Missing in Asynchronicity: A Kalman-EM Approach for Multivariate Realized Covariance Estimation

Fulvio Corsi* Stefano Peluso† Francesco Audrino‡

Revised version: November 2013

Abstract

Motivated by the need of a positive-semidefinite estimator of multivariate realized covariance matrices, we model noisy and asynchronous ultra-high-frequency asset prices in a state-space framework with missing data. We then estimate the covariance matrix of the latent states through a Kalman smoother and Expectation Maximization (KEM) algorithm. Iterating between the two EM steps, we obtain a covariance matrix estimate which is robust to both asynchronicity and microstructure noise, and positive-semidefinite by construction. We show the performance of the KEM estimator using extensive Monte Carlo simulations that mimic the liquidity and market microstructure characteristics of the S&P 500 universe as well as in an high-dimensional application on US stocks. KEM provides very accurate covariance matrix estimates and significantly outperforms alternative approaches recently introduced in the literature.

JEL classification: C13; C51; C52; C58

Keywords: High frequency data; Realized covariance matrix; Missing data; Kalman filter; EM algorithm.

*Ca' Foscari University of Venice and City University London, E-mail: fulvio.corsi@unive.it

†Corresponding author. University of Lugano, Swiss Finance Institute, E-mail: stefano.peluso@usi.ch. Via Giuseppe Buffi 13 CH-6904 Lugano. Tel.: +41 (0)58 666 44 95, Fax: +41 (0)58 666 46 47

‡University of St. Gallen, E-mail: francesco.audrino@unisg.ch

1 Introduction

Modeling and forecasting the conditional covariance matrix of asset returns is pivotal to many prominent financial problems such as risk management, asset allocation, and option pricing. It is now well recognized that the proper use of intra-day price observations leads to precise and accurate measurement and forecasting of the unobservable asset volatility, the so-called realized volatility approach. The idea of realized volatility measures goes back to the seminal work of Merton (1980), who showed that the integrated variance of a Brownian motion can be approximated by the sum of a large number of intra-day squared returns. This original intuition has been recently formalized and generalized in a series of papers applying quadratic variation theory.[1] These results allow one to exploit all the information contained in intra-day high-frequency data in the construction of a volatility measure.

However, the multivariate extensions of the realized volatility approach pose a series of difficult challenges that are still the subject of active research. First, market microstructure effects contaminate price observations, complicating the inference on the statistical properties of the true, efficient price process.

Second, the so-called non-synchronous trading effect (Lo and MacKinlay 1990) strongly affects the estimation of the realized covariance measures. Standard realized covariance measures constructed by imposing an artificially regularly spaced time series on high frequency data have an attenuation bias since the difference in the time stamps of the last ticks in each regularly spaced interval is not taken into account. This problem, which tends to increase with the sampling frequency, was first reported by Epps (1979) and is hence termed the Epps effect. Various approaches have been proposed in the literature to tackle this asynchronicity problem: incorporate lead and lag cross returns in the estimator (Scholes and Williams 1977; Cohen et al. 1983; Bollerslev and Zhang 2003; Bandi and Russell 2005), avoid any synchronization by directly using tick-by-tick data (De Jong and Nijman 1997; Hayashi and Yoshida 2005; Palandri 2006; Sheppard 2006; Voev and Lunde 2007; Corsi and Audrino 2012; Griffin and Oomen 2011), adopt the so-called refresh time scheme (Barndorff-Nielsen et al. 2011; Hautsch et al. 2012; Ait-Sahalia et al. 2010; Zhang 2011), and apply the multivariate Fourier method (Renò 2003; Mancino and Sanfelici 2011).

[1] See, e.g., Andersen et al. (2001, 2003); Barndorff-Nielsen and Shephard (2001, 2002a,b, 2005), and Comte and Renault (1998).

Third, the covariance matrix needs to be positive-semidefinite (psd). However, any kind of correction for these aforementioned microstructure effects will typically result in a covariance matrix which is not guaranteed to be psd. Notable exceptions are the multivariate realized kernel with refresh time of Barndorff-Nielsen et al. (2011) and the above mentioned multivariate Fourier method of Mancino and Sanfelici (2011). In both cases, however, the frequency at which all the realized covariance estimates are computed is dictated by the asset having the lowest liquidity, hence discarding in practice a considerable amount of information, especially for the most liquid assets. To alleviate this problem, Hautsch et al. (2012) proposed an extension of the multivariate realized kernel in which assets are first grouped according to liquidity and then the covariance matrix is estimated block-wise and regularized.

In this paper, we propose a new approach to the estimation of the covariance matrix based on the idea of viewing the asynchronicity problem as a missing values problem on a set of otherwise synchronous ultra-high-frequency series; i.e., in our view, data on a very high-frequency grid are synchronous, although many observations are missing. Then we consider the asynchronicity as arising from the fact that when some assets trade, the observations of some other assets may be missing. The advantage of this point of view is that powerful statistical methodology for dealing with missing observations can be employed to cope with the asynchronicity problem in asset prices. In fact, modeling the market microstructure noise as a measurement error on the true latent efficient price naturally leads to a state space model with the transition equation describing the dynamics of the latent efficient price and the observation equation modeling the contamination due to the market microstructure effects. This state space approach with missing values, in turn, naturally leads to an estimation methodology based on Kalman filter recursion within an Expectation Maximization (EM) algorithm.[2] EM effectively deals with the missing observation problem by iterating between two steps: the expectation step, which in our context reconstructs the latent and synchronized series of the efficient price processes by means of the Kalman recursion; and the maximization step, which searches for model parameter values (the entries of the covariance matrices in our case) that maximize the expected likelihood obtained with the reconstructed price series. In this way EM guarantees maximizing the likelihood of the observed data even in the presence of missing observations (see Dempster et al. 1977). Therefore, the proposed estimators can also be seen as an application of the QMLE approach to quadratic variation

[2]Alternatively, a Bayesian approach for sampling the missing observations can be employed using a Gibbs Sampler, as proposed in Peluso et al. (2013).

estimation recently proposed by Xiu (2010). We term this approach Kalman-EM or KEM for short.

The proposed KEM estimator has the important advantage of employing all the information available in all the price series, thereby making use of all the trades of any asset. Specifically, by reconstructing each latent price series using all the information contained in the other series, the KEM methodology pulls all the available multivariate information in computing each single pair of covariances while guaranteeing that the estimated covariance matrix is psd. Through an extensive simulation study, we show that the KEM approach is simultaneously able to effectively deal with both the asynchronicity (or missing value) problem and the market microstructure noise, providing realized covariance matrices which are both psd by construction and more accurate than any of the competing methodologies considered. Furthermore, as we show in an empirical application to US stocks, given its computational efficiency, KEM is also feasible in large dimensions and can be applied in practical situations involving several dozens (or even hundreds) of assets.

A similar approach has recently been investigated in two independent and concurrent papers by Liu and Tang (2012) and Shephard and Xiu (2012).[3] The focus of these studies, however, is more on the derivation of the theoretical asymptotic properties of the estimator, showing that it converges at the optimal rate. In our work we put greater emphasis on the methodological implementation of the KEM estimator and we show how it can be used in possibly large dimensional applications. Liu and Tang (2012) study a realized QML estimator of a multivariate exactly synchronized data set. They propose using refresh time type devices to achieve exact synchronicity. By contrast, and similarly to Shephard and Xiu (2012), in our approach we take into account asynchronous trading by considering all ticks observed in the market. To estimate the realized covariance matrix, Shephard and Xiu (2012) also introduce a missing value approach based on a Brownian motion observed with error, but their algorithm involves different computations.

The rest of the paper is organized as follows. Section 2 introduces our multivariate state space model. Section 3 describes the proposed KEM estimation approach. Section 4 is dedicated to the presentation of the results of an extensive simulation study which compares the KEM estimator with several competing approaches over a broad range of settings. Section 5 contains the empirical application of our proposed estimator to a portfolio of one hundred US stocks taken from the universe

[3] In the different context of now casting the macroeconomic activity, a similar approach has also been applied by Beber et al. (2013).

of the S&P 500 constituents, and Section 6 concludes.

2 Model

We start by specifying a general continuous-time process for the efficient log-price

$$d\mathbf{X}(t) \quad = \quad \mu_t dt + \Sigma d\mathbf{W}(t), \tag{1}$$

where $\mathbf{X}(t)$ is the d-dimensional latent log-price process free of noise, $\mathbf{W}(t)$ is the d-dimensional vector of independent Brownian motions, the drift μ_t is a vector of predictable locally bounded elements, and $\Sigma\Sigma' = Q$ is the constant diffusion coefficient matrix.

Model (1) can be regarded as too simplistic along two dimensions. First, we do not consider the presence of jumps. Nevertheless, the focus of our approach is on the estimation of the continuous part of the quadratic variation of the latent log-price process. In the presence of jumps, the problem of separating the continuous part from the jump part could be preemptively tackled by pre-testing the data for jumps using jump identification tests that are able to locate the position of jumps inside the day (such as Fan and Wang 2007; Lee and Mykland 2008). Our methodology can then be applied to the resulting jump-filtered series.

Second, we recognize that the assumption of a constant instantaneous volatility matrix is clearly restrictive. In general, $Q = Q_t$ is a multivariate time-varying process. However, our goal is the estimation of the daily integrated volatility $\int_0^\tau Q_t dt$. We test the robustness of our results in simulations allowing for time-varying stochastic volatility obtaining very accurate estimates of the integrated volatility. This result is not surprising given the recent studies of Xiu (2010) for the univariate case and Liu and Tang (2012) and Shephard and Xiu (2012) for the multivariate case, which showed that the QMLE of the integrated volatility on a misspecified model with constant volatility and zero drift remains consistent and optimal in terms of its rate of convergence for $\int_0^\tau Q_t dt$ under fairly general assumptions. In fact, the methodology we introduce is based on the Expectation Maximization (EM) algorithm, known to converge, under certain technical conditions, to the MLE of the incomplete-data likelihood. Thus, we reasonably expect our results to be valid in settings more general than (1), in particular when the volatility matrix is time-varying.

