Universität St.Gallen

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

Fulvio Corsi, Stefano Peluso and Francesco Audrino

January 2012 Discussion Paper no. 2012-2

School of Economics and Political Science,
Department of Economics

University of St. Gallen

Editor: Martina Flockerzi
University of St. Gallen
School of Economics and Political Science
Department of Economics
Varnbüelstrasse 19
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35
Email seps@unisg.ch

Publisher: School of Economics and Political Science
Department of Economics
University of St. Gallen
Varnbüelstrasse 19
CH-9000 St. Gallen
Phone +41 71 224 23 25

Electronic Publication: Fax +41 71 224 31 35
http://www.seps.unisg.ch

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

Fulvio Corsi, Stefano Peluso and Francesco Audrino

Authors' address:

Dr. Fulvio Corsi
University of St. Gallen and Swiss Finance Institute
Bodanstrasse 6
CH-9000 St. Gallen
Email fulvio.corsi@usi.ch

Stefano Peluso
University of Lugano and Swiss Finance Institute
Via Buffi 13
CH-6900 Lugano
Email stefano.peluso@usi.ch

Prof. Dr. Francesco Audrino
University of St. Gallen
Bodanstrasse 6
CH-9000 St. Gallen
Phone +41 71 2242431
Fax +41 71 2242894
Email francesco.audrino@unisg.ch

Abstract

Motivated by the need for an unbiased and 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. In the expectation step, by means of the Kalman filter with missing data, we reconstruct the smoothed and synchronized series of the latent price processes. In the maximization step, we search for covariance matrices that maximize the expected likelihood obtained with the reconstructed price series. Iterating between the two EM steps, we obtain a KEM-improved covariance matrix estimate which is robust to both asynchronicity and microstructure noise, and positive-semidefinite by construction. Extensive Monte Carlo simulations show the superior performance of the KEM estimator over several alternative covariance matrix estimates introduced in the literature. The application of the KEM estimator in practice is illustrated on a 10-dimensional US stock data set.

Keywords

High frequency data; Realized covariance matrix; Market microstructure noise; Missing data; Kalman filter; EM algorithm; Maximum likelihood.

JEL Classification

C13; C51; C52; C58.

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 forecast 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 which tends to increase with the sampling frequency. This is because in each regularly spaced interval, any difference in the time stamps of the last ticks of the two assets will correspond to a portion of the cross product returns that will not be accounted for in the computation of the realized covariance. In fact, the returns corresponding to this time difference are not matched (being ascribed to two different time intervals) and hence no longer contribute to the cross product summation. This reduction of the correlations absolute value when increasing 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

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

2005; Palandri 2006; Sheppard 2006; Voev and Lunde 2007; Corsi and Audrino 2008; Hautsch et al. 2009; Griffin and Oomen 2011), adopt the so-called refresh time scheme (Barndorff-Nielsen et al. 2011; Ait-Sahalia et al. 2010; Zhang 2011), and the multivariate Fourier method (Renò 2003; Mancino and Sanfelici 2011).

Third, the variance-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 multivariate Fourier method of Mancino and Sanfelici (2011). In both cases, however, the frequency at which all the realized variance-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.

In this paper, we propose a novel approach to the estimation of the variance-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 (say at one second) 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 might be missing. The advantage of this point of view is that standard statistical methodology used for dealing with missing observations can be employed to cope with the asynchronicity problem in asset prices. Moreover, modelling 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 modelling 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

[2]Alternatively, a Bayesian approach for sampling the missing observations can be employed using a Gibbs Sampler, as proposed in Peluso et al. (2011).

guarantees maximizing the likelihood of the observed data even in presence of missing observations (see Dempster et al. 1977). Therefore, the proposed estimators can be seen as an application of the QMLE approach to quadratic variation 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 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 the estimated variance-covariance matrix to be psd. Through an extensive simulation study, we show that the KEM approach is able to effectively deal, at the same time, with both the asynchronicity (or missing value) problem and the market microstructure noise, providing realized covariance matrices which are both psd by construction and the most accurate of the competing methodologies considered. Furthermore, 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.

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 competitive approaches over a broad range of settings. Section 5 contains the empirical application of our proposed estimators to a portfolio of 10 stocks, and Section 6 concludes.

2 Model

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

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

where $\mathbf{X}(t)$ is the d-dimensional latent log-price process free of noise, $\mathbf{W}(t)$ is the d-dimensional Brownian motion, and the drift μ_t and the diffusion coefficient $\Sigma_t\Sigma_t' = Q_t$ are functions smooth enough to guarantee the existence of a unique solution to (1).

