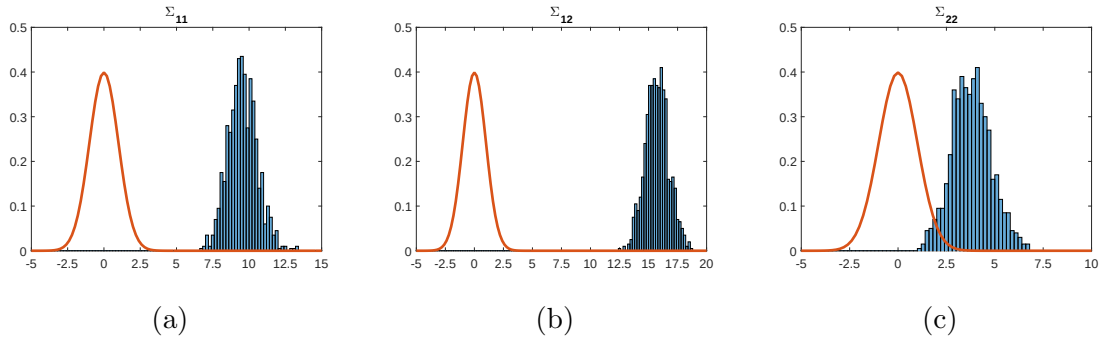


Figure 6: *Histograms of the elements of the matrix* $\hat{\theta}_Q$ *standardized by their sample standard deviations and computed using the Hayashi-Yoshida estimator in the scenario* $\bar{\delta} = 1, \Lambda = 0$ *over* $N = 1000$ *independent realizations. The red line is the standard normal distribution.*