In the empirical data, observed log transaction prices are, however, contaminated by microstructure

noise. We propose to model the system composed of the latent and observed log-price as a state-space model discretized on an ultra-high-frequency grid (of one second in the simulation and empirical analysis). Our ultra-high-frequency discrete time system then reads:

$$\mathbf{Y}_t = \mathbf{X}_t + \eta_t \quad \eta_t \sim N(\mathbf{0}, R), \tag{2}$$

$$\mathbf{X}_t = \mathbf{X}_{t-1} + \epsilon_t \quad \epsilon_t \sim N(\mathbf{0}, Q), \tag{3}$$

where $\mathbf{X}_t$ is the vector of true-latent-efficient prices, $\mathbf{Y}_t$ is the vector of observed prices, partitioned in its observed and missing components $[\mathbf{Y}_t^o, \mathbf{Y}_t^m]$ and η_t and ϵ_t are assumed to be uncorrelated and mutually independent errors. Although the normality assumption we are imposing for the error distributions might be empirically questionable, it greatly simplifies the methodological implementation of the KEM estimator by allowing the application of the standard Kalman filter recursion in the Expectation step of the EM estimation procedure (as detailed in the next section). We assume the covariance matrix R to be diagonal, meaning that the microstructure noises are uncorrelated across assets.[4]

The discretized model is a simple linear state space model, known as a local level model, consisting of the observation equation (2) and the state equation (3). The model can also be viewed as a particular case of a dynamic linear model (Elliott et al. 1995; Roweis and Ghahramani 1999; West and Harrison 1997), characterized in its general form by $\{A, C, R, Q\}$, respectively, the observation matrix, transition matrix, observation error variance matrix and transition error variance matrix, possibly time-varying. Then, our model is a time-invariant dynamic linear model with matrices $\{I_d, I_d, R, Q\}$.

3 Estimation methodology

Following Shumway and Stoffer (1982), the estimation of the linear Gaussian dynamic system in (2)-(3) is performed using the EM algorithm. Here we briefly review the idea of the powerful EM algorithm.

[4]The generalization to non-diagonal microstructure noise variance is feasible but with an additional computational effort: The sufficient statistics reported in Appendix A.3 and the iterative formula for the estimated microstructure noise variance become more involved, while the iterative formula for the estimated latent log price covariance remains the same. However, due to the intrinsic asynchronicity of market microstructure noise, the empirical matrix R tends to be approximately diagonal. Moreover, simulation results with non-diagonal R (reported in the web-based appendix) show that the deterioration of the precision of the estimator is limited and does not affect its ranking with respect to the competing estimators (i.e., KEM remains the best performing one).

3.1 The Kalman-EM algorithm

The objective of the EM algorithm (Dempster et al. 1977) is to maximize the likelihood of the observed data in the presence of hidden variables. In the problem under consideration, the hidden variables are $\tilde{\mathbf{X}} = [\mathbf{X}, \mathbf{Y}^m]$, that is, the true latent process $\mathbf{X}$ and the missing observations of the process $\mathbf{Y}^m$. Maximizing the likelihood as a function of Q and R is equivalent to maximizing the log-likelihood:

$$L(Q, R) = \ln P(\mathbf{Y}^o | Q, R) = \ln \int P(\tilde{\mathbf{X}}, \mathbf{Y}^o | Q, R) d\tilde{\mathbf{X}}.$$

Using any distribution π over the hidden variables, we can obtain a lower bound on L. In fact:

$$\begin{aligned}
L(Q, R) = \ln P(\mathbf{Y^o} | Q, R) &= \ln \int P(\tilde{\mathbf{X}}, \mathbf{Y^o} | Q, R) d\tilde{\mathbf{X}} \\
&= \ln \int \frac{P(\tilde{\mathbf{X}}, \mathbf{Y^o} | Q, R)}{\pi(\tilde{\mathbf{X}})} \pi(\tilde{\mathbf{X}}) d\tilde{\mathbf{X}} \\
&\geq \int \ln \left(\frac{P(\tilde{\mathbf{X}}, \mathbf{Y^o} | Q, R)}{\pi(\tilde{\mathbf{X}})} \right) \pi(\tilde{\mathbf{X}}) d\tilde{\mathbf{X}} \\
&= \int \ln P(\tilde{\mathbf{X}}, \mathbf{Y^o} | Q, R) \pi(\tilde{\mathbf{X}}) d\tilde{\mathbf{X}} - \int \ln \pi(\tilde{\mathbf{X}}) \pi(\tilde{\mathbf{X}}) d\tilde{\mathbf{X}} \\
&\equiv F(\pi, Q, R)
\end{aligned}$$

where the middle inequality is due to the Jensen inequality induced by the concavity of the log function. The EM algorithm alternates between maximizing F with respect to the distribution π and the parameters Q and R, respectively, holding the other fixed. Specifically, starting from some initial parameters Q_0 and R_0, at iteration $k+1$ we have:

$$\begin{aligned}
&\textbf{E step:} \quad \pi_{k+1} = \arg\max_{\pi} F(\pi_k, Q_k, R_k) \\
&\textbf{M step:} \quad Q_{k+1}, R_{k+1} = \arg\max_{Q_k, R_k} F(\pi_{k+1}, Q_k, R_k);
\end{aligned}$$

Thus, EM can be interpreted as coordinate ascent algorithm in F.

The maximum in the E-step results when π is exactly the conditional distribution of $\tilde{\mathbf{X}}$, i.e. $\pi_{k+1} = P(\tilde{\mathbf{X}} | \mathbf{Y}^o, Q_k, R_k)$ at which point the bound becomes an equality: $F(\pi, Q, R) = L(Q, R)$ (see Appendix A.1). Since $P(\tilde{\mathbf{X}} | \mathbf{Y}^o, Q_k, R_k)$ is the standard output of the Kalman filter, in our model the E-step is simply performed by running the Kalman filtering and smoothing recursion. The forward (filtering) and backward (smoothing) Kalman recursions for our model are reported in Appendix A.2.

The filtering recursion formulas are only slightly modified to take into account the missing data by entering zeros in the observation vector $\mathbf{Y}_t$ where data is missing and by zeroing out the corresponding row of the observation matrix (Shumway and Stoffer 1982), which, in our model, is simply the identity matrix (see Section 2).

The maximum in the M-step is obtained by maximizing the first term of $F(\pi, Q, R)$, since the second term (the entropy of π) does not depend on Q and R, i.e.,

$$\begin{aligned} Q_{k+1}, R_{k+1} &= \arg\max_{Q,R} \int_{\tilde{\mathbf{X}}} \log P(\mathbf{X}, \mathbf{Y}|Q_k, R_k) P(\tilde{\mathbf{X}}|\mathbf{Y^o}, Q_k, R_k) d\tilde{\mathbf{X}} \quad (4) \\ &= \arg\max_{Q,R} \mathbb{E}[\log P(\mathbf{X}, \mathbf{Y}|Q_k, R_k)] \quad (5) \end{aligned}$$

where the expectation is taken with respect to the distribution of $\tilde{\mathbf{X}}|\mathbf{Y^o}, Q_k, R_k$. Therefore, we maximize the expected log likelihood of the joint data (observed and hidden) under $\pi_{k+1} = P(\tilde{\mathbf{X}}|\mathbf{Y^o}, Q_k, R_k)$, that is, the distribution of the hidden variables conditional on the observations and parameters of the model. Given the estimated hidden variables obtained from the E-step, it becomes easy to solve for a new set of parameters. In fact, this reduces to the minimization of quadratic forms, given that the likelihood of the joint data is:

$$\begin{aligned} \log P(\mathbf{X}, \mathbf{Y}|Q, R) \quad \propto \quad & -\frac{T}{2} \ln |Q| - \frac{1}{2} \sum_{t=1}^{T} (\mathbf{X}_t - \mathbf{X}_{t-1})' Q^{-1} (\mathbf{X}_t - \mathbf{X}_{t-1}) \\ & -\frac{T}{2} \ln |R| - \frac{1}{2} \sum_{t=1}^{T} (\mathbf{Y}_t - \mathbf{X}_t)' R^{-1} (\mathbf{Y}_t - \mathbf{X}_t). \end{aligned}$$

Then, the maximization is analogous to the usual multivariate regression approach, which yields the following estimated values of Q and R at iteration $(k+1)$:[5]

$$\begin{aligned} Q_{k+1} &= (T-1)^{-1} \left[\sum_{t=2}^{T} (V_t^s + \mathbf{X}_t^s \mathbf{X}_t^{s'}) - \left(\sum_{t=2}^{T} (VL_t^s + \mathbf{X}_t^s \mathbf{X}_{t-1}^{s'}) \right) \right. \\ &\quad \left. \left(\sum_{t=2}^{T} (V_{t-1}^s + \mathbf{X}_{t-1}^s \mathbf{X}_{t-1}^{s'}) \right)^{-1} \left(\sum_{t=2}^{T} (VL_t^s + \mathbf{X}_t^s \mathbf{X}_{t-1}^{s'}) \right)' \right] \quad (6) \\ R_{k+1} &= T^{-1} \sum_{t=1}^{T} D_t \begin{pmatrix} (\mathbf{Y}_t^o - \mathbf{X}_t^{o,s})(\mathbf{Y}_t^o - \mathbf{X}_t^{o,s})' + V_t^{o,s} & \mathbf{0} \\ \mathbf{0} & R_t^m(k) \end{pmatrix} D_t', \quad (7) \end{aligned}$$

[5]The sufficient statistics suggested by Digalakis et al. 1993 for this maximization are obtained from the smoothing recursions and are reported in Appendix A.3.

where $\mathbf{X}_t^s$ is the smoothed log-price latent process and $\mathbf{X}_t^{o,s}$ are the components of $\mathbf{X}_t^s$ corresponding to $\mathbf{Y}_t^o$; V_t^s is the variance of the smoothing error, $V_t^{o,s}$ is the submatrix of V_t^s corresponding to $\mathbf{Y}_t^o$, and VL_t^s is the one-lag autocovariance of the smoothing error (for the expression of the smoothed quantities see Appendix A.2); R_t^m is the submatrix of R corresponding to $\mathbf{Y}_t^m$, and D_t is a permutation matrix that at each instant t orders the observed and then the missing components of $\mathbf{Y}_t$ (the original order is then re-established with D_t').