However, the observed log transaction prices are 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 a ultra-high-frequency grid (of one second in the simulation and empirical analysis) with the assumption of zero drift on the latent price and constant variance-covariance matrices. 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.

Assuming a drift equal to zero is a reasonable assumption if we consider that we are dealing with infra-day log-prices. Furthermore, we recognize that the volatility is a multivariate time-varying process but since we are interested in the estimation of a constant volatility matrix, we apply the recent approach of Xiu (2010) that shows that the QMLE of the volatility of a misspecified model with constant volatility remains consistent and optimal in terms of its rate of convergence under fairly general assumptions. Finally, we assume the covariance matrix R to be diagonal, meaning that the microstructure noises are uncorrelated across assets.[3]

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: the Kalman-EM algorithm

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

[3]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.

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:

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

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

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

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). Being $P(\tilde{\mathbf{X}}|\mathbf{Y}^o, Q_k, R_k)$ 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 looks as follows:

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

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

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

where $\mathbf{X}_t^s$ is the smoothed log-price latent process, $\mathbf{X}_t^{o,s}$ 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 $\mathbf{Y}_t^o$ (for

the expression of the smoothed quantities $\mathbf{X}_t^s$ and V_t^s 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 observed and then 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*
- *use 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 $V_t^{o,f}$ are the submatrices of, respectively, R and V_t^f, corresponding to $\mathbf{Y}_t^o$ (for the expression of the filtered quantities $\mathbf{X}_t^f$, V_t^f, and V_t^p 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 (or keep the same) 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).

As simply a more convenient and computationally efficient way of maximizing the likelihood of the

observed data, the KEM estimators of Q and R are ML estimators and therefore possess all requested asymptotic properties, that is, consistency, asymptotic normality and efficiency. As mentioned above, the MLE of the constant matrix Q can be interpreted as the QMLE for the quadratic variation of the (multivariate) efficient price process[4] by applying the results proved in Xiu (2010) to the reconstructed series of the synchronized efficient price $\mathbf{X}$.

Formally, we state the consistency of the KEM estimator in the following proposition, whose proof is given in Appendix A.4.

Proposition 3.1 *Suppose the following conditions are satisfied:*

- *the underlying multivariate latent price process satisfies* $d\boldsymbol{X}(t) = \Sigma_t d\boldsymbol{W}(t)$, *with* Σ_t *a positive and locally bounded Itô semimartingale process such that* $Q_t = \Sigma_t \Sigma_t'$,
- *the observed data log likelihood* $L(Q, R)$ *in* (8) *is unimodal and with a unique stationary point,*
- *the expected complete data log likelihood* $\mathbb{E}[\log P(\mathbf{X}, \mathbf{Y}|Q_k, R_k)]$, *with expectation taken with respect to the distribution of* $\tilde{\mathbf{X}}|\mathbf{Y^o}, Q_k, R_k$, *is continuously differentiable in* Q_k *and* R_k, *where* Q_k *and* R_k *are the EM estimates at iteration* k *given in* (6) *and* (7).

Then, given M_j *the number of observed prices* $Y^o_{i,j}$ *for the* j*-th asset in the interval* $[0, \tau]$, *and* $M := \min(M_1, \ldots, M_j, \ldots, M_d)$,

$$\left| \hat{Q} - \frac{1}{\tau} \int_0^\tau Q_t dt \right| \xrightarrow{p} 0 \quad as \quad M \to +\infty,$$

where $\hat{Q}$ *is the steady state version of the estimator* (6) *on the misspecified dynamics* $d\boldsymbol{X}(t) = \Sigma d\boldsymbol{W}(t)$, *with* $\Sigma\Sigma' = Q$ *constant.*

This result shows the consistency of the KEM estimator for the unknown quadratic covariation process $\int_0^\tau Q_t dt$. In our empirical investigations, τ will be fixed to one day.

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

[4] As such, in the presence of jumps, KEM will estimate the sum of the contribution to the quadratic variation coming from both the continuous part and the jump part, i.e. it will estimates the integrated volatility plus the sum of square jumps for variances and the integrated covariation plus the sum of cojumps for covariances. In this setting, the problem of separating the continuous part from the jump part could be tackled by pre-testing the data for jumps using jump identification tests which are able to locate the position of jumps inside the day (such as Fan and Wang 2007; Lee and Mykland 2008).

reconstruct the latent dynamics of the efficient price by exploiting the correlations of one asset with the dynamics of all the other assets. This multivariate signal extraction is particularly useful for the less liquid series that have fewer observations and can therefore benefit more from the information contained in the dynamics of the more liquid assets. As an illustration of that, in Figures 1 and 2 we plot the reconstructed latent price process together with the observed tick prices, for two of the least liquid assets in our empirical application (Hasbro and Nike, see Section 5 for more details) over the first 500 seconds of the trading day January 3, 2007.

