A Markov Chain Estimator of Multivariate Volatility from High Frequency Data

Peter Reinhard Hansen, Guillaume Horel, Asger Lunde and Ilya Archakov

Abstract We introduce a multivariate estimator of financial volatility that is based on the theory of Markov chains. The Markov chain framework takes advantage of the discreteness of high-frequency returns. We study the finite sample properties of the estimation in a simulation study and apply it to high-frequency commodity prices.

Keywords Markov chain · Multivariate volatility · Quadratic variation · Integrated variance · Realized variance · High frequency data

JEL Classification C10 · C22 · C80

The first and third authors acknowledge support from CREATES—Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation.

P.R. Hansen (✉)
European University Institute and CREATES,
Villa San Paolo, Via Della Piazzuola 43, 50133 Firenze, Italy
e-mail: reinhardhansen@gmail.com

G. Horel
Serenitas Capital,
175 Varick St, New York 10014, USA
e-mail: guillaume.horel@gmail.com

A. Lunde
University of Aarhus and CREATES,
Fuglesangs All 4, 8210 Aarhus V, Denmark
e-mail: alunde@econ.au.dk

I. Archakov
European University Institute,
Villa San Paolo, Via Della Piazzuola 43, 50133 Firenze, Italy
e-mail: ilya.archakov@eui.eu

M. Podolskij et al. (eds.), *The Fascination of Probability, Statistics and their Applications*, DOI 10.1007/978-3-319-25826-3_17

1 Introduction

This paper introduces the Markov chain estimator of multivariate volatility. Our analysis builds on the results by [19] who proposed the univariate Markov chain estimator. The multivariate extension poses new challenges related to asynchronicity and the potential need to enforce the estimator to be positive semidefinite.

The availability of high-frequency financial data has made it possible to estimate volatility over relatively short periods of time, such as an hour or a day. The main obstacle in obtaining precise estimators is the fact that high-frequency returns do not conform with conventional no-arbitrage models. The reason is that there is a great deal of autocorrelation in tick-by-tick returns. The apparent contradiction can be explained by market microstructure noise, which gives rise to the notion that the observed price is a noisy measure of the efficient price. In this paper, we introduce a multivariate volatility estimator that is built on the theory of Markov chains. The estimator utilizes the discreteness of high-frequency data, and the framework implicitly permits a high degree of serial dependence in the noise as well as dependence between the efficient price and the noise.

The use of high-frequency data for volatility estimation has been very active over the past two decades, since [3] used the realized variance to evaluate GARCH models. The realized variance is simply the sum of squared intraday returns, and its properties were detailed in [11], for the case where the semimartingale is observed without noise, which was extended to the multivariate context in [12]. The noise in high-frequency returns motivated a number of robust estimators, including the two-scale estimator by [33], the realized kernels by [8], and the pre-average estimator by [25]. Empirical features of the market microstructure noise were detailed in [22], which documented that the noise is both serially dependent and endogenous, in the sense that there is dependence between the underlying semimartingale and the noise. These empirical features motivated the development of the multivariate realized kernel in [9], which is an estimator that permits the noise to have both of these features.

An attractive feature of the Markov framework is that serially dependent and endogenous noise is a natural part of the framework. Moreover, the Markov chain estimator is simple to compute and the same is the case for the estimator of its asymptotic variance. It only takes basic matrix operations to compute the estimator and its confidence intervals.

To illustrate our estimator consider the case with two assets. The bivariate sequence of high-frequency returns is denoted by $\{\varDelta X_t\}_{i=1}^n$, and we define the $S \times 2$ matrix $\mathbf{x}$, where S is the number of states for $\varDelta X_t$, and each row of $\mathbf{x}$ corresponds to a possible realization of $\varDelta X_t$. For instance, the s-th row of $\mathbf{x}$ may equal $x_{s,\cdot} = (2, -1)$ that is the state where the first asset increased by 2 units, while the second asset went down by one unit. The $S \times S$ transition matrix, P, for a Markov chain of order $k = 1$ is given by

$$P_{r,s} = \Pr\left(\varDelta X_{t+1} = x_{s,\cdot} | \varDelta X_t = x_{r,\cdot}\right), \qquad r, s = 1, \ldots, S,$$

and its stationary distribution, $\pi = (\pi_1, \ldots, \pi_S)'$, is characterized by $\pi' P = \pi'$. We define $\Lambda_\pi = \operatorname{diag}(\pi_1, \ldots, \pi_S)$ and the *fundamental matrix* $Z = (I - P + \Pi)^{-1}$ where $\Pi = \iota\pi'$ with $\iota = (1, \ldots, 1)' \in \mathbb{R}^S$. From the maximum likelihood estimator of P we deduced estimates of π and Z, denoted $\hat{\pi}$ and $\hat{Z}$, see Sect. 3 for details. The multivariate Markov chain estimator is given by

$$\mathrm{MC} = nD^{-1}\left\{\mathbf{x}'(\Lambda_{\hat{\pi}}\hat{Z} + \hat{Z}'\Lambda_{\hat{\pi}} - \hat{\pi}\hat{\pi}' - \Lambda_{\hat{\pi}})\mathbf{x}\right\}D^{-1},$$

where $D = \operatorname{diag}(\delta_1, \delta_2)$ and $\delta_j^2 = n^{-1}\sum_{t=1}^n X_{j,t}^2$ is the sample average of the squared price of the j-th asset, $j = 1, 2$. The expression inside the curly brackets is the estimator of the long-run variance of a finite Markov chain, see [20]. The scaling involving D, is a transformation needed for the estimator to be an estimator of volatility of logarithmic prices. The scaling with the sample size, n, relates to the local-to-zero asymptotic scheme that arises under in-fill asymptotics.

Hansen and Horel [19] showed that filtering can resolve the problems caused by market microstructure noise under weak assumptions that essentially amounts to the noise process to be ergodic with finite first moment. This result is theoretical in nature, because the ideal filter requires knowledge about the data generating process. In order to turn the theoretical filtering result into an actual estimator, one needs to adopt a statistical model, and our approach is to model the increments of the process with a Markov chain model, which is a natural starting point given the discrete nature of high-frequency data.

The discreteness of financial data is a product of the so-called *tick size*, which defines the coarseness of the grid that prices are confined to. For example, the tick-size is currently 1 cent for most of the stocks that are listed on the New York Stock Exchange. The implication is that all transaction and quoted prices are in whole cents. The Markov estimator can also be applied to time series that do not live on a grid, by forcing the process onto a grid. While this will introduce rounding error, it will not affect the long-run variance of the process. Delattre and Jacod [16] studied the effect of rounding on realized variances for a standard Brownian motion, and [28] extended this analysis to log-normal diffusions.

The present paper adds to a growing literature on volatility estimation using high-frequency data, dating back to [34, 35]. Well known estimators include the realized variance, see [5, 8]; the two-scale and multi-scale estimators, see [32, 33]; the realized kernels, see [8, 9]. The finite sample properties of these estimators are analyzed in [6, 7], and the close relation between multi-scale estimators and realized kernels is established in [10]. Other estimators include those based on moving average filtered returns, see [4, 21, 30]; the range-based estimator, see [14]; the pre-averaging estimator, see [25]; the quantile-based estimator [13]; and the duration-based estimator, see [2].

The stochastic properties of market microstructure noise are very important in this context. Estimators that are robust to iid noise can be adversely affected by dependent noise. Hansen and Lunde [22] analyzed the empirical features of market microstructure noise and showed that serial dependence and endogenous noise are pronounced in high-frequency stock prices. Endogenous noise refers to the dependence between the noise and the efficient price. A major advantage of the Markov chain estimator is that dependent and endogenous noise is permitted in the framework. In fact, dependent and endogenous noise arises naturally in this context, see [18]. Thus estimation and inference are done under a realistic set of assumptions in regard to the noise.

The present paper is an extension of [19] to the multivariate context. This extension posed new challenges that are specific to the multivariate context. For instance, different assets are typically not traded at synchronous times. This non-synchronicity leads to the so-called Epps effect, which manifests itself by a bias towards zero for the realized covariance as the sampling frequency increases. See [31] for a study of the determinants of the Epps effect. Another issue that may arise in the multivariate context is a need for the estimator to be positive semidefinite, which is not guaranteed by all multivariate estimators. The asynchronicity poses few obstacles for the Markov chain estimator, albeit a large order of the Markov chain, or another remedy, may be needed if an illiquid asset is paired with a liquid asset.