Summarizing, we iterate between

- *Kalman filter/smoother to estimate the unknown hidden variables given the observations and current parameter values*
- *using this fictitious complete data to solve for new parameters in the expected log likelihood of the joint data.*

We monitor the convergence of KEM by computing recursively from the filtering iterations the prediction error decomposition form of the incomplete data normal log-likelihood $L(Q, R)$ (Gupta and Mehra (1974)) since

$$L(Q,R) \propto \sum_t \left[-\frac{1}{2}\ln(|V_t^{o,p} + R_t^o|) - \frac{1}{2}(\mathbf{Y}_t^o - \mathbf{X}_{t-1}^{o,f})'(V_t^{o,p} + R_t^o)^{-1}(\mathbf{Y}_t^o - \mathbf{X}_{t-1}^{o,f}) \right]. \tag{8}$$

where $\mathbf{X}_t^f$ is the filtered value of the log-price latent process, $\mathbf{X}_t^{o,f}$ is the component of $\mathbf{X}_t^f$ corresponding to $\mathbf{Y}_t^o$, V_t^p is the variance of the prediction error, and R_t^o and $V_t^{o,p}$ are the submatrices of, respectively, R and V_t^p, corresponding to $\mathbf{Y}_t^o$ (for the expression of these quantities see Appendix A.2).

With the choice $\pi(\tilde{\mathbf{X}}) = P(\tilde{\mathbf{X}}|\mathbf{Y^o}, Q_k, R_k)$, $F(\pi, Q, R) = L(Q, R)$ at the beginning of each M-step and since the E-step does not change Q and R, we are guaranteed not to decrease the likelihood after each combined EM-step. Thus, although it seems that we are maximizing the wrong likelihood (the expected likelihood of the joint data instead of that of the observed data), EM indeed guarantees to increase the correct likelihood of interest, with the advantage of important computational benefits compared to the standard ML approach. In fact, the application of nonlinear minimization methods to this problem is computationally demanding and extension to the case of missing data is very complicated. However, the EM approach only involves simple matrix multiplications at each step, and it can be easily extended to the case of missing data since the missing observations and the state

variables can be jointly treated as hidden variables (Digalakis et al. 1993).

3.2 Signal extraction of the latent price

An important byproduct of our approach is the signal extraction of the latent efficient price process $\mathbf{X}$ for each asset. By the KEM estimation procedure we can filter out the microstructure noise and reconstruct the latent dynamics of the efficient price by exploiting the correlations of one asset with the dynamics of all the other assets. This signal extraction is particularly useful for the less liquid series that have fewer observations because they can benefit more from the information contained in the dynamics of the more liquid assets. The multivariate nature of the proposed latent price reconstruction represents the main difference and advantage respect to the univariate filtering obtained with the so-called pre-averaging method introduced by Jacod et al. (2009), which performs a local average of the univariate time series of observed prices. Moreover, our approach automatically guarantees that the multivariate and two-sided local weighting scheme is optimally chosen by the Kalman filter-smoother algorithm. Therefore, our signal extraction procedure can be seen as a natural and computationally convenient multivariate extension of the Jacod et al. (2009) pre-averaging approach.

As an illustration, in Figure 1 we plot the reconstructed latent price process together with the observed tick prices, for one of the assets in our empirical application: We show the first 500 seconds of the trading day January 3, 2007 for Nike Inc.; see Section 5 for more details.

3.3 Asymptotic properties

As simply a more convenient and computationally efficient way of maximizing the likelihood of the observed data, the KEM estimator of Q can be interpreted as a QMLE for the quadratic variation of the efficient price process. Hence, under standard regularity conditions on the likelihood function (Wu, 1983), the estimator possesses all classical asymptotic properties. Although a rigorous proof of the asymptotic theory of the Gaussian state space model similar to ours can be found in Shephard and Xiu (2012), the basic intuition for the consistency of the estimator is quite simple: First, for the number of EM iterations tending to infinity, the estimator $\hat{Q}_T$ converges to the realized volatility of the latent process $\{\mathbf{X}\}$, where T denotes the total number of points reconstructed for the latent price process. Second for $T \to \infty$, the realized volatility of the latent process converges (as is well known

from the standard theory of realized volatility in absence of noise and asynchronicity) to the integrated covariance process $\int_0^\tau Q_t dt$, which is our quantity of interest.

4 Simulation Study

In this section we compare the performance of our KEM estimator with other estimators proposed in the literature by means of an extensive simulation analysis. Specifically, the competing estimators we consider are: (i) the pairwise estimator proposed by Ait-Sahalia et al. (2010) (AFX), (ii) the Multivariate Realized Kernel of Barndorff-Nielsen et al. (2011) (MRK), and (iii) a covariance matrix obtained with the Zhang et al. (2005) Two Scale estimators for the variance elements and the pairwise covariance terms estimated by Hayashi and Yoshida (2005) (HY).

In order to test the robustness of the estimator to the misspecification induced by intraday volatility dynamics, we generate the data using the stochastic volatility model of Heston (1993). The data generating process (DGP) is, for $i = 1, \cdots, d$ and $t = 1, \cdots, T$ (the total number of points reconstructed for the latent price process)

$$dX_{i,t} = \sigma_{i,t} dW_{i,t} \tag{9}$$

$$d\sigma^2_{i,t} = k_i(\bar{\sigma}^2_i - \sigma^2_{i,t}) + s_i \sigma_{i,t} dB_{i,t} \tag{10}$$

where $E(dW_{i,t} dW_{j,t}) = \rho_{ij} dt$, $E(dW_{i,t} dB_{j,t}) = \delta_{ij} \pi_i dt$. We generate the data through an Euler-Maruyama discretization scheme, where the first observation for the variance process is drawn from a Gamma distribution $\Gamma(2k_i\bar{\sigma}^2_i/s^2_i, s^2_i/2k_i)$ centered in the mean variance. We simulate the covariance matrix of 10 assets for $M = 500$ simulated sample paths. The KEM running times per path are a few seconds for portfolios of 2-3 assets, a few minutes for tens of assets, and up to 2.5 hours for the hundred assets we consider in our application, using all the data in one trading day.[6] All simulations are initialized from $P_0 = log([100, 40, 60, 80, 40, 20, 90, 30, 50, 60])$. We stress that the DGP (9)-(10) is misspecified, since it assumes $\sigma_{i,t}$ to be a stochastic process, while the methodologies under comparison assume $\sigma_{i,t} = \sigma_i$, constant in time.

Our KEM estimator, based on the synchronized efficient price process reconstructed at the one

[6] All codes were written in Matlab 7.11.0 (R2010b) and run (possibly in parallel) with Intel(R) Xeon(R) CPU X7460 @ 2.66 GHz.

second frequency, is psd by construction. If the covariance matrix obtained with the pairwise AFX and HY estimation procedures are not psd, we project them onto the space of psd matrices using the methodology proposed in Higham (2002).[7] Convergence of the EM algorithm is assumed when the relative percentage increase of the log-likelihood (8) is below 0.00001. The true DGP variance matrices Q and R used in the simulation frameworks are reported in Table 1 and closely mimic the ones obtained when applying the KEM method on real data.

We perform the study in six simulation settings, reported in Table 2, which differ in terms of noise-to-signal ratio and mean and standard deviation of the percentage of missing observations. We are then able to reproduce a broad range of empirically realistic cases which combine moderate and severe market microstructure noise contamination with a wide spectrum of missing probability distributions across the 10 assets. We characterize the scenarios according to three summary indicators: the average noise-to-signal and the average probability and the standard deviation of the missing observations. The average noise-to-signal is simply the ratio between the diagonal values of R, the covariance matrix of the microstructure error, and the diagonal values of Q, the covariance matrix of the latent log-price process. It indicates how much noise contaminates the observed asset prices on the market, relative to the true signal that, on the other hand, drives the dynamics of the underlying true value of the assets. It goes from 0.78 in the *Standard* scenario, and it is increased to 2.58 in the *High noise* scenarios (rows 2, 4 and 6 in Table 2). The average missing probability is the number of seconds without any observed price on the market, over the total number of seconds observed in a trading day, averaged over all assets and all the trading days of the simulation study. It is an indicator of liquidity: it starts from 0.32 in the *Standard* scenario and arrives at 0.67 in *High missings* scenarios (rows 3-5), with peaks of assets simulated with up to 85% of missing values. Finally, we distinguish our simulation settings according to the standard deviation of missing observations. This is the rooted sum of the deviations of the missing probabilities of each stock from the average missing probability. It controls for liquidity heterogeneity: a large value indicates the presence in the portfolio of assets with very different liquidity profiles, that is very liquid and very illiquid stocks. It is fixed at 0.11 in the *Standard* scenario, and moved to 0.35 in *Dispersed missings* scenarios in rows 5 and 6.