4 Simulation Study

In this section, by means of an extensive simulation analysis, we compare the performance of our KEM estimator with other estimators proposed in the literature. 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 Two Scale estimators of Zhang et al. (2005) for the variance elements and the pairwise covariance terms estimated by Hayashi and Yoshida (2005) (HY). We generate the data using the stochastic volatility model of Heston (1993). The data generating process (DGP) is, for $i = 1, \cdots, d$ and $t = 1, \cdots, T$

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

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

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, by extrapolation, we assume a computing time of a few hours for hundreds of assets, using all the data in one trading day.[5] All simulations are initialized from $P_0 = log([100, 40, 60, 80, 40, 20, 90, 30, 50, 60])$.

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

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

one second frequency, is psd by construction. If the variance-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).[6] 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 are reported in Table 1 and have been obtained by averaging the results of the application of our KEM methodology to the 10 stocks employed in our empirical analysis (see Section 5).

We perform the study in six simulations 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. As a synthesis measure of the performance we choose the Frobenius norm of the matrix difference between the estimated and the true (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 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.

The sample means and standard deviations of the Frobenius distances are summarized in Table 3. In Figures 3 to 8 we graphically compare the kernel density of the $M = 500$ simulated Frobenius distances obtained from each competing estimator.

Starting from the standard setting featuring a moderate level of microstructure noise and a high and homogeneous level of liquidity across the assets, Figure 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 (even for the less liquid ones) leads to more accurate realized covariance estimates (while maintaining the positive-semidefiniteness of it). These results also confirm that MRK guarantees the positive-semidefiniteness of the estimated covariance matrix at the price of discarding a considerable amount of information.

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

In the high noise setting, the noise-to-signal ratio is raised from the 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,since it is not robust to microstructure noise, is the worst performing while KEM is again the best estimator although now closely followed by AFX which then proves in addition to be highly robust to microstructure noise (see Figure 4).

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 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 5).

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, being better equipped to deal with noise. However, the KEM estimator still remains the preferred one, confirming its robustness to both sources of errors (Figure 6).

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 (Figures 7 and 8). 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 under both settings. 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 especially from the more liquid ones (hence, doing somewhat the opposite of what KEM tends to do). 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 performances.

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 has been 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, in line with the results illustrated in Figures 3 to 8, 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).

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 variance-covariance matrix comes at no cost in terms of precision of the estimates.

5 Application

In our empirical analysis we use a data set from tickdata.com consisting of tick-by-tick data for the 10 stocks described in Table 4 over the period 3 Jan 2007 - 21 Nov 2007. The data set contains both liquid stocks (such as Exxon, Citigroup, and Microsoft) with an average probability of missing at the one second frequency of no more than 0.7 and much less liquid stocks (such as Hasbro, Harley Davidson, Nike Inc. and Tektronix) with an average probability of missing larger than 0.8.

The estimated (with KEM) average (across stocks) signal-to-noise ratio for our empirical data is 1.23, with mean and standard deviation of missing probabilities equal to 0.78 and 0.11, respectively. So *High missings* is the simulation scenario closest to our empirical data. From the simulation results

in the previous section, we found that MRK was the least accurate among the considered estimators due to the large amount of information discarded, as discussed above. This was confirmed by the fact that MRK was the first methodology deleted in the MCS analysis. For the sake of readability of the results, in the following empirical analysis we then decided not to report the results of MRK.

In Tables 5 and 6 we report the average estimated variance matrix for AFX and HY, respectively, both projected onto the space of psd matrices; while the average covariance matrix Q estimated with KEM was already reported in Table 1.

As illustrative pictures, in Figures 9 and 10 we plot the time series of the daily estimates of, respectively, the annualized variance of Intel (INTC) and the annualized covariance between Alcoa (AA) and Exxon Mobile Corporation (XOM), estimated with KEM, AFX and HY (the last two methodologies are projected to impose the positiveness of the resulting matrix). Similar plots can be obtained for all other (combinations of) stocks.