The outline of this paper is as follows. The Markov chain framework is presented in Sect. 2, and the estimator in Sect. 3. In Sect. 4 we present two composite estimators that estimate every element of the matrix separately. In Sect. 5 we propose a novel projection methods that may be needed to ensure that the composite estimators are positive semidefinite. The properties of the estimators are evaluated in Sect. 6 with a simulation study, and an empirical application to commodity prices is presented in Sect. 7.

2 The Markov Chain Framework

Let $\{X_t\}$ denote a d-dimensional process, whose returns ΔX_t can take S distinct values in $\mathbb{R}^d$. For notational convenience we take ΔX_t to be a row-vector. The possible states for the k-tuple, $\Delta\mathscr{X}_t = \{\Delta X_{t-k+1}, \ldots, \Delta X_t\}$, are indexed by $s = 1, \ldots, S^k$, where the s-th state corresponds to the case where $\Delta\mathscr{X}_t = \mathbf{x}_s$, which is an $1 \times kd$ vector. See the example below.

We make the following assumption about the increments of the process.

Assumption 1 The increments $\{\Delta X_t\}_{t=1}^n$ are ergodic and distributed as a homogeneous Markov chain of order $k < \infty$, with $S < \infty$ states.

The homogeneity assumption is unlikely to be valid in the context of high-frequency data. Fortunately the assumption is not critical for our results, because by increasing the order, k, of the homogeneous Markov chain that is imposed on the high-frequency returns, the resulting estimator becomes robust to inhomogeneity, see [19]. This feature of the Markov chain estimator is demonstrated in our simulation study in Sect. 6.

The transition matrix, P, is given by

$$P_{r,s} = \Pr(\Delta\mathscr{X}_{t+1} = \mathbf{x}_s | \Delta\mathscr{X}_t = \mathbf{x}_r), \qquad \text{for } r, s = 1, \ldots, S^k,$$

and the corresponding *stationary distribution,* $\pi \in \mathbb{R}^{S^k}$, which is unique under Assumption 1, is defined by $\pi' P = \pi'$. The *fundamental matrix* by [26] is defined by

$$Z = (I - P + \Pi)^{-1},$$

where $\Pi = \iota\pi'$ with $\iota = (1, \dots, 1)' \in \mathbb{R}^{S^k}$ so that each row of Π is simply π'.

The $S^k \times d$ matrix, f, is defined to be the last d columns of $\mathbf{x}$. So f_s is the value that (the latest observation of) ΔX_t has in state s. (Recall that a state represents a realization of k consecutive returns). Finally, we define the diagonal matrix $\Lambda_\pi = \text{diag}(\pi_1, \dots, \pi_{S^k})$.

The following example illustrates the multivariate Markov chain estimation in the case where $d = 2$ and $S = 2$, and $k = 1, 2$.

Example 1 Suppose that we have two assets and that all price changes are up or down by a single unit. If the order of the Markov chain is $k = 1$, then the transition matrix, P, is a 4×4 matrix, and we can define the state matrix as

$$\mathbf{x} = f = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \end{pmatrix}.$$

If, instead, the order is $k = 2$, then we have $S^2 = 16$ states, and consequently P will be an 16×16 matrix and f an 16×2 matrix. For instance, we may order the states as below, so that a row of $\mathbf{x}$ corresponds to a state value for $(\Delta X_{t-1}, \Delta X_t)$ and the corresponding row of f will have just the state value for ΔX_t:

$$\mathbf{x} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 \\ -1 & -1 & -1 & -1 \end{pmatrix} \qquad f = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \\ 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \\ 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \\ 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \end{pmatrix}.$$

Although the transition matrix is a 16 × 16 matrix, it has at most four non-zero elements in each row. The reason is that many transitions are impossible. For, instance if $\Delta\mathscr{X}_t = \{(1,1),(1,1)\}$ then the next state will have to be $\{(1,1),(*,*)\}$, and a transition to, $\{(-1,-1),(1,1)\}$ say, is impossible, and thus have probability zero. So the transition matrix, P, will be increasingly sparse as k increases.

The underlying idea of the Markov chain estimator is a martingale decomposition of

$$X_t = Y_t + \mu_t + U_t,$$

where $\{Y_t, \mathscr{F}_t\}$ is a martingale with increments $\Delta Y_t' = e_{s_t}' Zf - e_{s_{t-1}}' PZf$, $\mu_t = t\mu$ with $\mu = \mathrm{E}(\Delta X_t)$, and U_t is a stationary ergodic bounded process.

The asymptotic scheme that will be used in the present context is the following:

$$f = n^{-1/2}\xi, \qquad \text{with} \quad \xi \in \mathbb{R}^{S^k \times d} \quad \text{fixed.} \tag{1}$$

This local-to-zero asymptotic scheme is similar to those used in [16, 28], and is natural under in-fill asymptotics. In the present context, it guarantees almost sure convergence of the estimator.

Under this scheme, it follows from [20] (and the ergodic theorem) that

Proposition 1 *Suppose that Assumption 1 holds, then under the asymptotic scheme (1), we have*

$$\sum_{t=1}^{n} \Delta Y_t \Delta Y_t' \overset{a.s.}{\to} \xi' Z'(\Lambda_\pi - P'\Lambda_\pi P)Z\xi = \xi'(\Lambda_\pi Z + Z'\Lambda_\pi - \pi\pi' - \Lambda_\pi)\xi,$$

as $n \to \infty$.

Proof By Assumption 1 it follows that $\vartheta_t' = e_{s_t}' Z\xi - e_{s_{t-1}}' PZ\xi$ is an ergodic Markov chain (of order $k+1$) with

$$\mathrm{E}\vartheta_t\vartheta_t' = \xi' Z'(\Lambda_\pi - P'\Lambda_\pi P)Z\xi = \xi'(\Lambda_\pi Z + Z'\Lambda_\pi - \pi\pi' - \Lambda_\pi)\xi,$$

where the first identity follows from [18, Theorem 2] and the second from [20, Lemma 1]. By the ergodic theorem it follows that $\frac{1}{n}\sum_{t=1}^{n}\vartheta_t\vartheta_t' = \sum_{t=1}^{n}\Delta Y_t \Delta Y_t'$ converges to $\xi'(\Lambda_\pi Z + Z'\Lambda_\pi - \pi\pi' - \Lambda_\pi)\xi$ almost surely (and in mean).

An implication of Proposition 1 is that

$$\Sigma^{\#} = \xi'(\Lambda_\pi Z + Z'\Lambda_\pi - \pi\pi' - \Lambda_\pi)\xi,$$

is the quadratic variation of the martingale component. The same quantity is also the long-run variance of $\varDelta X_t'$ in the sense that

$$\operatorname{var}(X_n' - X_0') = \operatorname{var}(\sum_{t=1}^{n} \varDelta X_t') \to \varSigma^{\#}, \qquad \text{as} \quad n \to \infty.$$

There are different ways to construct a multivariate volatility estimator using Markov chain methods, and we shall present three distinct estimators and highlight each of their advantages.

3 The Markov Estimator

Let $\hat{P}$ be the maximum likelihood estimator of P and let $\hat{\pi}$ be its corresponding eigenvector, $\hat{\pi}'\hat{P} = \hat{\pi}'$. Furthermore, let $\hat{\varPi} = \iota\hat{\pi}'$ and $\hat{Z} = (I - \hat{P} + \hat{\varPi})^{-1}$. The expression for the long-run variance of the Markov chain motivates the estimator

$$\mathrm{MC}^{\#} = nf'(\varLambda_{\hat{\pi}}\hat{Z} + \hat{Z}'\varLambda_{\hat{\pi}} - \hat{\pi}\hat{\pi}' - \varLambda_{\hat{\pi}})f,$$

for which we have the following asymptotic distribution.

Proposition 2 *Suppose that Assumption 1 holds, then under the asymptotic scheme (1), we have*

$$n^{1/2}(\mathrm{MC}^{\#} - \varSigma^{\#}) \xrightarrow{d} N(0, \varOmega),$$

where the asymptotic covariance between the (i, j)th and (l, m)th elements is

$$\varOmega_{ij,kl} = \sum_{r,s,v} [V(r)]_{s,v}[\varXi(r, s)]_{i,j}[\varXi(r, v)]_{l,m}, \tag{2}$$

with $V(s) = \frac{1}{\pi_r}(\varLambda_{e_r'P} - P'e_re_r'P)$ and

$$\begin{aligned}\varXi(r, s) = {} & \pi_r\xi'Z'(\varLambda_{z_s} - P'\varLambda_{z_s}P)Z\xi \\ & +\pi_r\xi'(\pi\pi' - 2\pi z_s' - 2z_s\pi')\xi + \xi'[\varLambda_\pi Ze_rz_s' + z_se_r'Z'\varLambda_\pi]\xi,\end{aligned}$$

and where $z_s' = e_s'Z$ is the s-th row of Z.