As for the choice of a synthesis measure of performance, Laurent et al. (2013) show that the only two distances that guarantee consistency between orderings based on estimated and true covariance matrices in a multivariate setting are the Frobenius norm and the Stein distance. The Frobenius norm

[7] We notice that in our simulation settings, the results of the projected matrices are almost indistinguishable from those of the unprojected ones.

is the matrix difference between the estimated and the true one used to generate the data Q

$$Frob = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d}(\hat{\sigma}_{ij} - \sigma_{ij})^2}$$

where $\hat{\sigma}_{ij}$ is the ij-th element of the estimated $\hat{Q}$ and σ_{ij} is the ij-th element of the true Q used in the simulation, for the various methodologies in the different settings. [8]

The sample means and standard deviations of the Frobenius distances are summarized in Table 3. As illustrative examples, in Figures 2 and 3 we graphically compare the kernel density of the $M = 500$ simulated Frobenius distances obtained from each competing estimator for two of the more realistic scenarios: high missings and dispersed missings with high noise. Similar plots for the other scenarios can be found in a web-based appendix.

Starting from the standard setting featuring a moderate level of microstructure noise and a high and homogeneous level of liquidity across the assets, Table 3 shows a clear ranking among the considered estimators, with KEM being distinctly the most accurate followed by AFX, HY, and MRK, respectively. These results confirm that the KEM ability to pull information from the most liquid assets to improve covariance estimations leads to more accurate realized covariance estimates. They also confirm that MRK guarantees the positive-semidefiniteness of the estimated covariance matrix at the price of discarding a considerable amount of information.

In the high noise setting, the noise-to-signal ratio is raised from the value 0.78 of the standard setting to 2.58. This level of noise contamination is quite extreme for standard empirical data, providing a useful stress test for the realized covariance estimators. In this setting HY is the worst performing estimator since it is not robust to microstructure noise. Although now closely followed by AFX, KEM is again the best estimator.

In the high missing setting we instead raise the number of missing observations compared to the standard setting by increasing the average probability of missing from 0.32 to 0.67, thus effectively more than doubling it. Within this setting we can stress test the ability of our KEM approach to

[8] We also considered the Stein difference introduced by James and Stein (1961) and defined as

$$Stein = tr\left(\hat{Q}^{-1}Q\right) - \ln\left|\hat{Q}^{-1}Q\right| - d$$

to assess differences across the methods obtaining very similar results. For the sake of brevity, we do not report the results; these are available from the authors upon request.

reconstruct the latent efficient price process of the 10 stocks from a much smaller number of observed prices. Interestingly, the KEM estimator is still able to outperform the HY one which, in this setting, becomes the best one among the remaining estimators as it is especially suited to deal with sparse data and asynchronicity (Figure 2).

By combining both a high level of noise and missings, we construct a very challenging setting which underlines the ability of the estimators to be contemporaneously robust to strong market microstructure noise and low asset liquidity. Now AFX performs better than HY and MRK, as it is better equipped to deal with noise. However, the KEM estimator still remains the preferred one, confirming its robustness to both sources of errors.

Another direction in which we stress test our estimators is by intensifying the problem of asynchronicity among the assets. For this purpose we construct a simulation setting in which the liquidity of the different assets is more dispersed, i.e., we increase the standard deviation of the distribution of the missing probabilities across the 10 assets. Within this setting we consider two levels of noise-to-signal ratios, moderate and high. We term these two settings dispersed missing and dispersed missing high noise, respectively (see Figure 3 for the latter). Even under these settings the KEM estimator is able to pull information from the most liquid assets to obtain better covariance estimates, again proving to be, by far, the best-performing estimator. In contrast, the approaches based on the refresh time (AFX and MRK) suffer from the limitations of this scheme which, dictated by the less liquid assets, discards a large amount of information. In fact, contrary to the standard case, in the dispersed missing setting HY outperforms AFX as it is better able to deal with a high level of asynchronicity. In the dispersed missing high noise case, however, the lack of robustness of HY to microstructure noise heavily degrades its performance.

To assess statistically whether the differences in the performances are relevant, we run a series of formal tests. We start by computing Model Confidence Sets (MCS) following the procedure introduced by Hansen et al. (2003) and (2011). The MCS approach was introduced with the goal of characterizing the best subset of models with respect to a pre-specified performance measure out of a set of competing ones. In recent years it has become a standard for this kind of task. Not surprisingly, we find that in all simulation settings, the MCS at all relevant confidence levels exclusively consists of the KEM approach. What changes from setting to setting is only the sequence in which the other models are

deleted from the confidence set (as reported in Table 3). Full details on the MCS results can be found in the web-based appendix.

Summarizing the results of the simulation study across the different settings, the KEM estimator consistently emerges as the most accurate measure of realized covariance among the considered estimators. This is due to its high level of robustness to both microstructure noise and asynchronicity and to its capacity of effectively exploiting all the multivariate information available: none of the observations is discarded and all are used in each single covariance estimate. Therefore, contrary to what happens in the MRK approach, the positive-semidefiniteness of the KEM estimated covariance matrix comes at no cost in terms of the precision of the estimates.

Finally, to test for the robustness of the estimators to other disturbing factors that may arise in empirical applications, we replicate the simulation study with a more realistic data-generating process. In detail, we consider a Heston-type stochastic volatility model as before that also features intraday volatility patterns and jumps in the price process. As the simulation results reported in the web-based appendix show, including these factors in the data-generating process does not alter the estimation outcomes and, importantly, the KEM method continues to significantly outperform all its competitors.

5 Application

In our empirical analysis we use a data set from the TAQ database, consisting of tick-by-tick data for 100 US stocks, over the period Jan 1, 2006 - Dec 31, 2008. We first filter the tick-by-tick data to remove wrongly reported observations that may distort our analysis. In particular, we use as a simple filter the following rule: We first compute a robust indicator of the daily price variability as the sample standard deviation of the observed prices falling between the 25^{th} and 75^{th} percentile of the empirical distribution. We then remove all prices that are outside a confidence interval constructed by two times this robust measure of variability and centered around both the previous and next price tick.[9]

[9] More formally, take $(Y_{i,1}, \ldots, Y_{i,T})$ to be the T observed log prices of asset i in a generic day. Sort the prices to get the ordered sample $(Y_{i,(1)}, \ldots, Y_{i,(T)})$ and compute the threshold d as the sample standard deviation of $(Y_{i,(\lfloor 0.25T \rfloor)}, \ldots, Y_{i,(\lfloor 0.75T \rfloor)})$, where $\lfloor \cdot \rfloor$ is the floor function. If $|Y_{i,t} - Y_{i,t-1}| > 2d$ and $|Y_{i,t} - Y_{i,t+1}| > 2d$, then $Y_{i,t}$ is removed from the series. This filtering strategy can be improved along several dimensions: see, for example, Brownlees and Gallo (2006). Nevertheless, by visual inspection of the resulting filtered series, we see that the simple filter we use is effective in removing wrong observations.

In Table 4 we report some descriptive statistics for a sub-sample of companies. Our data set contains both liquid stocks such as Apple, Exxon, and General Electric with an average probability of missing observations at the one second frequency around 0.4, and much less liquid stocks such as Biogen or Prudential Financial with an average probability of missing observations larger than 0.75.

The KEM estimated average noise-to-signal ratio for the empirical data is 1.15, with mean and standard deviation of missing probabilities equal to 0.67 and 0.14, respectively. Thus, High missings is the simulation scenario closest to our empirical data.

5.1 Estimated variances and correlations

As illustrative pictures, in Figure 4 we plot the time series of the the annualized correlation between Exxon Mobile Corporation (XOM) and Chevron Corporation (CVX), estimated with KEM, MRK, AFX, and HY. The last two methodologies are projected to impose the positiveness of the resulting matrix. XOM and CVX are two of the most liquid assets in our sample, both operating in the U.S. energy sector. In Figure 5 we plot the same graph for two assets with less liquidity: the estimated correlation between Lowe's Companies, Inc. (LOW) and Apache Corp. (APA). Similar plots can be obtained for all the other stocks.

The four methods return realized variance estimates that lie close to one another: two illustrative examples are shown in the web-based appendix. Some differences among the estimators are worth highlighting when considering realized correlation estimates: First, with the only exception of MRK for XOM/CVX correlations, all considered estimators identify a surge in the value of correlations during the financial crisis. In fact, the MRK XOM/CVX correlation estimates before the financial crisis seem to be somehow too large, causing the MRK measure to miss the steady increase of correlations identified by the other estimators. As can be seen in Figure 5, this increase is not always smooth and/or monotonic. Second, HY correlation estimates between highly liquid assets tend to be smaller in absolute value than the alternative approaches. This could be a consequence of the HY downward bias on ultra-high-frequency data that has been recently documented by Griffin and Oomen (2011). Third, the AFX and MRK estimates appear to be somehow less smooth than those obtained using KEM, generating large and extreme values and a much more erratic correlation dynamics.