For the variance, the three methods return estimates which are almost indistinguishable, showing that the KEM methodology also provides realized variance estimates which are comparable to those provided by the renowned Two Scale estimators of Zhang et al. (2005). Although a performance comparison on empirical data is always problematic (lacking a reference target), we can make several observations. First, all three considered estimators identify a surge in the value of covariance corresponding to the financial turmoil of August and October 2007. Second, KEM covariance estimates tend to be slightly higher than HY. This could be a consequence of HY downward bias on empirical data recently documented by Griffin and Oomen (2011). Third, during periods of higher covariance, the AFX estimator appears to be noisier, generating large and extreme (compared to HY and KEM) values and a much more erratic covariance dynamics. We can therefore conclude that the empirical results align with those obtained in our Monte Carlo simulation analysis.

6 Conclusions

In this paper we proposed a novel 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: reconstruct the smoothed and synchronized series of the latent price processes (E-step) and use this fictitious complete data set to easily maximize the complete data likelihood obtaining new estimates for the parameters of interest, i.e. the variance-covariance matrix of the latent price process and 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 portfolio applications such as portfolio selection and risk management.

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. and Shephard, N. (2004). Econometric analysis of realized covariation; high frequency based covariance, regression, and correlation in financial economics. *Econometrica*, 72:885–925.

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.

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.

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. (2008). Realized covariance tick-by-tick in presence of rounded time stamps and general microstructure effects. Unpublished manuscript, University of St. Gallen.

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.

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.

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., 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. (2009). A blocking and regularization approach to high-dimensional realized covariance estimation. *Journal of Applied Econometrics*.

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.

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

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.

Peluso, S., Corsi, F., and Mira, A. (2011). 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.

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 $t = 1, \cdots, T$:

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

where $\mathbf{X}_t^p$ is the predicted value of the log-prices 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 and V_t^f is the variance of the filtering error.

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

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

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 Vt_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:

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

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

A.4 Proof of Proposition 3.1

Suppose that $L(Q,R)$ is unimodal with (Q^*,R^*) being the only stationary point and that $\mathbb{E}[\log P(\mathbf{X},\mathbf{Y}|Q_k,R_k)]$ is continuously differentiable in Q_k and R_k. Then, by Corollary 1 of Wu (1983), for any EM sequence $\{Q_k,R_k\}$, (Q_k,R_k) converges to $(\hat{Q},\hat{R})$, that is the unique maximizer (Q^*,R^*) of $L(Q,R)$. Furthermore, let us define the t multivariate intra-τ return of the efficient log-price process as $\boldsymbol{X}_t-\boldsymbol{X}_{t-1},\ t=1,\ldots,M$. Then

$$\begin{aligned}\frac{1}{\tau}\sum_{t=1}^{M}(\boldsymbol{X}_t-\boldsymbol{X}_{t-1})(\boldsymbol{X}_t-\boldsymbol{X}_{t-1})' &= \arg\max_{Q} P(\mathbf{X},\mathbf{Y}|Q,R)\\ &= \arg\max_{Q}\int P(\tilde{\mathbf{X}},\mathbf{Y}^o|Q,R)d\tilde{\mathbf{X}}\\ &= \arg\max_{Q} L(Q,R).\end{aligned}$$

Then, $\frac{1}{\tau}\sum_{t=1}^{M}(\boldsymbol{X}_t-\boldsymbol{X}_{t-1})(\boldsymbol{X}_t-\boldsymbol{X}_{t-1})' \xrightarrow{p} Q^*$ by uniqueness of the maximizer of $L(Q,R)$. This means that $\hat{Q}$ and $\frac{1}{\tau}\sum_{t=1}^{M}(\boldsymbol{X}_t-\boldsymbol{X}_{t-1})(\boldsymbol{X}_t-\boldsymbol{X}_{t-1})'$ are asymptotically equivalent, and we know from Theorem 1 in Barndorff-Nielsen and Shephard (2004) that, for $M\to+\infty$,

$$\left|\sum_{t=1}^{M}(\boldsymbol{X}_t-\boldsymbol{X}_{t-1})(\boldsymbol{X}_t-\boldsymbol{X}_{t-1})'-\int_0^{\tau}Q_t dt\right|\xrightarrow{p}0.$$