Proof Follows from [20, Theorem 2] by adapting their expressions (substitute ξ for f and $\xi'\pi$ for μ).

Remark 1 We note that in the univariate case, $\mathrm{MC}^{\#}$ simplifies to $nf'\varLambda_{\hat{\pi}}(2\hat{Z} - \hat{\varPi} - I)f$, which was the estimator proposed by [19].

3.1 Volatility of Logarithmic Prices

The object of interest is, typically, the volatility of log-prices, rather than $\Sigma^{\#}$, which is the volatility of the price process in levels. An exact estimator can be obtained with the Markov framework, by first extracting the Martingale component of X_t, however, for the univariate case [19] show that the following estimator,

$$\mathrm{MC} = \frac{\mathrm{MC}^{\#}}{\frac{1}{n}\sum_{t=1}^{n} X_{T_t}^2},$$

is virtually identical to the realized variance of the filtered logarithmic prices that are deduced from the estimated Markov chain. The resulting approximate estimator of the quadratic variation has several advantages, such as computational simplicity. In the present multivariate context, we adopt the following estimator:

$$\mathrm{MC} = D^{-1}\mathrm{MC}^{\#}D^{-1}, \tag{3}$$

with $D = \mathrm{diag}\,(\delta_1, \ldots, \delta_d)$ and $\delta_j^2 = n^{-1}\sum_{t=1}^{n} X_{j,t}^2$ $j = 1, \ldots, d$. Our simulation in Sect. 6 shows that this approximate estimator is more accurate than other realized measures.

Alternatively one could use the estimator $\mathrm{MC}_{i,j}^{\#}/\frac{1}{n}\sum_{t=1}^{n} X_{i,t}X_{j,t}$, for $i, j = 1, \ldots, d$, but we prefer (3) because positively definiteness of $\mathrm{MC}^{\#}$ is passed onto MC, and in practice $\delta_i\delta_j \simeq \frac{1}{n}\sum_{t=1}^{n} X_{i,t}X_{j,t}$ because the prices do not vary much in relation to their average level over the estimation window, which is typically a trading day.

4 Composite Markov Estimators

The number of possible states increases exponentially with the dimension of the process, d. Consequently, the dimension of P can become unmanageable even with moderate values of S, k, and d. For instance, with $d = 10$ assets, and price changes ranging from -4 to 4 cents, $S = 9$, and a Markov chain of order $k = 2$, the transition matrix would be $(S^d)^k \times (S^d)^k = 9^{20} \times 9^{20}$, which is impractical.

As an alternative, one can construct a composite estimator, that combines lower dimensional Markov estimator, which is in the spirit of [23, 29]. In this section we consider two such estimators. The first is constructed from univariate estimators, using a simple transformation for the estimation of covariances. The second estimator is constructed from bivariate Markov estimators, which has the advantage that standard errors of each element will be readily available. We will make use of these standard errors in the next section.

4.1 The 1-Composite Markov Estimator

In this section we introduce a composite estimator that is based on univariate Markov estimators. The identity

$$\mathrm{cov}(X,Y)=\frac{\mathrm{var}(X+Y)-\mathrm{var}(X-Y)}{4}$$

motivates the estimator

$$\mathrm{MC}^{\#_1}_{i,j}=\tfrac{1}{4}(\mathrm{MC}^{\#}_{X_i+X_j}-\mathrm{MC}^{\#}_{X_i-X_j}),$$

where $\mathrm{MC}^{\#}_{X_i+X_j}$ and $\mathrm{MC}^{\#}_{X_i-X_j}$ are the univariate Markov chain estimator, applied to the time series $X_{i,t}+X_{j,t}$ and $X_{i,t}-X_{j,t}$, respectively. Note that the diagonal terms, $i=j$, simplifies to $\frac{1}{4}\mathrm{MC}^{\#}_{2X_i}=\mathrm{MC}^{\#}_{X_i}$. This approach to polarization-based estimation of the covariance is well known. In the context of high-frequency data it was first used in [24, Sect. 3.6.1] who also explored related identities. More recently it has been used in [1].

The 1-Composite estimator is mapped into estimators of the volatility of log-returns using the same diagonal matrix, D, as in (3), thus $\mathrm{MC}^1=D^{-1}\mathrm{MC}^{\#_1}D^{-1}$.

4.1.1 Pre-Scaling

If one seeks to estimate the covariance of two assets, whose increments are on different grid sizes, it can be advantageous to use differentiated scaling of the assets, specifically

$$\mathrm{cov}(X,Y)=\frac{\mathrm{var}(aX+bY)-\mathrm{var}(aX-bY)}{4ab},$$

where a and b are constants. This can, in some cases, greatly reduce the number of states, which is computationally advantageous.

4.2 The 2-Composite Markov Estimator

In this subsection we introduce a composite estimator that uses bivariate $\mathrm{MC}^{\#}$ estimates. For all pairs of assets we compute the correlation along with an estimate of its asymptotic variance, which will be used in the next section.

We simply estimate the bivariate Markov process $(X_{i,t},X_{j,t})'$, and obtain the estimator of Sect. 3, $\mathrm{MC}^{\#}$, which is a 2×2 matrix. The covariance terms we seek is the lower-left (or upper-right) element

$$\mathrm{MC}^{\#_2}_{i,j} = \begin{cases} \mathrm{MC}^{\#}(X_i) & \text{if} \quad i = j, \\ \mathrm{MC}^{\#}_{1,2}(X_i, X_j) & \text{if} \quad i \neq j, \end{cases}$$

where $\mathrm{MC}^{\#}_{1,2}(X_i, X_j)$ is the upper right element of the 2×2 matrix $\mathrm{MC}^{\#}$, for the bivariate process, (X_i, X_j). In contrast to the covariance estimated with the 1-composite estimator, the standard error of $\mathrm{MC}^{\#}_{1,2}(X_i, X_j)$ is readily available from (2).

Analogous to the other estimators, the 2-composite estimator is mapped into estimators of the volatility of log-returns with $\mathrm{MC}^2 = D^{-1}\mathrm{MC}^{\#_2} D^{-1}$.

4.3 Advantages and Drawbacks of Composite Estimators

The advantages of the composite estimators are threefold.

- Computational: The state space for a univariate series is smaller than that of a multivariate.
- Dimension: Enables the construction of covariance matrices of any dimension, whereas the multivariate approach is limited to relatively low dimensions.
- No need to synchronize the observation times for each of the asset, e.g. by refresh time, see [9].

The drawbacks of the composite estimators include:

- Positive semidefinite estimate is not guaranteed
- Estimate of asymptotic variance is not readily available.

The dimension of the transition matrix (and fundamental matrix) increases rapidly with the dimension of the process d, and at some point it becomes computationally impossible to manipulate the relevant expressions that are needed for the computation of the Markov estimator. In our empirical analysis with $k = 5$, the dimension of P was about 500–1000 for $d = 1$, about 3000–5000 for $d = 2$, and about 8000–10,000 for $d = 3$. The problem with non-psd appears to be relatively rare in practice when d is small. We have only seen one case where a 5×5 estimate was non-psd estimate. The occurrence is more common in higher dimensions. Of the 251 14×14 estimators we obtained for 2013, 14 of them were non-psd.

5 Enforcing Positivity

While $\mathrm{MC}^{\#}$ is a quadratic form that yields a positive semidefinite estimator, there is no reason to expect that the composite estimators, $\mathrm{MC}^{\#_1}$ and $\mathrm{MC}^{\#_2}$, will be positive semidefinite (PSD) in finite samples. This problem is often encountered in estimation of high-dimensional variance-covariance matrices.

One can project a non-PSD estimate, by solving the following semi-definite program for the variable Σ

$$\min_{\Sigma} \|\Sigma - A\|_{\mathrm{Fro}} \quad \text{subject to} \quad \Sigma \geq 0. \tag{4}$$

The solution can be found efficiently by computing the spectral decomposition of the matrix A, and drop all negative eigenvalues, i.e. map the symmetric matrix, $A = Q\mathrm{diag}(\lambda_1, \ldots, \lambda_d)Q'$ into $Q\mathrm{diag}(\lambda_1^+, \ldots, \lambda_d^+)Q'$, where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of A and $x^+ = \max(x, 0)$. Such an estimator will, due to the zero eigenvalues, be on the boundary of the space of psd matrices, which motivated [27] to impose an additional constraint, $\mathrm{diag}(\Sigma) = \mathrm{diag}(A)$.