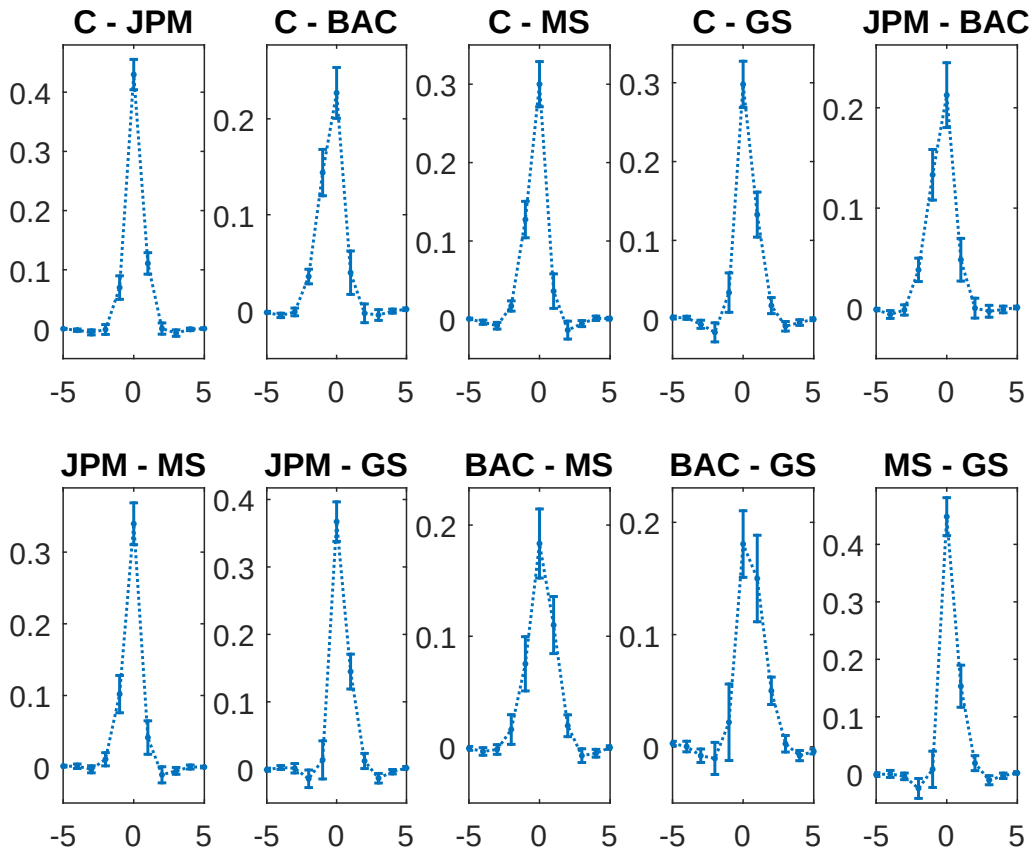


Figure 7: *Cross autocorrelations for all the couples of assets in Group I. The correlograms are computed by averaging those obtained by estimating the model on all the business days of 2014. Error bars denote* 95% *confindence intervals. Correlations at positive lags imply that the second asset displayed in the title is leading the first and the other way around for negative lags.*

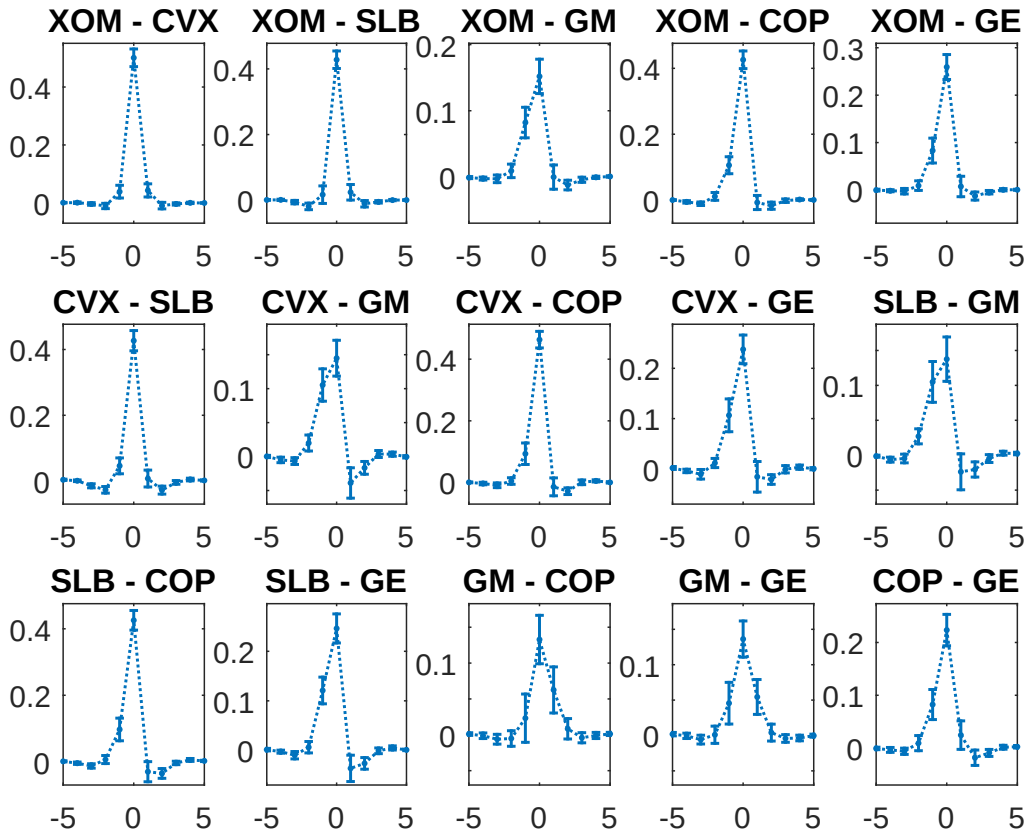
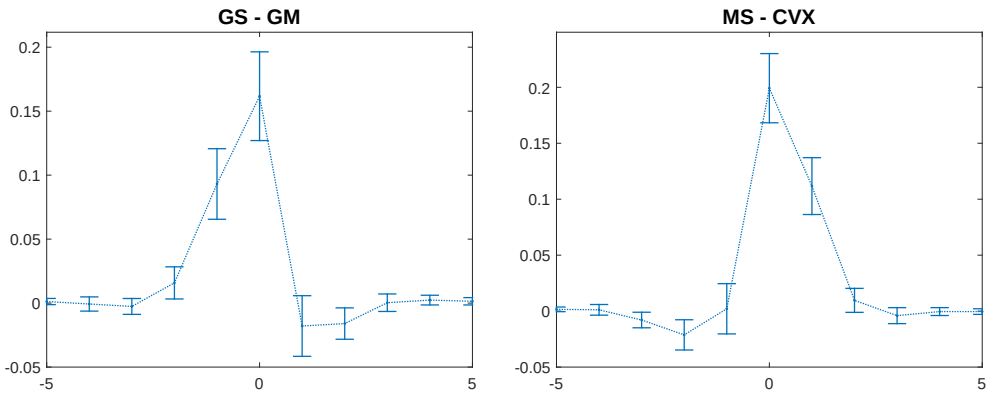