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

,

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

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 our data-set for the period January 3, 2007 to November 21, 2007, 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 new 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 ($\times 10^3$)	Prob Miss
Alcoa Inc.	AA	3.5941	0.0055	0.0230	2.6120	5.4715	0.7662
Citigroup Inc.	C	3.9024	0.0046	0.1278	2.7887	7.7763	0.6677
Hasbro Inc.	HAS	3.3851	0.0048	-0.0604	3.6145	2.1984	0.9060
Harley Davidson	HDI	4.0729	0.0045	0.0256	2.8292	2.7621	0.8820
Intel Corp.	INTC	3.1323	0.0042	-0.0723	2.7104	6.9434	0.7033
Microsoft Corp.	MSFT	3.3964	0.0035	0.0293	2.8854	6.9505	0.7030
Nike Inc.	NKE	4.2077	0.0042	0.0383	2.9120	3.1495	0.8654
Pfizer Inc.	PFE	3.2369	0.0032	0.0514	2.7540	6.6511	0.7158
Tektronix	TEK	3.4380	0.0030	-0.0518	3.2857	1.0143	0.9567
Exxon Mobil Corp.	XOM	4.4071	0.0043	-0.1611	2.5500	8.3083	0.6449

Table 4: Summary statistics of 10 US stocks tick-by-tick log prices for the period January 3, 2007 to November 21, 2007, downloaded from tickdata.com. The columns report for each stock: name, symbol, average missing probability per day, 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.

$$Q = \begin{bmatrix} 0.1169 & 0.0107 & 0.0093 & 0.0096 & 0.0087 & 0.0070 & 0.0106 & 0.0070 & 0.0064 & 0.0125 \\ 0.0107 & 0.0572 & 0.0083 & 0.0081 & 0.0071 & 0.0069 & 0.0091 & 0.0065 & 0.0058 & 0.0124 \\ 0.0093 & 0.0083 & 0.0815 & 0.0099 & 0.0073 & 0.0073 & 0.0112 & 0.0061 & 0.0095 & 0.0092 \\ 0.0096 & 0.0081 & 0.0099 & 0.0722 & 0.0072 & 0.0065 & 0.0096 & 0.0061 & 0.0073 & 0.0091 \\ 0.0087 & 0.0071 & 0.0073 & 0.0072 & 0.0562 & 0.0112 & 0.0074 & 0.0059 & 0.0067 & 0.0081 \\ 0.0070 & 0.0069 & 0.0073 & 0.0065 & 0.0112 & 0.0400 & 0.0070 & 0.0054 & 0.0065 & 0.0074 \\ 0.0106 & 0.0091 & 0.0112 & 0.0096 & 0.0074 & 0.0070 & 0.0723 & 0.0058 & 0.0074 & 0.0099 \\ 0.0070 & 0.0065 & 0.0061 & 0.0061 & 0.0059 & 0.0054 & 0.0058 & 0.0342 & 0.0044 & 0.0067 \\ 0.0064 & 0.0058 & 0.0095 & 0.0073 & 0.0067 & 0.0065 & 0.0074 & 0.0044 & 0.0678 & 0.0070 \\ 0.0125 & 0.0124 & 0.0092 & 0.0091 & 0.0081 & 0.0074 & 0.0099 & 0.0067 & 0.0070 & 0.0539 \end{bmatrix}$$

Table 5: 10-dimensional annualized diffusion Q matrix estimated using the pairwise estimator proposed by Ait-Sahalia et al. (2010) to our data-set for the period January 3, 2007 to November 21, 2007, projected to impose positiveness and averaged over time.

$$Q = \begin{bmatrix}
0.1161 & 0.0108 & 0.0092 & 0.0090 & 0.0091 & 0.0073 & 0.0102 & 0.0071 & 0.0065 & 0.0125 \\
0.0108 & 0.0574 & 0.0084 & 0.0079 & 0.0073 & 0.0069 & 0.0092 & 0.0064 & 0.0057 & 0.0125 \\
0.0092 & 0.0084 & 0.0811 & 0.0104 & 0.0074 & 0.0071 & 0.0107 & 0.0057 & 0.0097 & 0.0089 \\
0.0090 & 0.0079 & 0.0104 & 0.0726 & 0.0073 & 0.0065 & 0.0096 & 0.0060 & 0.0073 & 0.0088 \\
0.0091 & 0.0073 & 0.0074 & 0.0073 & 0.0562 & 0.0115 & 0.0076 & 0.0050 & 0.0062 & 0.0080 \\
0.0073 & 0.0069 & 0.0071 & 0.0065 & 0.0115 & 0.0399 & 0.0067 & 0.0051 & 0.0063 & 0.0072 \\
0.0102 & 0.0092 & 0.0107 & 0.0096 & 0.0076 & 0.0067 & 0.0723 & 0.0054 & 0.0078 & 0.0099 \\
0.0071 & 0.0064 & 0.0057 & 0.0060 & 0.0050 & 0.0051 & 0.0054 & 0.0342 & 0.0044 & 0.0068 \\
0.0065 & 0.0057 & 0.0097 & 0.0073 & 0.0062 & 0.0063 & 0.0078 & 0.0044 & 0.0680 & 0.0068 \\
0.0125 & 0.0125 & 0.0089 & 0.0088 & 0.0080 & 0.0072 & 0.0099 & 0.0068 & 0.0068 & 0.0538
\end{bmatrix}$$

Table 6: 10-dimensional annualized diffusion Q 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) to our data-set for the period January 3, 2007 to November 21, 2007, projected to impose positiveness and averaged over time.