In this paper we propose a novel projection that takes advantage of standard errors of the individual elements of the matrix A when these are available. Thus let ω_{ij} be (an estimate of) the standard errors of A_{ij}. Then we solve the following program

$$\min_{\Sigma} \sum_{i,j=1}^{d} \left(\frac{\Sigma_{ij} - A_{ij}}{\omega_{ij}} \right)^2 \quad \text{subject to} \quad \Sigma \geq 0. \tag{5}$$

The solution can be obtained using semidefinite programming solvers that are readily available, including the `cvx` software for Matlab by [17]. The optimization problem can be supplemented with the constraint $\mathrm{diag}(\Sigma) = \mathrm{diag}(A)$, which would produce a constrained estimate with strictly positive eigenvalues, except in pathological cases, e.g. if A is psd with zero eigenvalues to begin with.

The projection in (5) is appealing because it attempts to influence accurately measured elements of A less than those that are relatively inaccurate. An even more appealing projection along these lines would also account for correlations across elements. In the present context, such cross correlations are only available for the estimator MC#. However, since this estimator, MC#, is psd per construction, there is no need for a projection of this estimator.

6 Simulation

In this section we compare the 1-composite Markov estimator against some benchmark. Diagonal elements are compared with the realized variance (RV) and the realized kernel (RK). Off-diagonal elements are compared with the realized covariance (RC).

6.1 Efficient Price

Our simulations are based on two designs for the latent price process, Y_t. In the first design, Y_t is simply sampled from a Brownian motion with constant volatility. In the second design, Y_t is drawn from a stochastic volatility model, which is known as the Dothan model in the literature on interest rates, similar to that used in [8]. Specifically we simulate

$$\log Y_{i,t} = \log Y_{i,t} + \sigma_{i,t} V_{i,t}, \qquad i = 1, 2,$$

where $V_{i,t} = \gamma Z_{i,t} + \sqrt{1-\gamma^2} W_{i,t}$ with $(Z_{1,t}, Z_{2,t}, W_{1,t}, W_{2,t})$ being iid Gaussian, all having unit variance and zero correlation, with the exception that $\mathrm{cov}(W_{1,t}, W_{2,t}) = \rho$.

In the design with stochastic volatility, the volatility, $\sigma_{i,t}$, correlates with $Z_{i,t}$, so that γ controls the leverage effect of the volatility on the stock prices. Specifically,

$$\sigma_{i,t} = \sqrt{\Delta}\left\{\exp\left(\beta_0 + \beta_1 \tau_{i,t}\right)\right\},$$

where $\tau_{i,t} = \exp(\alpha\Delta)\tau_{i,t-1} + \sqrt{\frac{\exp(2\alpha\Delta)-1}{2\alpha}} Z_{i,t}$, with $\tau_{i,1}$ drawn from its unconditional distribution, and $\Delta = \frac{1}{N}$ with $N = 23{,}400$. Additional details about the specification is given in the Appendix.

The values of the parameters in both designs are summarized in Table 1.

6.2 Noise

We will use two specifications for the noise. The first is pure rounding noise, so that

$$X_t = \delta[Y_t/\delta],$$

where $[a]$ is the rounding of a to the integers so that the parameter δ controls the coarseness of the rounding error.

The second specification has an additive noise component in addition to the rounding error, specifically

$$X_t = \delta[(Y_t + U_t)/\delta],$$

Table 1 Parameters values for simulating the efficient price process, Y_t

	β_0	β_1	α	ρ	γ
Constant volatility	0	0	–	–0.3	0
Stochastic volatility	–0.3125	0.125	–0.025	–0.3	0.5

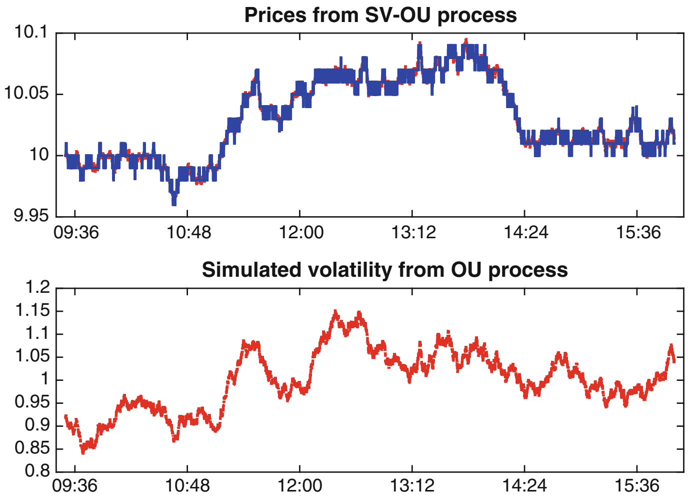


Fig. 1 A typical sample path of the simulated stochastic volatility process. The *upper panel* displays the price process, Y_t, and the observed process, X_t, that is subject to noise and rounding error. The *lower panel* displays the corresponding volatility process, $\sigma^2(t)$

where U_t are iid and uniformly distributed. The idea is that it would more closely resemble the bid ask bounce (due to the additional jitter introduced by U_t, we will either round up or down).

In our simulation study we use $\delta = 0.01$ to emulate rounding errors to a grid, and the noise is $U_t \sim \text{iid}U[-\frac{1}{3}, \frac{1}{3}]$ which adds additional (mean-reverting) jitter to the returns.

In Fig. 1 we show an example of a realization of the process with stochastic volatility using the design in Table 1. The upper panel has Y_t and X_t, where the latter is clearly identified by it being confined to the grid values. The lower panel displays the corresponding volatility process, specifically we plot $\sigma^2(t) = \sigma^2_{i,t}/\Delta$.

6.3 Estimators and Tuning Parameters

We consider the realized variance computed with different sampling frequencies. To imitate a 24 h period with second-by-second price observations, we generate 23,400 noisy high-frequency returns in each simulation.

The realized variance (RV) and the realized covariance (RC) is computed using different sampling frequencies. The choice of sampling frequency entails a bias-variance trade-off, because the bias arising from the noise is most pronounced at high sampling frequencies, while the variance of the estimator increases as the sampling frequency is lowered. Thus for the RV and the RC we sample every H-th price observation where $H \in \{1, 3, 5, 10, 15, 30, 60, 120, 240\}$.

The multivariate realized kernel (MRK) follows the implementation in [9], which is based on the Parzen kernel, and an automatic selection of the bandwidth parameter. This estimator is also applied to high-frequency returns based on the various sample frequencies. The MRK should, in principle, be most accurate when based on returns sampled at the highest frequency, $H = 1$.

The tuning parameter for the Markov chain estimator is the order of the Markov chain, k, and we apply this estimator for $k \in \{1, \dots, 5\}$.

6.4 Simulation Results

We report bias and the root mean squared error (RMSE) for each of the estimators using the various choices for their respective tuning parameters. The results are based on 10,000 simulations. The results are presented in Tables 2 and 3 for the case with constant volatility and stochastic volatility, respectively.

Consider first the case with constant volatility in Table 2. With pure rounding error we note that the Markov chain estimator tend to outperform both the kernel estimator and the realized variance in terms of the mean squared error. Similarly for the covariance, the MC 1-composite estimator dominates the RC and performs on par with the MRK. The Markov estimator is somewhat insensitive to the choice of k, so even with a non-optimal choice for k, the Markov estimator is fairly accurate. The MRK is similarly insensitive to the choice for H. In contrast, the RV and the RC are very sensitive to the choice of H, and suffer from large biases when H is small.

Turning to the case with both additive noise and rounding error. This design generates increments with rather different features. While $k = 1$ was the optimal choice with rounding error, the best configuration is now $k = 3$ or $k = 4$. The RV performs even worse in this design, the RC just as bad as in the previous design, whereas MRK performs as well as in the previous design, and is on par with the Markov estimator. This comparison is again made with hindsight as assume that relatively good choices of tuning parameters, for k and H, respectively, are used. For the covariance, we observe that the RMSE of the Markov estimator is predominantly driven by a bias.