5.2 Portfolio selection

Following the strategy introduced by Engle and Colacito (2006) and extensively used thereafter, we perform a comparison of the accuracy of the different realized covariance estimators for portfolio selection.[10]

The idea is that an investor solves the classical mean-variance portfolio allocation problem to get the optimal weights of the different constituents of the portfolio. That is, the investor minimizes the expected variance of the portfolio $\sigma_t^2 = \omega_t' Q_t \omega_t$ under the constraint that the weights ω_t add to some scalar value $\overline{\omega}$:

$$\min_{\omega_t} \sigma_t^2 = \min_{\omega_t} \omega_t' Q_t \omega_t, \quad \text{s.t. } \omega_t^{'} i = \overline{\omega},$$

where i denotes a vector of ones. Clearly, to solve such an optimization problem, the investor should know the conditional covariance matrix Q_t of the assets composing the portfolio under investigation. Given that this matrix is unknown, conditional covariances are predicted and denoted by $\hat{Q}_{t,j}$, where j identifies the method used for prediction. In this study we adopt the simple strategy that the next day covariance predictions are given by today's covariance estimates obtained from the different methods considered above.[11] Portfolio weights are then constructed minimizing $\sigma_{t,j}^2 = \omega_{t,j}' \hat{Q}_{t,j} \omega_{t,j}$. To ensure the consistency of the ranking of the models and to avoid misleading results, we set $\overline{\omega} = 1$ and comparisons are based on the global minimum variance portfolio (see, for example, the discussion in Hansen and Lunde 2006 or Patton and Sheppard 2009).

Engle and Colacito (2006) show that in order to quantify the gains from superior covariance prediction, one can look at the ratio

$$\frac{\sigma_{t,j} - \sigma_t}{\sigma_t} \geq 0.$$

This ratio can be interpreted as the percentage reduction in portfolio investment that could have been achieved by knowing the true covariance matrix.

Moreover, when two alternative methods are at one's disposal, one can test differences in the following way: For each period, the mean-variance problem is solved and portfolio weights are con-

[10] As customary in the realized volatility literature, all the daily measures refer to the open-to-close period of the daily market activity.

[11] This is a simple strategy that can be improved by also taking into account the time-varying dynamics of the covariances. However, this would need to specify a time series model for the dynamics of the high-dimensional conditional covariance matrices and is far beyond the scope of this study.

structed based on each covariance matrix prediction. One can perform a test on the series of the differences of the variances on the two portfolios $u_t = \sigma^2_{t,j_1} - \sigma^2_{t,j_2}$ using generalizations of the standard Diebold and Mariano (1995) test. The null hypothesis of equal portfolio variance is simply a test that the mean of the differences u_t is zero. Significantly positive (negative) means are in favor of method j_2 (j_1). In case of multiple comparisons, the model confidence set approach introduced in Section 4 can be performed.

Results of pairwise tests on the portfolio variances constructed from the hundred US stocks in the S&P500 universe for each competing approach against the KEM method are summarized in Table 5, Panel A. The out-of-sample time period goes from January 1, 2006 to December 31, 2008. Results clearly show the superiority of the KEM approach: all alternative methods are significantly outperformed at the 99% confidence level. The efficiency gains, going from 50 to 310 basis points depending on the alternative method, are economically significant and slightly larger than the usual range found in the literature.

In Panel B of Table 5 we summarize the results of the MCS constructions. Once again, results are clearly in favor of the KEM method. In particular, the KEM approach is the only one belonging to both 90% and 95% MCS. At the 99% confidence level, the second-best performing MRK method also belongs to the confidence set.

Given that the last period in our sample corresponding to the last months of 2008 is a very noisy period, we finally redo the whole portfolio choice application excluding the last 100 observations. Results of the different tests do not change qualitatively but support the superiority of the KEM approach even more clearly.

6 Conclusions

In this paper we proposed a new view on the problem of asynchronicity in the realized covariance estimation by considering it as a missing data problem. Together with the treatment of the market microstructure noise as a measurement error problem, this naturally leads to a state-space model with missing observations for which an EM-type of approach is particularly suited. We then estimate the covariance matrix of the latent price process with a Kalman-EM (KEM) algorithm, which iterates between the following two steps: (i) reconstruct the smoothed and synchronized series of the latent

price processes (E-step) and (ii) use this fictitious complete data set to easily maximize the complete data likelihood, obtaining new estimates for the parameters of interest, i.e., the covariance matrices of the latent price process and of the microstructure noise (M-step). The proposed KEM estimator is then robust to both asynchronicity and microstructure noise, and psd by construction.

We perform an extensive Monte Carlo simulation analysis reproducing a broad spectrum of empirically realistic cases in terms of both level of market microstructure noise and distribution of missing probabilities. Across all different settings, the KEM estimator consistently outperforms several competing estimators introduced in the previous literature. The reason for this superior performance is the ability of KEM to exploit all the information contained in the tick-by-tick multivariate series to reconstruct the latent efficient price process of each series. In this way the KEM estimator effectively pulls information from all the available series in computing each single pair of covariances and remains highly robust to both microstructure noise and asynchronicity (i.e., missing data). As it is also psd by construction and computationally efficient, the KEM estimator is highly suited for possibly high-dimensional portfolio applications such as portfolio selection and risk management, as we have shown in an application to one hundred stocks belonging to the S&P500 universe.

Acknowledgements

The authors would like to thank Prof. Antonietta Mira for fruitful discussions and the seminar participants at the 5th CSDA International Conference on Computational and Financial Econometrics (CFE 2011) for insightful comments. We also thank the editor, Tim Bollerslev, and two anonymous referees who helped us improve the paper.

References

Ait-Sahalia, Y., Fan, J., and Xiu, D. (2010). High-Frequency Covariance Estimates with Noisy and Asynchronous Financial Data. *Journal of the American Statistical Association*, 105:1504–1517.

Andersen, T. G., Bollerslev, T., Diebold, F., and Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, 96:42–55.

Andersen, T. G., Bollerslev, T., Diebold, F., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71:579–625.

Bandi, F. and Russell, J. (2005). Realized covariation, realized beta and microstructure noise. Unpublished paper, Graduate School of Business, University of Chicago.

Barndorff-Nielsen, O. and Shephard, N. (2002a). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:253–280.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and 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.

Barndorff-Nielsen, O. E. and Shephard, N. (2001). Non-gaussian ornstein-uhlembech-based models and some of their rses in financial economics. *Journal of the Royal Statistical Society*, Series B(63):167–241.

Barndorff-Nielsen, O. E. and Shephard, N. (2002b). Estimating quadratic variation using realized variance. *Journal of Applied Econometrics*, 17:457–477.

Barndorff-Nielsen, O. E. and Shephard, N. (2005). How accurate is the asymptotic approximation to the distribution of realized volatility? In Andrews, D. W. F. and Stock, J. H., editors, *Identification and Inference for Econometric Models. A Festschrift in Honour of T.J. Rothenberg*, pages 306–331. Cambridge University Press.

Beber, A., Brandt, M. W., and Luisi, M. (2013). Distilling the macroeconomic news flow.

Bollerslev, T. and Zhang, B. Y. B. (2003). Measuring and modeling systematic risk in factor pricing models using high-frequency data. *Journal of Empirical Finance*, 10(5):533–558.

Brownlees, C. T. and Gallo, G. M. (2006). Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics & Data Analysis*, 51(4):2232–2245.

Cohen, K., Hawawini, G. A., Maier, S. F., R., R. S., and D., D. W. (1983). Friction in the trading process and the estimation of systematic risk. *Journal of Financial Economics*, 12:263–278.

Comte, F. and Renault, E. (1998). Long memory in continuous time stochastic volatility models. *Mathematical Finance*, 8:291–323.

Corsi, F. and Audrino, F. (2012). Realized covariance tick-by-tick in presence of rounded time stamps and general microstructure effects. *Journal of Financial Econometrics*, 10:591–616.

De Jong, F. and Nijman, T. (1997). High frequency analysis of lead-lag relationships between financial markets. *Journal of Empirical Finance*, 4(2-3):259–277.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood Estimation from Incomplete Data. *Journal of Royal Statistical Society (B)*, 39:1–38.

Diebold, F. and Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13:253–263.

Digalakis, V., Rohlicek, J., and Ostendorf, M. (1993). ML Estimation of a Stochastic Linear System with the EM Algorithm and its Application to Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 1:431–442.

Elliott, R., Aggoun, L., and Moore, J. (1995). *Hidden Markov models: estimation and control*, volume 29. Springer.

Engle, R. F. and Colacito, R. (2006). Testing and valuing dynamic correlations for asset allocation. *Journal of Business & Economic Statistics*, 24(2):238–253.

Epps, T. (1979). Comovements in Stock Prices in the Very Short Run. *Journal of the American Statistical Association*, 74:291–296.

Fan, J. and Wang, Y. (2007). Multi-scale jump and volatility analysis for high-frequency financial data. *Journal of the American Statistical Association*, 102:1349–1362.

Griffin, J. and Oomen, R. (2011). Covariance measurement in the presence of non-synchronous trading and market microstructure noise. *Journal of Econometrics*, 160(1):58–68.

Gupta, N. and Mehra, R. (1974). Computational Aspects of Maximum Likelihood Estimation and Reduction in Sensitivity Function Calculations. *IEEE Trans. Automatic Control*, AC-19:774–783.

Hansen, P. and Lunde, A. (2006). Consistent ranking of volatility models. *Journal of Econometrics*, 131:97–121.

Hansen, P., Lunde, A., and Nason, J. (2003). Choosing the best volatility models: The model confidence set approach. *Oxford Bulletin of Economics and Statistics*, 65:839–861.

Hansen, P., Lunde, A., and Nason, J. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Hautsch, N., Kyj, L., and Oomen, R. (2012). A blocking and regularization approach to high-dimensional realized covariance estimation. *Journal of Applied Econometrics*, 27:625–645.

Hayashi, T. and Yoshida, N. (2005). On Covariance Estimation of Non-Synchronously Observed Diffusion Processes. *Bernoulli*, 11:359–379.

Heston, S. (1993). A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *The Review of Financial Studies*, 6:327–343.