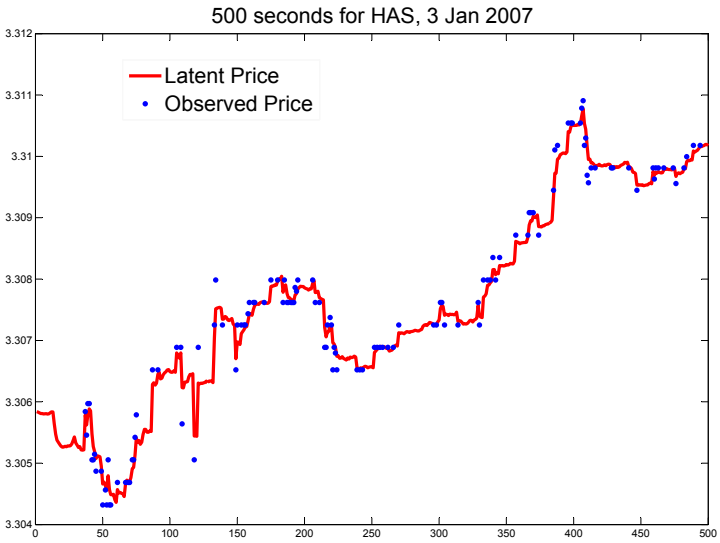

Figure 1: Reconstructed latent log-price process with KEM and observed log prices for the Hasbro Inc. stock (HAS). Shown are the first 500 seconds on January 3, 2007.

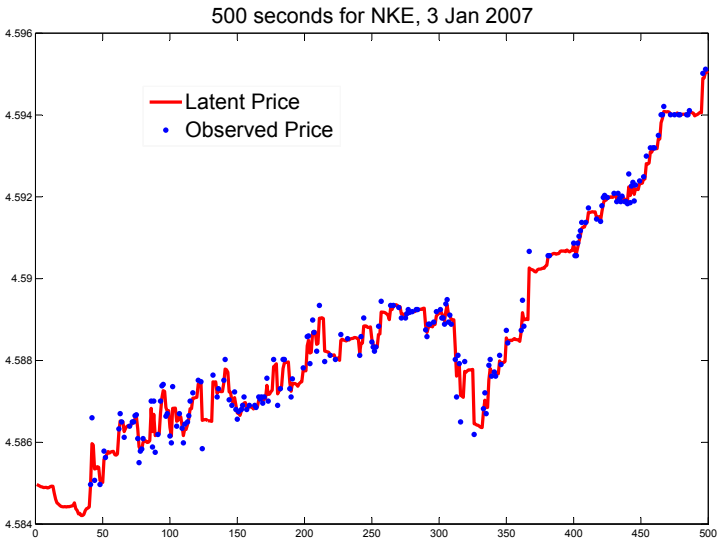

Figure 2: Reconstructed latent log-price process with KEM and observed log prices for the Nike stock (NKE). Shown are the first 500 seconds on January 3, 2007.

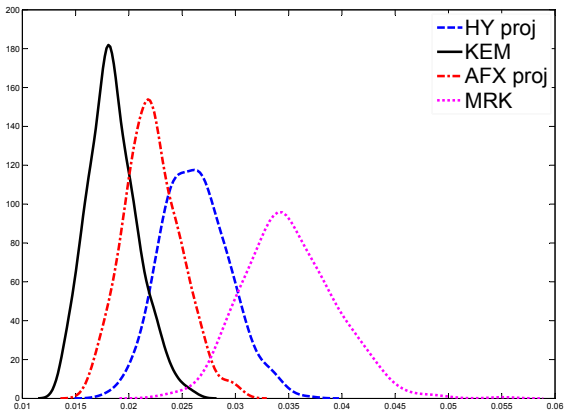

Figure 3: Kernel density estimates of the Frobenius distances: standard setting. $M = 500$ simulated values of $Frob_k = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d}(\hat{\sigma}_{ij}^k - \sigma_{ij})^2}$, where $\hat{\sigma}_{ij}^k$ is the ij-th element of Q estimated with methodology k, $k \in \{KEM, HY, AFX, MRK\}$ and σ_{ij} is the ij-th element of the true Q used in the simulation. The competing methodologies are the new 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.