Next we turn to the result in Table 3 which is for the case with stochastic volatility. The Markov chain estimator is based on fitting a homogeneous Markov chain to the observed increment. For this reason it might be expected that the Markov estimator is not well suited for the design with time varying volatility, see Fig. 1. However, even in the case with stochastic volatility that induced an inhomogeneous model for the increments, we see that MC performs well. The RMSEs are, as expected, a bit larger. Interestingly, it is the design with pure rounding errors that results in the largest RMSEs. Both the Markov estimator and the MRK appear to benefit from the additional layer of noise that is added prior to the rounding error.

Table 2 Simulation results for the case with constant volatility

Panel A: Constant volatility and pure rounding error

	Variance				Covariance			
	MC				MC			
k	Bias	Rmse			Bias	Rmse		
1	0.002	0.109			–0.008	0.075		
2	0.004	0.144			–0.053	0.093		
3	0.007	0.174			–0.030	0.098		
4	0.005	0.202			–0.020	0.107		
5	0.003	0.221			–0.010	0.120		
	RV		MRK		RC		MRK	
H	Bias	Rmse	Bias	Rmse	Bias	Rmse	Bias	Rmse
1	3.752	3.763	0.140	0.182	–0.450	0.451	–0.008	0.074
3	3.445	3.455	0.058	0.136	–0.373	0.377	–0.007	0.088
5	3.136	3.145	0.037	0.136	–0.313	0.320	–0.007	0.097
10	2.486	2.494	0.017	0.146	–0.215	0.229	–0.008	0.110
15	2.014	2.021	0.010	0.155	–0.156	0.178	–0.009	0.119
30	1.222	1.229	0.002	0.175	–0.082	0.120	–0.011	0.135
60	0.646	0.657	–0.005	0.200	–0.042	0.100	–0.014	0.154
120	0.324	0.350	–0.015	0.234	–0.020	0.104	–0.018	0.182
240	0.161	0.231	–0.024	0.283	–0.010	0.128	–0.019	0.224

Panel B: Stochastic volatility with noise and rounding error

	Variance				Covariance			
	MC				MC			
k	Bias	Rmse			Bias	Rmse		
1	0.281	0.302			–0.066	0.115		
2	0.140	0.173			–0.137	0.151		
3	0.056	0.131			–0.078	0.104		
4	0.025	0.130			–0.076	0.103		
5	0.011	0.141			–0.049	0.092		
	RV		MRK		RC		MRK	
H	Bias	Rmse	Bias	Rmse	Bias	Rmse	Bias	Rmse
1	9.697	9.708	0.112	0.159	–0.451	0.456	–0.006	0.075
3	8.086	8.095	0.044	0.131	–0.374	0.389	–0.005	0.090
5	6.829	6.838	0.027	0.135	–0.314	0.338	–0.006	0.099
10	4.720	4.727	0.013	0.149	–0.213	0.252	–0.007	0.113
15	3.488	3.495	0.009	0.160	–0.158	0.204	–0.008	0.122
30	1.859	1.866	0.005	0.181	–0.081	0.140	–0.011	0.138
60	0.938	0.949	0.002	0.208	–0.039	0.114	–0.015	0.160
120	0.467	0.491	–0.007	0.243	–0.020	0.115	–0.018	0.188
240	0.233	0.293	–0.016	0.291	–0.009	0.137	–0.021	0.230

[a] Panel A has simulation results for the case where the underlying volatility is constant and the observed process is only subject to rounding error. Panel B presents the corresponding results for the case with both noise and rounding error

Table 3 Simulation results for the case with stochastic volatility

Panel A: Stochastic volatility and pure rounding error

	Variance				Covariance			
	MC				MC			
k	Bias	Rmse			Bias	Rmse		
1	0.005	0.125			–0.020	0.091		
2	0.003	0.165			–0.024	0.084		
3	0.003	0.214			–0.019	0.089		
4	0.002	0.260			–0.010	0.098		
5	–0.001	0.327			–0.007	0.108		
	RV		MRK		RC		MRK	
H	Bias	Rmse	Bias	Rmse	Bias	Rmse	Bias	Rmse
1	3.070	3.325	0.132	0.196	–0.325	0.443	–0.007	0.072
3	2.812	3.032	0.058	0.168	–0.270	0.370	–0.006	0.086
5	2.555	2.741	0.037	0.180	–0.227	0.313	–0.007	0.094
10	2.017	2.141	0.018	0.198	–0.155	0.221	–0.008	0.105
15	1.636	1.721	0.011	0.222	–0.114	0.173	–0.009	0.113
30	1.016	1.056	0.000	0.249	–0.059	0.112	–0.011	0.129
60	0.571	0.622	–0.010	0.295	–0.031	0.098	–0.014	0.151
120	0.304	0.371	–0.022	0.374	–0.016	0.103	–0.016	0.181
240	0.157	0.314	–0.031	0.498	–0.009	0.127	–0.016	0.223

Panel B: Stochastic volatility with noise and rounding error

	Variance				Covariance			
	MC				MC			
k	Bias	Rmse			Bias	Rmse		
1	0.219	0.272			–0.055	0.143		
2	0.117	0.185			–0.093	0.131		
3	0.055	0.175			–0.053	0.097		
4	0.034	0.193			–0.050	0.094		
5	0.021	0.212			–0.034	0.090		
	RV		MRK		RC		MRK	
H	Bias	Rmse	Bias	Rmse	Bias	Rmse	Bias	Rmse
1	9.619	9.639	0.103	0.202	–0.333	0.461	–0.007	0.073
3	8.010	8.028	0.042	0.170	–0.275	0.392	–0.005	0.087
5	6.757	6.773	0.026	0.177	–0.230	0.340	–0.006	0.095
10	4.656	4.668	0.013	0.197	–0.160	0.256	–0.007	0.108
15	3.432	3.443	0.008	0.214	–0.116	0.205	–0.007	0.117
30	1.825	1.836	0.002	0.253	–0.062	0.142	–0.009	0.134
60	0.924	0.942	–0.004	0.311	–0.032	0.116	–0.011	0.157
120	0.462	0.508	–0.013	0.389	–0.017	0.113	–0.014	0.185
240	0.232	0.357	–0.023	0.488	–0.009	0.132	–0.016	0.226

[a] Panel A has simulation results for the case where the underlying volatility is stochastic and the observed process is only subject to rounding error. Panel B presents the corresponding results for the case with both noise and rounding error

7 Empirical Analysis

7.1 *Data Description*

We apply the Markov chain estimator to high-frequency commodity prices, that have previously been used in [15]. We confine our empirical analysis to 2013 data and consider in our study high frequency data for 14 assets. The 14 assets include the exchange traded fund, SPY, that tracks the S&P 500 index, and 13 commodity futures. We refer to [15] for detailed information about the data, including the procedures used for cleaning the high frequency data for outliers and other anomalies. Summary statistics for the 14 assets are presented in (Table 4).

Of the 15 commodities analyzed in [15], we drop two of these series for computational reasons. Specifically, we dropped "Heating Oil" (HO) because it has an unusually large number of distinct second-to-second price increments and "Feeder Cattle" (FC) because it is substantially less liquid compared with the other commodities. Thus, in addition to the SPY, we use the following 13 commodities in our empirical analysis: Crude Light (CL), Natural Gas (NG), Gold (GC), Silver (SV), Copper (HG), Live Cattle (LC), Lean Hogs (LH), Coffee (KC), Sugar (SB), Cotton (CT), Corn (CN), Soybeans (SY) and Wheat (WC).

We exclusively apply the Markov estimators to high-frequency data from the time interval 10:00–14:00 eastern standard time, because all assets are actively traded in this period. The high-frequency prices for eight selected assets for March 18th, 2013 are displayed in Fig. 2.

7.2 *Empirical Results*

First we present detailed results for March 18th, 2013 (to celebrate the occasion for writing this paper). Daily estimates (for the 10:00–14:00 interval) for all trading days in 2013 will be presented in figures.