Higham, J. (2002). Computing the Nearest Correlation Matrix: a Problem from Finance. *IMA Journal of Numerical Analysis*, 22:329–343.

Jacod, J., Li, Y., Mykland, P. A., Podolskij, M., and Vetter, M. (2009). Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic Processes and their Applications*, 119(7):2249–2276.

James, W. and Stein, C. (1961). Estimation with Quadratic Loss. In *Proc. Fourth Berkley Symp. on Math. Statist. and Prob.*, volume 1, pages 361–379.

Laurent, S., Rombouts, J., and Violante, F. (2013). On loss functions and ranking forecasting performances of multivariate volatility models. *Journal of Econometrics*, 173(1):1–10.

Lee, S. and Mykland, P. (2008). Jumps in financial markets: A new nonparametric test and jump dynamics. *Review of Financial studies*, 21(6):2535.

Liu, C. and Tang, C. (2012). A Quasi-Maximum Likelihood Approach to Covariance Matrix with High-Frequency Data. working paper, National University of Singapore.

Lo, A. and MacKinlay, C. A. (1990). An econometric analysis of nonsynchronous trading. *Journal of Econometrics*, 45(1-2):181–211.

Mancino, M. and Sanfelici, S. (2011). Estimating covariance via fourier method in the presence of asynchronous trading and microstructure noise. *Journal of Financial Econometrics*, 9(2):367.

Merton, R. C. (1980). On estimating the expected return on the market: an exploratory investigation. *Journal of Financial Economics*, 8:323–61.

Palandri, A. (2006). Consistent realized covariance for asynchronous observations contaminated by market microstructure noise. Unpublished Manuscript.

Patton, A. and Sheppard, K. (2009). Evaluating volatility and correlation forecasts. In Andersen, T., Davis, R., Kreiss, J., and Mikosch, T., editors, *Handbook of Financial Time Series*. Springer, Berlin.

Peluso, S., Corsi, F., and Mira, A. (2013). A Bayesian High-Frequency Estimator of the Multivariate Covariance of Noisy and Asynchronous Returns. working paper, Swiss Finance Institute.

Renò, R. (2003). A closer look at the Epps effect. *International Journal of Theoretical and Applied Finance*, 6(1):87–102.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear gaussian models. *Neural computation*, 11(2):305–345.

Scholes, M. and Williams, J. (1977). Estimating betas from nonsynchronous data. *Journal of Financial Economics*, 5:181–212.

Shephard, N. and Xiu, D. (2012). Econometric analysis of multivariate realised qml: Efficient positive semi-definite estimators of the covariation of equity prices. *Chicago Booth Research Paper*, (12-14).

Sheppard, K. (2006). Realized covariance and scrambling. Unpublished Manuscript.

Shumway, R. and Stoffer, D. (1982). An Approach to Time Series Smoothing and Forecasting using the EM Algorithm. *Journal of Time Series Analysis*, 3:253–264.

Voev, V. and Lunde, A. (2007). Integrated covariance estimation using high-frequency data in the presence of noise. *Journal of Financial Econometrics*, 5:68–104.

West, M. and Harrison, P. (1997). *Bayesian Forecasting and Dynamic Models*. New York: Springer Verlag.

Wu, C. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103.

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

Zhang, L. (2011). Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics*, 160(1):33–47.

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

A Appendix

A.1 E-step optimal distribution

We show that $\pi(\tilde{\mathbf{X}}) = P(\tilde{\mathbf{X}}|\mathbf{Y}^o, Q, R)$ is the optimal distribution which saturates the bound, i.e., $F(\pi, Q, R) = L(Q, R)$.

$$\begin{aligned}
F(\pi, Q, R) &= \int_{\tilde{\mathbf{X}}} \ln\left(\frac{P(\tilde{\mathbf{X}}, \mathbf{Y^o}|Q, R)}{\pi(\tilde{\mathbf{X}})}\right) \pi(\tilde{\mathbf{X}}) d\tilde{\mathbf{X}} \\
&= \int_{\tilde{\mathbf{X}}} \ln\left(\frac{P(\tilde{\mathbf{X}}, \mathbf{Y^o}|Q, R)}{P(\tilde{\mathbf{X}}|\mathbf{Y}^o, Q, R)}\right) P(\tilde{\mathbf{X}}|\mathbf{Y}^o, Q, R) d\tilde{\mathbf{X}} \\
&= \int_{\tilde{\mathbf{X}}} \ln P(\mathbf{Y^o}|Q, R) P(\tilde{\mathbf{X}}|\mathbf{Y}^o, Q, R) d\tilde{\mathbf{X}} \\
&= \ln P(\mathbf{Y^o}|Q, R) \int_{\tilde{\mathbf{X}}} P(\tilde{\mathbf{X}}|\mathbf{Y}^o, Q, R) d\tilde{\mathbf{X}} \\
&= L(Q, R) \cdot 1
\end{aligned}$$

A.2 Kalman filter and smoothing recursions

The Kalman filter recursion formulas are, for $t = 1, \cdots, T$:

$$\begin{aligned}
\mathbf{X}_t^p &= \mathbf{X}_{t-1}^f \\
V_t^p &= V_{t-1}^f + Q \\
K_t &= V_t^p Z_t (Z_t V_t^p Z_t + R)^{-1} \\
\mathbf{X}_t^f &= \mathbf{X}_t^p + K_t(\mathbf{Y}_t - Z_t \mathbf{X}_t^p) \\
V_t^f &= V_t^p - K_t Z_t V_t^p
\end{aligned}$$

where $\mathbf{X}_t^p$ is the predicted value of the log-price latent process, V_t^p is the variance of the prediction error, K_t is the filtering correction term, $\mathbf{X}_t^f$ is the filtered value of the log-price latent process, V_t^f is the variance of the filtering error and Z_t is a diagonal matrix with i-th diagonal element equal to 1 if $Y_{i,t}$ is observed or equal to 0 otherwise, $i = 1, \ldots, d$.

The smoothing recursions are, for $t = T-1, \cdots, 1$:

$$\begin{aligned}
J_t &= V_t^f (V_{t+1}^p)^{-1} \\
\mathbf{X}_t^s &= \mathbf{X}_t^f + J_t(\mathbf{X}_{t+1}^s - \mathbf{X}_t^f) \\
V_t^s &= V_t^f + J_t(V_{t+1}^s - V_{t+1}^p)J_t' \\
VL_{t+1}^s &= V_{t+1}^f J_t' + J_{t+1}(VL_{t+2}^s - V_{t+1}^f)J_t', \qquad t < T-1
\end{aligned}$$

where J_t is the smoothing correction term, $\mathbf{X}_t^s$ is the smoothed log-price latent process, V_t^s is the variance of the smoothing error and VL_t^s is the one-lag autocovariance of the smoothing error.

A.3 KEM sufficient statistics

Thanks to the assumption of diagonal R, the quantities necessary to compute the sufficient statistics are the following:

$$\begin{aligned}
E_r(\mathbf{X}_t\mathbf{X}_t'|\mathbf{Y}^o) &= V_t^s + \mathbf{X}_t^s\mathbf{X}_t^{s'} \\
E_r(\mathbf{X}_t\mathbf{X}_{t-1}'|\mathbf{Y}^o) &= E_r\{(\mathbf{X}_t - \mathbf{X}_t^s)(\mathbf{X}_t - \mathbf{X}_t^s)'|\mathbf{Y}^o\} + \mathbf{X}_t^s(\mathbf{X}_{t-1}^s)' \\
&= VL_t^s + \mathbf{X}_t^s(\mathbf{X}_{t-1}^s)' \\
E_r(\mathbf{Y}_t^o\mathbf{Y}_t^{o'}|\mathbf{Y}^o) &= \mathbf{Y}_t^o\mathbf{Y}_t^{o'} \\
E_r(\mathbf{Y}_t^m\mathbf{Y}_t^{m'}|\mathbf{Y}^o) &= R_t^m + V_t^{m,s} + \mathbf{X}_t^{m,s}\mathbf{X}_t^{m,s'} \\
E_r(\mathbf{Y}_t^o\mathbf{X}_t^{o'}|\mathbf{Y}^o) &= \mathbf{Y}_t^o\mathbf{X}_t^{o,s'} \\
E_r(\mathbf{Y}_t^m\mathbf{X}_t^{m'}|\mathbf{Y}^o) &= V_t^{m,s} + \mathbf{X}_t^{m,s}\mathbf{X}_t^{m,s'} \\
E_r(\mathbf{X}_t^o\mathbf{X}_t^{o'}|\mathbf{Y}^o) &= V_t^{o,s} + \mathbf{X}_t^{o,s}\mathbf{X}_t^{o,s'}.
\end{aligned}$$

where R_t^m is the submatrix of R corresponding to $\mathbf{Y}_t^m - \mathbf{X}_t^m$.