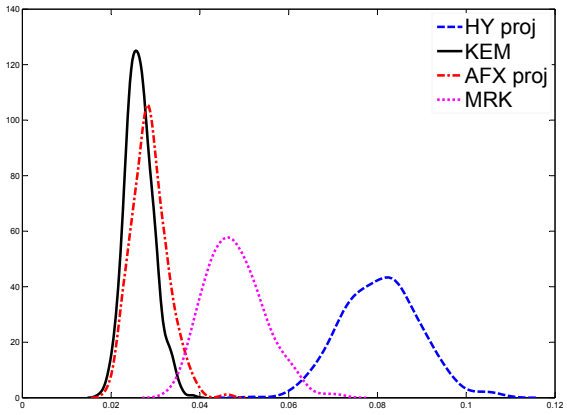

Figure 4: Kernel density estimates of the Frobenius distances: high noise setting. $M = 500$ simulated values of $Frob_k = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d}(\hat{\sigma}_{ij}^k - \sigma_{ij})^2}$, where $\hat{\sigma}_{ij}^k$ is the ij-th element of Q estimated with methodology k, $k \in \{KEM, HY, AFX, MRK\}$ and σ_{ij} is the ij-th element of the true Q used in the simulation. The competing methodologies are the new 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.

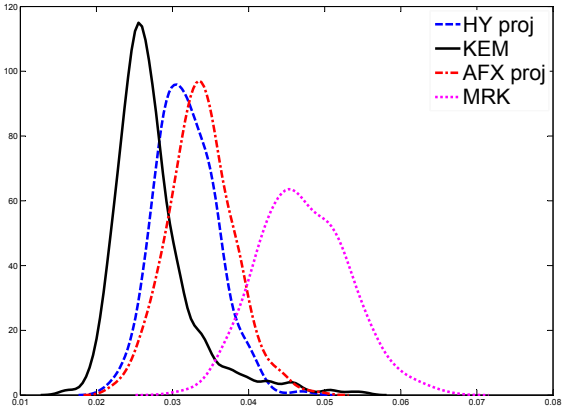

Figure 5: Kernel density estimates of the Frobenius distances: high missings setting. $M = 500$ simulated values of $Frob_k = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d}(\hat{\sigma}_{ij}^k - \sigma_{ij})^2}$, where $\hat{\sigma}_{ij}^k$ is the ij-th element of Q estimated with methodology k, $k \in \{KEM, HY, AFX, MRK\}$ and σ_{ij} is the ij-th element of the true Q used in the simulation. The competing methodologies are the new 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.

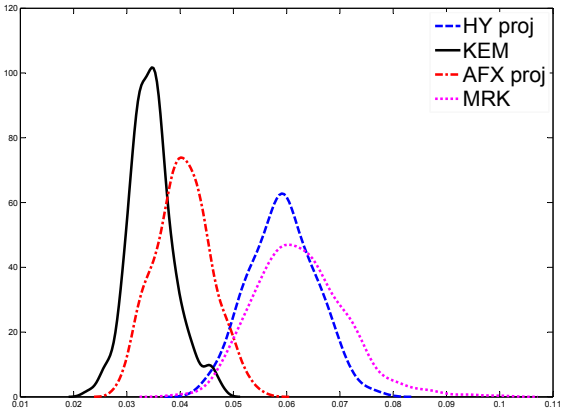

Figure 6: Kernel density estimates of the Frobenius distances: high missings, high noise setting. $M = 500$ simulated values of $Frob_k = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d}(\hat{\sigma}_{ij}^k - \sigma_{ij})^2}$, where $\hat{\sigma}_{ij}^k$ is the ij-th element of Q estimated with methodology k, $k \in \{KEM, HY, AFX, MRK\}$ and σ_{ij} is the ij-th element of the true Q used in the simulation. The competing methodologies are the new 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.