7.2.1 Daily Estimates for March 18, 2013

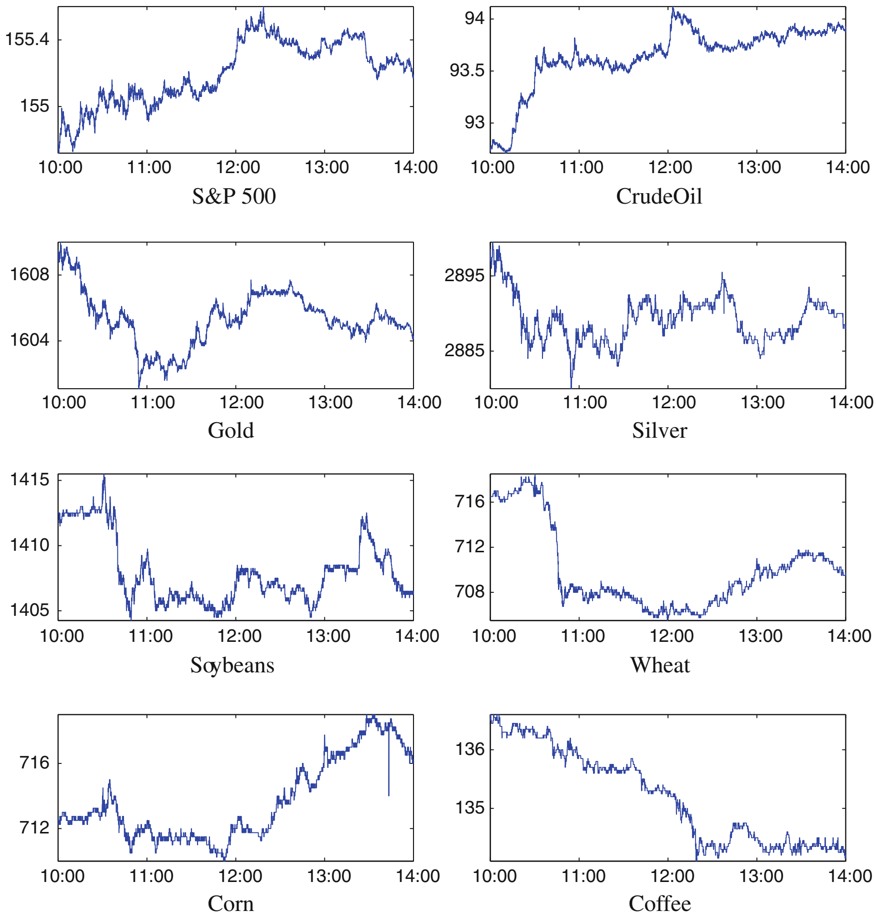
In Table 5 we present five estimators of the volatility matrix for five assets. There are relatively large discrepancies between the two realized variances, which may be due to sampling error or market microstructure noise. The Markov estimators are largely in agreement about the correlations, but the full estimator yields a smaller estimate of the diagonal elements in some cases. This may be caused by the estimator being somewhat unreliable, as it is based on $n = 8{,}700$ observations and the underlying transition matrix is an $8{,}600 \times 8{,}600$ matrix in this case. Further research is needed to characterize the limitations of the full estimator in practice.

Table 4 Summary statistics for the 251 trading days in 2013

Data summary statistics

	SPY	CL	NG	GC	SV	HG	LC	LH	KC	SB	CT	CN	SY	WC
Transactions per day (full trading day)	19178	21993	9761	21671	9607	9738	2835	2775	2922	4013	3325	7058	8882	4684
Transactions per day: n (10:00 AM–2:00 PM)	10117	8634	4495	6320	3024	2706	2315	2213	1613	2125	1940	4362	5011	2895
Primitive states:(10:00 AM–2:00 PM)	11	19	11	17	11	8	8	8	11	7	23	11	13	9
Tick size	0.01	0.01	0.001	0.1	0.5	0.05	0.025	0.025	0.05	0.01	0.01	0.25	0.25	0.25
Volatility share: κ (10 AM–2 PM)/24 h	0.31	0.36	0.40	0.24	0.20	0.20	0.47	0.25	0.44	0.37	0.46	0.24	0.25	0.52
Annual volatility (2013)	11 %	19 %	31 %	20 %	32 %	19 %	12 %	24 %	27 %	18 %	21 %	33 %	25 %	22 %

The average number of price observations per day, and within the 10:00 AM to 2:00 PM window, n, are reported in the first two rows, followed by the average number of primitive states, S, and the tick size for each of the assets. The second last row reports κ—the fraction of daily volatility that occurs during the 10:00 AM to 2:00 PM window on average. The average volatility for 2013 is reported in the last row



S&P 500

CrudeOil

Gold

Silver

Soybeans

Wheat

Corn

Coffee

Fig. 2 High frequency prices for eight selected commodities on March 18th, 2013 during the period from 10:00 AM to 2:00 PM

In Tables 6 and 7 we present estimates of the full 14 × 14 matrix. The realized variances are in Table 6 and the two composite Markov estimators are in Table 7. Joint estimation of the full 14 × 14 covariance matrix is not expected to be precise because the number of observed states (and the dimensionality of the transition matrix) is equal to the number of observations in that case. As in the previous example, the two realized variance estimators produce quite different values whereas the composite Markov estimators produce rather similar results. In general, signs and magnitudes of the elements of Markov covariance matrices are largely in agreement with those of the realized variances.

Table 5 Two realized variances and the three variants of the Markov chain estimator are presented

RV_{5min}: *Realized variance with 5-min sampling*

	SPY	CL	GC	SV	KC
S&P 500	**81.16**	79.39	–8.26	–20.03	–15.24
Light crude	*0.48*	**344.08**	–36.52	–1.39	–2.29
Gold	*–0.08*	*–0.16*	**146.11**	164.45	–47.74
Silver	*–0.13*	*–0.00*	*0.77*	**316.22**	–36.77
Coffee "C"	*–0.08*	*–0.01*	*–0.20*	*–0.10*	**407.69**

RV_{10min}: *Realized variance with 10-min sampling*

	SPY	CL	GC	SV	KC
S&P 500	**68.92**	90.57	1.98	–20.04	–40.79
Light crude	*0.59*	**342.05**	–69.66	–57.07	–11.96
Gold	*0.03*	*–0.43*	**78.43**	50.73	–35.05
Silver	*–0.20*	*–0.26*	*0.48*	**139.93**	–17.48
Coffee "C"	*–0.23*	*–0.03*	*–0.19*	*–0.07*	**446.70**

MC: *Markov chain estimator (full)*

	SPY	CL	GC	SV	KC
S&P 500	**80.65**	88.17	–24.05	–77.56	–1.89
Light crude	*0.49*	**407.24**	–75.01	–124.63	49.36
Gold	*–0.26*	*–0.35*	**109.99**	155.36	–32.78
Silver	*–0.45*	*–0.32*	*0.77*	**370.69**	–12.07
Coffee "C"	*–0.01*	*0.15*	*–0.19*	*–0.04*	**257.15**

MC^1: *Markov chain estimator 1-composite*

	SPY	CL	GC	SV	KC
S&P 500	**116.84**	75.77	–10.69	4.58	21.01
Light crude	*0.35*	**391.86**	14.34	61.58	41.60
Gold	*–0.09*	*0.07*	**116.87**	151.25	–16.83
Silver	*0.02*	*0.16*	*0.72*	**380.17**	–0.71
Coffee "C"	*0.09*	*0.10*	*–0.08*	*–0.00*	**421.94**

MC^2: *Markov chain 2-composite*

	SPY	CL	GC	SV	KC
S&P 500	**116.84**	80.35	–12.92	–7.44	13.76
Light Crude	*0.38*	**391.86**	–1.45	39.30	29.37
Gold	*–0.11*	*–0.01*	**116.87**	149.72	–27.00
Silver	*–0.04*	*0.10*	*0.71*	**380.17**	–59.90
Coffee "C"	*0.06*	*0.07*	*–0.12*	*–0.15*	**421.94**

The estimators are for the 10:00 AM–2:00 PM period on March 18, 2013 for five selected assets. Variances and covariances are annualized and further scaled by 10^4. Correlations are in the lower triangle in italic font