Figure 8: *Cross autocorrelations for all the couples of assets in Group II. The correlograms are computed by averaging those obtained by estimating the model on all the business days of 2014. Error bars denote* 95% *confindence intervals. Correlations at positive lags imply that the second asset displayed in the title is leading the first and the other way around for negative lags.*



(a) (b)

Figure 9: *Cross autocorrelations between stocks belonging to different groups.*

Appendix

A Proof of Proposition 1

We will use the following matrix differentiation rules:

$$\nabla_A \mathrm{tr}(AB) = B' \tag{A.1}$$

$$\nabla_A \mathrm{tr}(ABA'C) = CAB + C'AB' \tag{A.2}$$

$$\nabla_A |A| = |A|(A^{-1})' \tag{A.3}$$

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

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

$$G_1(F, Q) = -\frac{1}{2}\mathrm{Tr}[Q^{-1}(MCM' - \tilde{B}\tilde{\phi}' - \tilde{\phi}\tilde{B}' + \tilde{\phi}A\tilde{\phi}')] \tag{A.4}$$

where we have defined $\tilde{B} \equiv MB$ and $\tilde{\phi} \equiv M\phi$. 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:

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

This can be done using eq. (A.1) and (A.2). One obtains:

$$\begin{aligned}
\nabla_F \overline{G}_1(F) = Q^{-1}[&-2(B_{11} - B_{12} - A_{11} + A_{12}) \\
&+ 2F(A_{11} + A_{22} - A_{21} - A_{12})]
\end{aligned} \tag{A.5}$$

and therefore:

$$\hat{F} = \Gamma\Theta^{-1} \tag{A.6}$$

■

B Proof of Proposition 2

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

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

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

$$\hat{Q} = \frac{\hat{\Upsilon}}{n} \tag{B.2}$$

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

$$\begin{aligned} &\nabla_H G_3(H) = \\ &= \nabla_H \left[-\frac{n}{2} \log |H| - \frac{1}{2} \mathrm{Tr}(H^{-1} \mathrm{diag}(\Lambda)) \right] \\ &= \frac{n}{2} H - \frac{1}{2} \Lambda \end{aligned} \tag{B.3}$$

and therefore:

$$\hat{H} = \frac{\mathrm{diag}(\Lambda)}{n} \tag{B.4}$$

■

C Kalman filter and smoothing recursions

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

$$\bar{X}_t^{t-1} = \phi \bar{X}_{t-1}^{t-1} \tag{C.1}$$

$$\bar{P}_t^{t-1} = \phi \bar{P}_{t-1}^{t-1} \phi' + \bar{Q} \tag{C.2}$$

$$K_t = \bar{P}_t^{t-1} M' (M \bar{P}_t^{t-1} M' + H)^{-1} \tag{C.3}$$

$$\bar{X}_t^t = \bar{X}_t^{t-1} + K_t (Y_t - M \bar{X}_t^{t-1}) \tag{C.4}$$

$$\bar{P}_t^t = \bar{P}_t^{t-1} - K_t H \bar{P}_t^{t-1} \tag{C.5}$$

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

$$J_{t-1} = \bar{P}_{t-1}^{t-1} \phi' (\bar{P}_t^{t-1})^{-1} \tag{C.6}$$

$$\bar{X}_{t-1}^n = \bar{X}_{t-1}^{t-1} + J_{t-1} (X_t^n - \phi \bar{X}_{t-1}^{t-1}) \tag{C.7}$$