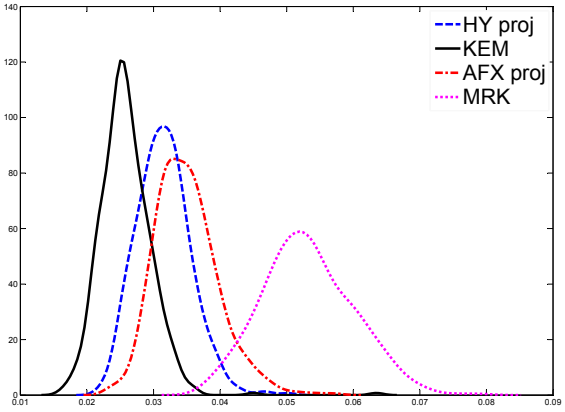

Figure 7: Kernel density estimates of the Frobenius distances: dispersed missings setting. $M = 500$ simulated values of $Frob_k = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d}(\hat{\sigma}_{ij}^k - \sigma_{ij})^2}$, where $\hat{\sigma}_{ij}^k$ is the ij-th element of Q estimated with methodology k, $k \in \{KEM, HY, AFX, MRK\}$ and σ_{ij} is the ij-th element of the true Q used in the simulation. The competing methodologies are the new 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.

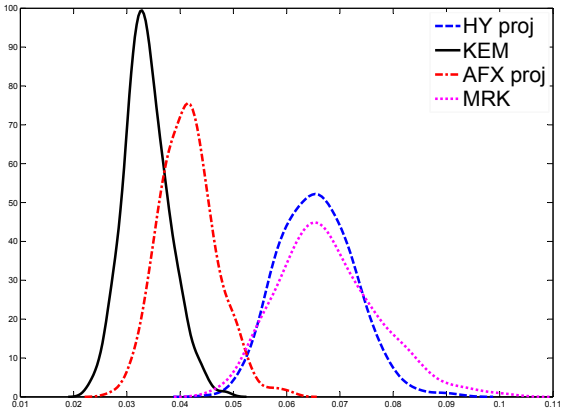

Figure 8: Kernel density estimates of the Frobenius distances: dispersed missings, high noise setting. $M = 500$ simulated values of $Frob_k = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d}(\hat{\sigma}_{ij}^k - \sigma_{ij})^2}$, where $\hat{\sigma}_{ij}^k$ is the ij-th element of Q estimated with methodology k, $k \in \{KEM, HY, AFX, MRK\}$ and σ_{ij} is the ij-th element of the true Q used in the simulation. The competing methodologies are the new 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.

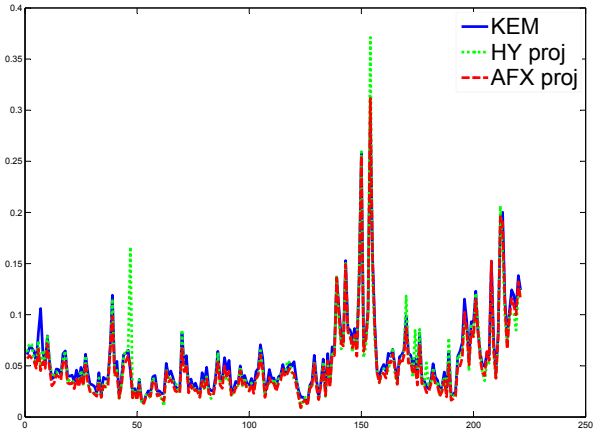

Figure 9: Intel Corporation (INTC) estimated variances for the time period between January 3, 2007 and November 21, 2007. KEM, projected AFX, and projected HY denote the new proposed KEM approach, the projected pairwise estimator proposed by Ait-Sahalia et al. (2010), and a projected covariance matrix 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), respectively.

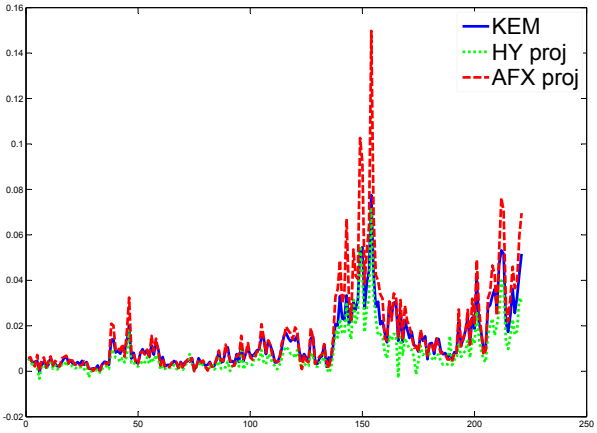

Figure 10: Estimated covariances between Alcoa Inc. (AA) and Exxon Mobil Corporation (XOM) for the time period between January 3, 2007 and November 21, 2007. KEM, projected AFX, and projected HY denote the new proposed KEM approach, the projected pairwise estimator proposed by Ait-Sahalia et al. (2010), and a 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), respectively.