Table 6 The table presents two realized variances for the 14 assets

$RV_{5\text{min}}$	SPY	CL	NG	GC	SV	HG	LC	LH	KC	SB	CT	CN	SY	WC
S&P 500	**81.16**	79.39	18.76	–8.26	–20.03	44.26	–23.18	–1.03	–15.24	23.70	46.58	12.24	16.52	–21.65
Light crude	*0.48*	**344.08**	80.92	–36.52	–1.39	73.65	–7.29	–69.62	–2.29	–20.68	11.62	5.80	32.45	23.80
Natural gas	*0.07*	*0.15*	**805.41**	–4.63	60.09	159.46	–94.66	–21.79	53.54	–50.70	–54.01	–85.28	22.49	37.54
Gold	*-0.08*	*–0.16*	*–0.01*	**146.11**	164.45	6.66	–8.35	–4.19	–47.74	–39.13	11.51	–43.91	–21.55	–32.41
Silver	*–0.13*	*–0.00*	*0.12*	*0.77*	**316.22**	59.49	–21.80	9.80	–36.77	–51.15	16.38	–61.21	–46.76	–22.01
Copper	*0.30*	*0.24*	*0.34*	*0.03*	*0.20*	**269.51**	–70.43	13.29	89.35	13.11	24.30	–53.07	68.05	8.94
Live cattle	*–0.19*	*–0.03*	*–0.25*	*–0.05*	*–0.09*	*–0.32*	**180.30**	18.30	–23.07	–16.29	–26.96	70.81	–6.47	80.24
Lean hogs	*–0.00*	*–0.15*	*–0.03*	*–0.01*	*0.02*	*0.03*	*0.05*	**618.27**	–54.91	–7.75	9.10	–44.27	–88.09	–1.78
Coffee “C”	*–0.08*	*–0.01*	*0.09*	*–0.20*	*–0.10*	*0.27*	*–0.09*	*–0.11*	**407.69**	34.62	12.53	123.43	62.26	180.13
Sugar #1	*0.16*	*–0.07*	*–0.11*	*–0.20*	*–0.17*	*0.05*	*–0.07*	*–0.02*	*0.10*	**274.87**	65.41	–29.63	28.12	46.86
Cotton #2	*0.24*	*0.03*	*–0.09*	*0.04*	*0.04*	*0.07*	*–0.09*	*0.02*	*0.03*	*0.18*	**478.52**	87.19	102.05	53.99
Corn	*0.05*	*0.01*	*–0.12*	*–0.14*	*–0.14*	*–0.13*	*0.21*	*–0.07*	*0.24*	*–0.07*	*0.16*	**640.54**	288.11	412.80
Soybeans	*0.09*	*0.09*	*0.04*	*–0.09*	*–0.13*	*0.20*	*–0.02*	*–0.17*	*0.15*	*0.08*	*0.23*	*0.56*	**415.85**	232.42
Wheat	*–0.10*	*0.05*	*0.06*	*–0.11*	*–0.05*	*0.02*	*0.25*	*–0.00*	*0.38*	*0.12*	*0.10*	*0.69*	*0.48*	**553.00**

(continued)

Table 6 (continued)

RV_{10min}	SPY	CL	NG	GC	SV	HG	LC	LH	KC	SB	CT	CN	SY	WC
S&P 500	**68.92**	90.57	–7.81	1.98	–20.04	26.06	3.25	24.47	–40.79	–5.38	35.40	–3.91	–24.18	–32.89
Light crude	*0.59*	**342.05**	98.06	–69.66	–57.07	90.71	9.94	21.31	–11.96	–29.12	88.15	–8.19	20.25	18.49
Natural gas	*–0.04*	*0.22*	**578.93**	45.71	101.82	160.88	–78.18	–104.34	104.28	–42.88	47.80	–36.26	66.09	51.45
Gold	*0.03*	*–0.43*	*0.21*	**78.43**	50.73	9.61	–17.97	–8.42	–35.05	–2.83	–28.70	–38.85	–18.87	–32.80
Silver	*–0.20*	*–0.26*	*0.36*	*0.48*	**139.93**	54.77	–23.71	–15.59	–17.48	–13.00	–33.72	–39.72	–6.86	–23.42
Copper	*0.18*	*0.28*	*0.38*	*0.06*	*0.26*	**317.27**	–102.54	114.21	123.27	35.28	–37.42	–78.76	48.24	49.66
Live cattle	*0.03*	*0.04*	*–0.25*	*–0.16*	*–0.16*	*–0.45*	**164.68**	30.43	–25.38	2.54	88.65	151.23	65.74	148.58
Lean hogs	*0.14*	*0.06*	*–0.21*	*–0.05*	*–0.06*	*0.31*	*0.12*	**418.34**	–94.24	41.35	31.23	–113.20	–66.97	4.08
Coffee “C”	*–0.23*	*–0.03*	*0.21*	*–0.19*	*–0.07*	*0.33*	*–0.09*	*–0.22*	**446.70**	60.91	–62.30	108.18	100.85	279.64
Sugar #1	*–0.04*	*0.10*	*0.11*	*0.02*	*0.07*	*0.13*	*0.01*	*0.13*	*0.19*	**242.43**	145.50	26.27	–11.14	108.60
Cotton #2	*0.19*	*0.21*	*0.09*	*–0.14*	*–0.13*	*–0.09*	*0.31*	*0.07*	*–0.13*	*0.42*	**503.18**	54.13	73.48	147.24
Corn	*–0.02*	*–0.02*	*–0.06*	*–0.17*	*–0.13*	*–0.17*	*0.45*	*–0.21*	*0.19*	*–0.06*	*0.09*	**699.55**	403.33	542.96
Soybeans	*–0.14*	*0.05*	*0.13*	*–0.10*	*–0.03*	*0.13*	*0.24*	*–0.15*	*0.22*	*–0.03*	*0.15*	*0.71*	**463.78**	452.48
Wheat	*–0.14*	*0.03*	*0.07*	*–0.13*	*–0.07*	*0.10*	*0.39*	*0.01*	*0.45*	*0.24*	*0.22*	*0.70*	*0.72*	**860.64**

Variances and covariances are annualized and further scaled by 10^4. Correlations are in the lower triangle in italic font

Table 7 The table presents two Markov estimates of the variance-covariance for the 14 assets

MC[1]	SPY	CL	NG	GC	SV	HG	LC	LH	KC	SB	CT	CN	SY	WC
S&P 500	**116.84**	75.77	–34.48	–10.69	4.58	73.32	2.39	–5.64	21.01	–6.10	2.00	–4.96	2.38	1.81
Light crude	*0.35*	**391.86**	–23.61	14.34	61.58	101.52	11.42	–31.64	41.60	5.07	11.34	14.14	5.57	3.12
Natural gas	*–0.11*	*–0.04*	**921.74**	–1.88	–14.76	30.77	–26.48	8.46	29.60	–10.02	–61.60	52.74	–6.90	–1.04
Gold	*–0.09*	*0.07*	*–0.01*	**116.87**	151.25	9.18	–4.74	7.84	–16.83	10.41	–36.42	28.66	5.55	19.28
Silver	*0.02*	*0.16*	*–0.02*	*0.72*	**380.17**	31.04	30.64	30.20	–0.71	–3.73	85.57	–1.61	5.51	–17.91
Copper	*0.35*	*0.27*	*0.05*	*0.04*	*0.08*	**365.15**	–8.35	–21.68	–4.71	–36.98	13.93	–5.18	–23.51	–7.07
Live cattle	*0.02*	*0.05*	*–0.07*	*–0.04*	*0.13*	*–0.04*	**142.36**	40.37	4.44	8.11	–1.72	13.95	–18.30	27.09
Lean hogs	*–0.02*	*–0.06*	*0.01*	*0.03*	*0.06*	*–0.05*	*0.14*	**606.71**	4.55	–2.74	–22.64	–0.43	6.19	–19.86
Coffee "C"	*0.09*	*0.10*	*0.05*	*–0.08*	*–0.00*	*–0.01*	*0.02*	*0.01*	**421.94**	10.27	91.99	–1.41	39.64	–39.64
Sugar #1	*–0.03*	*0.01*	*–0.02*	*0.05*	*–0.01*	*–0.10*	*0.04*	*–0.01*	*0.03*	**345.21**	–37.34	3.76	–6.25	32.86
Cotton #2	*0.01*	*0.02*	*–0.09*	*–0.14*	*0.19*	*0.03*	*–0.01*	*–0.04*	*0.19*	*–0.09*	**544.65**	134.15	68.07	73.68
Corn	*–0.02*	*0.03*	*0.08*	*0.12*	*–0.00*	*–0.01*	*0.05*	*–0.00*	*–0.00*	*0.01*	*0.26*	**475.69**	204.56	276.23
Soybeans	*0.01*	*0.01*	*–0.01*	*0.03*	*0.01*	*–0.06*	*–0.08*	*0.01*	*0.10*	*–0.02*	*0.15*	*0.49*	**365.44**	132.19
Wheat	*0.01*	*0.01*	*–0.00*	*0.09*	*–0.05*	*–0.02*	*0.11*	*–0.04*	*–0.10*	*0.09*	*0.16*	*0.63*	*0.34*	**407.53**

(continued)

Table 7 (continued)