$$\bar{P}_{t-1}^n = \bar{P}_{t-1}^{t-1} + J_{t-1} (\bar{P}_t^n - \bar{P}_t^{t-1}) J_{t-1}' \tag{C.8}$$

for $t = n, \ldots, 1$. The covariance $\bar{P}^n_{t,t-1}$ in eq. (19) can be computed using the following backward recursion:

$$\bar{P}^n_{t-1,t-2} = \bar{P}^{t-1}_{t-1} J'_{t-2} + J_{t-1}(\bar{P}^n_{t,t-1} - \phi \bar{P}^{t-1}_{t-1}) J'_{t-2} \tag{C.9}$$

where $t = n, \ldots 2$ and $\bar{P}^n_{n,n-1} = (I - K_n M)\phi \bar{P}^{n-1}_{n-1}$.

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), 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 (492), 1504–1517.

Amihud, Y., Mendelson, H., 1987. Trading mechanisms and stock returns: An empirical investigation. The Journal of Finance 42 (3), 533–553.

Andersen, T., Bollerslev, T., 1997. Intraday periodicity and volatility persistence in financial markets. Journal of Empirical Finance 4 (2-3), 115–158.

Andrade, S., Chang, C., Seasholes, M., 5 2008. Trading imbalances, predictable reversals, and cross-stock price pressure. Journal of Financial Economics 88 (2), 406–423.

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 (2), 149 – 169.

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).

Buccheri, G., Bormetti, G., Corsi, F., Lillo, F., 2017. 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.

Caballé, J., Krishnan, M., 1994. Imperfect competition in a multi-security market with risk neutrality. Econometrica 62 (3), 695–704.

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 (5), 709 – 731.

Corsi, F., Peluso, S., Audrino, F., 2015. Missing in asynchronicity: A kalman-em approach for multivariate realized covariance estimation. Journal of Applied Econometrics 30 (3), 377–397.

Corsi, F., Pirino, D., Renò, R., 2010. Threshold bipower variation and the impact of jumps on volatility forecasting. Journal of Econometrics 159 (2), 276 – 288.

Damodaran, A., 1993. A simple measure of price adjustment coefficients. The Journal of Finance 48 (1), 387–400.

de Jong, F., Nijman, T., 1997. High frequency analysis of lead-lag relationships between financial markets. Journal of Empirical Finance 4 (2), 259 – 277.

Douc, R., Moulines, r., Stoffer, D., 2013. Nonlinear Time Series: Theory, Methods and Applications with R Examples. Chapman & Hall.

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

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

Foucault, T., Cespa, G., Apr. 2011. Dealer Attention, liquidity spillovers, and endogenous market segmentation. In: Séminaire de recherche, Rotterdam School of Management. Rotterdam, Netherlands.

Griffin, J. E., Oomen, R. C., 2011. Covariance measurement in the presence of non-synchronous trading and market microstructure noise. Journal of Econometrics 160 (1), 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 (1), 29–55.

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 (4), 1035–48.

Hasbrouck, J., Seppi, D. J., 2001. Common factors in prices, order flows, and liquidity. Journal of Financial Economics 59 (3), 383 – 411.

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

Hoffmann, M., Rosenbaum, M., Yoshida, N., 05 2013. Estimation of the lead-lag parameter from non-synchronous data. Bernoulli 19 (2), 426–461.

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

O. E. Barndorff-Nielsen, P. Reinhard Hansen, A. L. N. S., 2009. Realized kernels in practice: trades and quotes. The Econometrics Journal 12 (3), C1–C32.

Pasquariello, P., Vega, C., 2015. Strategic cross-trading in the u.s. stock market*. Review of Finance 19 (1), 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 (3), 665–697.

Roll, R., 1984. A simple implicit measure of the effective bid-ask spread in an efficient market. The Journal of Finance 39 (4), 1127–1139.

Shephard, N., Xiu, D., Dec. 2016. Econometric analysis of multivariate realised qml: Estimation of the covariation of equity prices under asynchronous trading. Chicago booth research paper no. 12-14.

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 (4), 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 (1), 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., 03 1983. On the convergence properties of the em algorithm. Ann. Statist. 11 (1), 95–103.