$$Q = \begin{bmatrix} 0.1165 & 0.0109 & 0.0100 & 0.0094 & 0.0090 & 0.0078 & 0.0104 & 0.0071 & 0.0069 & 0.0130 \\ 0.0109 & 0.0570 & 0.0086 & 0.0083 & 0.0075 & 0.0071 & 0.0095 & 0.0067 & 0.0062 & 0.0129 \\ 0.0100 & 0.0086 & 0.0814 & 0.0103 & 0.0075 & 0.0072 & 0.0110 & 0.0062 & 0.0097 & 0.0093 \\ 0.0094 & 0.0083 & 0.0103 & 0.0722 & 0.0076 & 0.0066 & 0.0101 & 0.0061 & 0.0076 & 0.0093 \\ 0.0090 & 0.0075 & 0.0075 & 0.0076 & 0.0561 & 0.0118 & 0.0076 & 0.0059 & 0.0071 & 0.0085 \\ 0.0078 & 0.0071 & 0.0072 & 0.0066 & 0.0118 & 0.0398 & 0.0069 & 0.0055 & 0.0065 & 0.0075 \\ 0.0104 & 0.0095 & 0.0110 & 0.0101 & 0.0076 & 0.0069 & 0.0719 & 0.0062 & 0.0081 & 0.0103 \\ 0.0071 & 0.0067 & 0.0062 & 0.0061 & 0.0059 & 0.0055 & 0.0062 & 0.0342 & 0.0046 & 0.0069 \\ 0.0069 & 0.0062 & 0.0097 & 0.0076 & 0.0071 & 0.0065 & 0.0081 & 0.0046 & 0.0681 & 0.0070 \\ 0.0130 & 0.0129 & 0.0093 & 0.0093 & 0.0085 & 0.0075 & 0.0103 & 0.0069 & 0.0070 & 0.0540 \end{bmatrix}$$

,

$$R = diag(0.0505, 0.0222, 0.2011, 0.0937, 0.1425, 0.0822, 0.0606, 0.1040, 0.1719, 0.0072)$$

Table 1: 10-dimensional annualized Q and R matrices used to generate the data in our simulations. The estimates are obtained through the application of the KEM methodology to a subset of our 100-dimensional real data set, averaged over time.

setting	avg noise-to-signal	avg missings prob	std dev missings
Standard	0.78	0.32	0.11
High noise	2.58	0.32	0.11
High missings	0.78	0.67	0.11
High missings, high noise	2.58	0.67	0.11
Dispersed missings	0.78	0.49	0.35
Dispersed missings, high noise	2.58	0.49	0.35

Table 2: Simulation settings. Setting missing probabilities $v = \{1/2, 1/3, 1/2, 1/4, 1/4, 1/3, 1/5, 1/4, 1/3, 1/4\}$, and matrices Q and R as in Table 1, the simulation scenarios are the following: (a) *Standard*: missing probabilities v and noise matrix R. (b) *High noise*: missing probabilities v and noise matrix $R + 0.35$. (c) *High missings*: missing probabilities $v + 0.35$ and noise matrix R. (d) *High missings, high noise*: missing probabilities $v + 0.35$ and noise matrix $R + 0.35$. (e) *Dispersed missings*: more dispersed missing probabilities $w = \{0, 0.5, 0.8, 0.9, 0.25, 0, 0.5, 0.8, 0.9, 0.25\}$ and noise matrix R. (f) *Dispersed missings, high noise*: missing probabilities w and noise matrix $R + 0.35$.

	KEM	HY	AFX	MRK
(a) *Standard*				
Mean	0.0185	0.0262	0.0222	0.0351
Std	0.0023	0.0031	0.0027	0.0043
MCS deletion rank		2	3	1
(b) *High noise*				
Mean	0.0264	0.0807	0.0286	0.0479
Std	0.0032	0.0086	0.0040	0.0067
MCS deletion rank		1	3	2
(c) *High missings*				
Mean	0.0275	0.0318	0.0337	0.0472
Std	0.0053	0.0040	0.0043	0.0059
MCS deletion rank		3	2	1
(d) *High missings, high noise*				
Mean	0.0347	0.0592	0.0405	0.0625
Std	0.0042	0.0062	0.0052	0.0085
MCS deletion rank		1	3	2
(e) *Dispersed missings*				
Mean	0.0259	0.0315	0.0350	0.0532
Std	0.0039	0.0040	0.0048	0.0068
MCS deletion rank		3	2	1
(f) *Dispersed missings, high noise*				
Mean	0.0337	0.0656	0.0415	0.0678
Std	0.0042	0.0070	0.0054	0.0095
MCS deletion rank		1	3	2

Table 3: Expected values and standard deviations of the Frobenius distances between the estimated and the true covariance matrix in the simulation settings summarized in Table 2. MCS rank denotes the place in the deletion sequence when constructing the model confidence set (MCS) of the different approaches; 1,2, and 3 indicate that the model is eliminated from the MCS at all relevant confidence levels in the first, second, and third step, respectively. The competing methodologies are the newly introduced KEM approach (KEM), a covariance matrix obtained with Two Scale estimators of Zhang et al. (2005) for the variance elements and the pairwise covariance terms estimated by Hayashi and Yoshida (2005) (HY), the pairwise estimator proposed by Ait-Sahalia et al. (2010) (AFX), and the Multivariate Realized Kernel of Barndorff-Nielsen et al. (2011) (MRK). HY and AFX are projected to impose positiveness of the estimated covariance matrices.

Name	Symbol	Mean	Std	Skew	Kurt	nobs ($\times 10^3$)	% Miss
Apple Inc.	AAPL	4.6669	0.0076	-0.0619	2.7622	13.4291	0.4261
Exxon Mobil Corp.	XOM	4.3357	0.0050	-0.0942	2.7159	14.2379	0.3916
Microsoft Corp.	MSFT	3.3179	0.0055	-0.0052	3.0515	8.3028	0.6452
IBM Corp.	IBM	4.5896	0.0043	-0.0488	3.1178	9.0729	0.6123
General Electric Company	GE	3.4894	0.0045	0.0929	2.9935	13.1563	0.4378
Biogen Idec Inc.	BIIB	3.9514	0.0062	-0.0139	2.9312	3.8158	0.8369
EOG Resources Inc.	EOG	4.3789	0.0083	-0.0303	2.7425	6.6027	0.7178
Prudential Financial Inc.	PRU	4.3219	0.0073	-0.0050	2.8226	5.6944	0.7567
The TJX Companies, Inc.	TJX	3.3298	0.0059	-0.0140	2.9776	5.7315	0.7551
Dell Inc.	DELL	3.1348	0.0065	-0.0087	2.8379	5.9364	0.7463

Table 4: Summary statistics of 10 US stocks' tick-by-tick log prices for the period January 1, 2006 to December 31, 2008, downloaded from TAQ. The columns report for each stock: name, symbol, average mean log price per day, average standard deviation of log price per day, average skewness of log price per day, average kurtosis of log price per day, average number of observations per day, average % of missing observations per day.

Panel A: Pairwise tests on portfolio variances

KEM against	Mean	Efficiency Gain
HY	0.0103*	1.0745
	(0.0014)	
AFX	0.0363*	3.0788
	(0.0062)	
MRK	0.0060*	0.5103
	(0.0021)	

Panel B: Model Confidence Set results

	EPA test statistic	
	range	semi-quadratic
1^{st} step:	6.6899 (0)	97.4815 (0)
2^{nd} step:	6.6641 (0)	51.7604 (0)
3^{rd} step:	2.3759 (0.0185)	5.6449 (0.0187)

Model	Worst performing index		
	1^{st} step	2^{nd} step	3^{rd} step
KEM	-5.6828	-6.3973	-2.3759
HY	-1.5234	3.1579	–
AFX	3.9709	–	–
MRK	-2.3175	0.3025	2.3759

Table 5: Panel A: Results of pairwise tests on portfolio variances constructed from one hundred US stocks belonging to the S&P500 universe for the null-hypothesis of equal predictive ability between the newly introduced KEM approach (KEM) and different competitors: a covariance matrix obtained with Two Scale estimators of Zhang et al. (2005) for the variance elements and the pairwise covariance terms estimated by Hayashi and Yoshida (2005) (HY), the pairwise estimator proposed by Ait-Sahalia et al. (2010) (AFX), and the Multivariate Realized Kernel of Barndorff-Nielsen et al. (2011) (MRK). HY and AFX are projected to impose positiveness of the estimated covariance matrices. The out-of-sample period goes from January 2006 to December 2008. Heteroscedastic consistent standard errors are reported in parentheses. Asterisks denote rejection of the null-hypothesis at the 0.01 significance level. Positive values are in favor of the KEM approach.
Panel B: Results of the model confidence set approach. Upper panel: Values of equal predictive ability (EPA) tests using the range statistic and the semi-quadratic statistic. Corresponding p-values computed using a block-bootstrap procedure with $10,000$ replications are given in parentheses. Lower panel: Worst performing index results for the construction of the confidence model sets. If the null hypothesis of EPA is rejected, the model with the largest worst performing index value is eliminated.

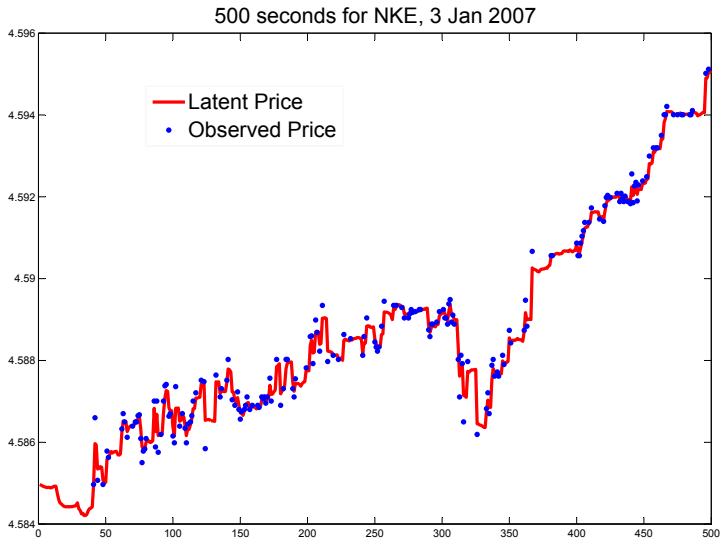

Figure 1: Reconstructed latent log-price process with KEM and observed log prices for the Nike Inc. stock (NKE). Shown are the first 500 seconds on January 3, 2007.

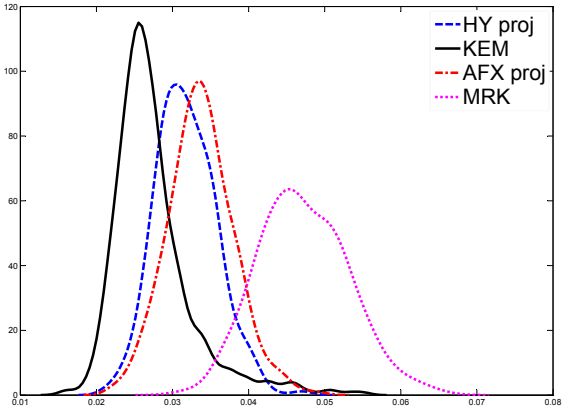

Figure 2: Kernel density estimates of the Frobenius distances: high missings setting. $M = 500$ simulated values of $Frob_k = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d}(\hat{\sigma}_{ij}^k - \sigma_{ij})^2}$, where $\hat{\sigma}_{ij}^k$ is the ij-th element of Q estimated with methodology k, $k \in \{KEM, HY, AFX, MRK\}$ and σ_{ij} is the ij-th element of the true Q used in the simulation. The competing methodologies are the newly introduced KEM approach (KEM), a covariance matrix obtained with Two Scale estimators of Zhang et al. (2005) for the variance elements and the pairwise covariance terms estimated by Hayashi and Yoshida (2005) (HY), the pairwise estimator proposed by Ait-Sahalia et al. (2010) (AFX), and the Multivariate Realized Kernel of Barndorff-Nielsen et al. (2011) (MRK). If necessary, HY and AFX are projected to impose positiveness.

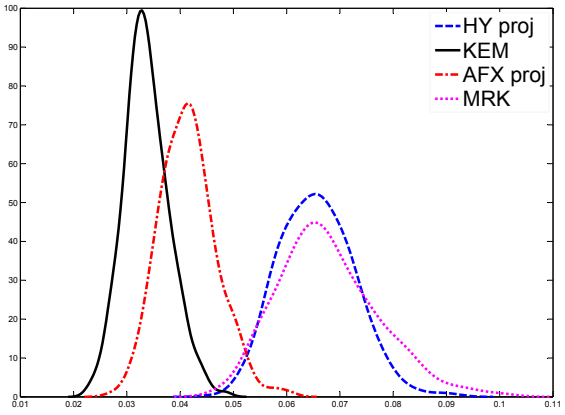

Figure 3: Kernel density estimates of the Frobenius distances: dispersed missings, high noise setting. $M = 500$ simulated values of $Frob_k = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d}(\hat{\sigma}_{ij}^k - \sigma_{ij})^2}$, where $\hat{\sigma}_{ij}^k$ is the ij-th element of Q estimated with methodology k, $k \in \{KEM, HY, AFX, MRK\}$ and σ_{ij} is the ij-th element of the true Q used in the simulation. The competing methodologies are the newly introduced KEM approach (KEM), a covariance matrix obtained with Two Scale estimators of Zhang et al. (2005) for the variance elements and the pairwise covariance terms estimated by Hayashi and Yoshida (2005) (HY), the pairwise estimator proposed by Ait-Sahalia et al. (2010) (AFX), and the Multivariate Realized Kernel of Barndorff-Nielsen et al. (2011) (MRK). If necessary, HY and AFX are projected to impose positiveness.

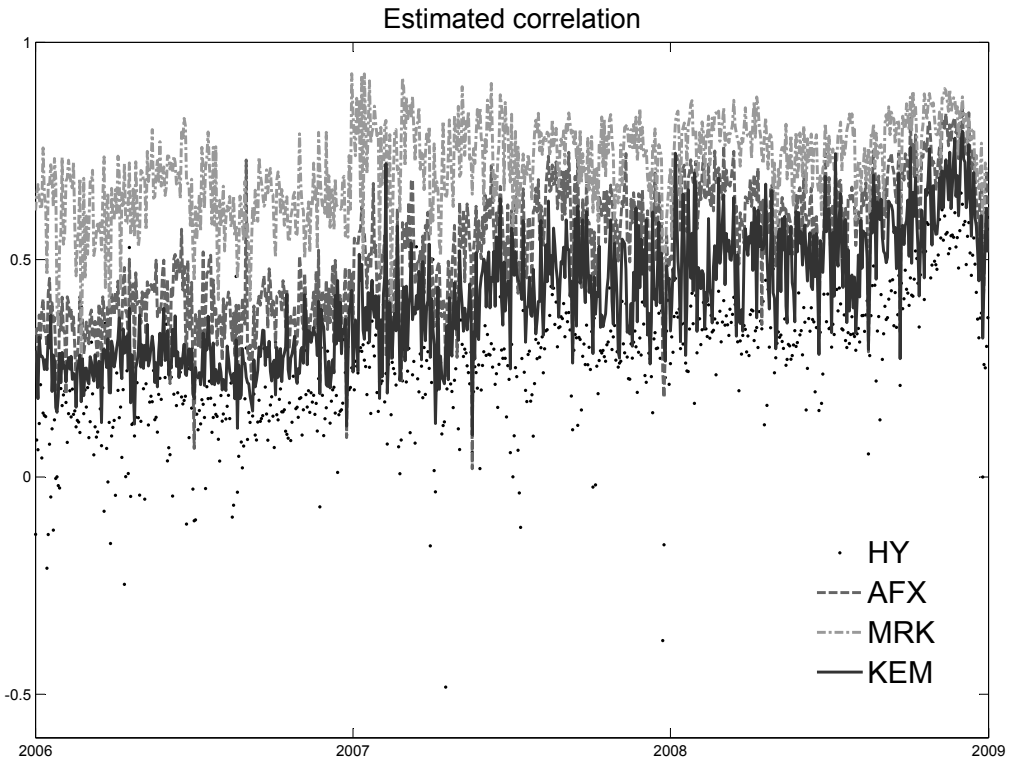

Figure 4: Estimated annualized correlation between Exxon Mobile Corporation (XOM) and Chevron Corporation (CVX). KEM is the newly proposed KEM estimator, AFX is the projected pairwise estimator proposed by Ait-Sahalia et al. (2010), HY the projected covariance matrix obtained with Two Scale estimators of Zhang et al. (2005) for the variance elements and the pairwise covariance terms estimated by Hayashi and Yoshida (2005), MRK is the Multivariate Realized Kernel estimator proposed in Barndorff-Nielsen et al. (2011).

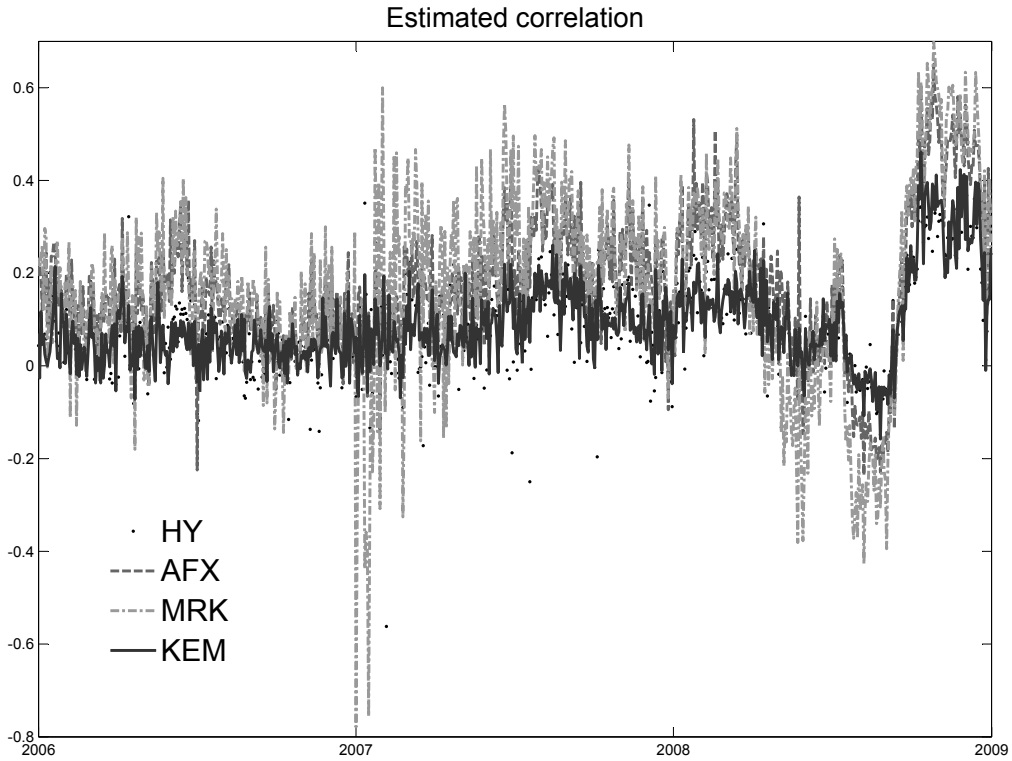

Figure 5: Estimated annualized correlation between Lowe's Companies, Inc. (LOW) and Apache Corp. (APA). KEM is the newly proposed KEM estimator, AFX is the projected pairwise estimator proposed by Ait-Sahalia et al. (2010), HY the projected covariance matrix obtained with Two Scale estimators of Zhang et al. (2005) for the variance elements and the pairwise covariance terms estimated by Hayashi and Yoshida (2005), MRK is the Multivariate Realized Kernel estimator proposed in Barndorff-Nielsen et al. (2011).