MC^2	SPY	CL	NG	GC	SV	HG	LC	LH	KC	SB	CT	CN	SY	WC
S&P 500	**116.84**	80.35	–12.55	–12.92	–7.44	76.88	8.40	–5.69	13.76	1.67	42.68	–9.93	10.89	0.47
Light crude	*0.38*	**391.86**	7.67	–1.45	39.30	125.36	6.38	–11.56	29.37	–7.56	89.63	–25.61	22.06	–3.37
Natural gas	*–0.04*	*0.01*	**921.74**	1.64	–44.28	84.25	–32.41	–10.85	74.22	–53.72	–56.59	34.81	–12.94	–22.93
Gold	*–0.11*	*–0.01*	*0.00*	**116.87**	149.72	15.23	–2.08	25.73	–27.00	–4.72	–37.49	0.15	5.05	–3.15
Silver	*–0.04*	*0.10*	*–0.07*	*0.71*	**380.17**	23.70	4.60	16.03	–59.90	23.58	–60.49	–36.39	–1.79	–38.08
Copper	*0.37*	*0.33*	*0.15*	*0.07*	*0.06*	**365.15**	–29.91	17.98	24.53	0.16	80.92	–7.86	–34.35	–16.89
Live cattle	*0.07*	*0.03*	*–0.09*	*–0.02*	*0.02*	*–0.13*	**142.36**	18.71	–20.08	4.14	–15.27	9.25	–19.75	31.61
Lean hogs	*–0.02*	*–0.02*	*–0.01*	*0.10*	*0.03*	*0.04*	*0.06*	**606.71**	69.63	7.01	426.46	–27.19	19.11	35.20
Coffee “C”	*0.06*	*0.07*	*0.12*	*–0.12*	*–0.15*	*0.06*	*–0.08*	*0.14*	**421.94**	23.83	5.64	–15.59	54.31	–39.85
Sugar #1	*0.01*	*–0.02*	*–0.10*	*–0.02*	*0.07*	*0.00*	*0.02*	*0.02*	*0.06*	**345.21**	95.41	11.69	–28.41	29.00
Cotton #2	*0.17*	*0.19*	*–0.08*	*–0.15*	*–0.13*	*0.18*	*–0.05*	*0.74*	*0.01*	*0.22*	**544.65**	83.64	19.19	68.24
Corn	*–0.04*	*–0.06*	*0.05*	*0.00*	*–0.09*	*–0.02*	*0.04*	*–0.05*	*–0.03*	*0.03*	*0.16*	**475.69**	236.91	320.47
Soybeans	*0.05*	*0.06*	*–0.02*	*0.02*	*–0.00*	*–0.09*	*–0.09*	*0.04*	*0.14*	*–0.08*	*0.04*	*0.57*	**365.44**	162.80
Wheat	*0.00*	*–0.01*	*–0.04*	*–0.01*	*–0.10*	*–0.04*	*0.13*	*0.07*	*–0.10*	*0.08*	*0.14*	*0.73*	*0.42*	**407.53**

Variances and covariances are annualized and further scaled by 10^4. Correlations are in the lower triangle in italic font

7.2.2 Estimates for Pooled March Data

Finally we have pooled the high-frequency data for all of March and estimated the 14 × 14 matrix that reflects the volatility in March, 2013 that occurred during the 10:00 AM–2:00 PM trading periods. With 20 trading days in March, 2013, this adds up to 80 h of high-frequency data. Precision is expected to improve with the larger sample size, although the dimensions of the underlying transition matrices are expected to increase as a larger number of states and transitions will be observed in a larger sample, and the latter can potentially cause computational difficulties. For the 2-composite estimator with $k = 4$ we observed between 15000 and 30000 distinct states in the pooled data set. Another challenges for the Markov estimator in the pooled sample is that a larger degree of inhomogeneity may be expected. Hansen and Horel [19] showed that an inhomogeneous Markov process can be approximated by a homogeneous Markov process, by increasing the order of the Markov chain. So a larger k may be needed in the pooled data, which also poses computational challenges.

In Tables 8 and 9 we report estimates for the 14 × 14 covariance matrix computed with the realized variances and the two composite Markov estimators. In contrast to the data for March 18th, 2013, the realized variances are largely in agreement for the pooled data. Albeit some differences are observed between the 5- and 10-min realized variances. The composite Markov estimators are in disagreement in some cases, which we attributed to the different order of the Markov chain that were used. The 1-composite estimator was computed with $k = 5$ whereas 2-composite was estimated with $k = 3$, for computational reasons. Naturally, one could use a higher order to compute the diagonal elements, but we used the same order for all entries of the 2-composite estimator to illustrate the differences that arise in this case. The 1-composite Markov estimator produces estimates that are generally in agreement with those of the realized variance, both in terms of magnitude and signs of covariances.

7.2.3 Daily Estimates for 2013

We have estimated variances and covariances for the 10:00–14:00 interval for all trading days during 2013. Some selected series are presented in Figs. 3 and 4.

In Fig. 3 we plot annualized volatilities for SPY, Crude Oil, Gold, and Wheat based on the Markov estimator with $k = 5$, and these are benchmarked with those of the realized variance with 10-min sampling. The estimates are quite similar, both for the very liquid assets, SPY, Crude Oil, and Gold, and the relatively illiquid assets, Wheat, whose high-frequency data have pronounced bid-ask bounces.

In Fig. 4 we plot daily estimates of the correlations for both the 1- and 2-composite Markov estimators. These are benchmarked with the realized correlations based on 10-min sampling. We observed that Gold/Silver are highly correlated and its correlation is highly persistent over the year 2013. More moderate correlations are observed for SPY/Crude Oil and Soybeans/Wheat, and these series exhibit a higher

Table 8 The table presents two realized variances for the 14 assets for the month of March, 2013

$\text{RV}_{5\text{min}}$	SPY	CL	NG	GC	SV	HG	LC	LH	KC	SB	CT	CN	SY	WC
S&P 500	**98.20**	72.55	7.53	–29.49	–40.80	59.85	11.82	8.20	14.35	–0.77	11.53	12.48	6.01	14.05
Light crude	*0.42*	**304.64**	41.55	4.72	27.94	106.73	17.78	34.32	46.45	5.08	15.23	45.38	25.09	37.27
Natural gas	*0.03*	*0.08*	**828.53**	–0.90	5.96	43.62	5.33	5.39	3.84	–21.71	4.43	18.29	28.41	–0.53
Gold	*–0.27*	*0.02*	*–0.00*	**123.93**	201.43	21.89	–5.48	–18.43	–10.53	–3.25	–3.96	8.03	14.74	6.89
Silver	*–0.18*	*0.07*	*0.01*	*0.80*	**512.90**	81.49	2.16	–45.85	15.06	13.05	–0.98	4.98	30.76	8.66
Copper	*0.38*	*0.38*	*0.09*	*0.12*	*0.22*	**257.37**	7.16	–10.27	49.36	11.44	–2.78	40.49	38.99	52.05
Live cattle	*0.08*	*0.07*	*0.01*	*–0.03*	*0.01*	*0.03*	**199.14**	153.29	4.09	–6.71	8.24	11.19	10.78	15.68
Lean hogs	*0.03*	*0.06*	*0.01*	*–0.05*	*–0.07*	*–0.02*	*0.36*	**935.02**	9.53	–60.11	20.26	7.31	35.86	–1.77
Coffee “C”	*0.06*	*0.10*	*0.01*	*–0.04*	*0.03*	*0.12*	*0.01*	*0.01*	**657.53**	74.46	16.09	63.27	22.37	61.59
Sugar #1	*–0.00*	*0.01*	*–0.03*	*–0.01*	*0.02*	*0.03*	*–0.02*	*–0.08*	*0.12*	**551.36**	2.53	79.07	42.36	72.83
Cotton #2	*0.05*	*0.04*	*0.01*	*–0.02*	*–0.00*	*–0.01*	*0.03*	*0.03*	*0.03*	*0.00*	**488.85**	–69.86	–17.14	–85.78
Corn	*0.04*	*0.08*	*0.02*	*0.02*	*0.01*	*0.08*	*0.03*	*0.01*	*0.08*	*0.11*	*–0.10*	**996.64**	335.43	711.51
Soybeans	*0.03*	*0.06*	*0.04*	*0.06*	*0.06*	*0.10*	*0.03*	*0.05*	*0.04*	*0.08*	*–0.03*	*0.45*	**552.12**	270.99
Wheat	*0.05*	*0.07*	*–0.00*	*0.02*	*0.01*	*0.11*	*0.04*	*–0.00*	*0.08*	*0.11*	*–0.14*	*0.79*	*0.40*	**818.54**

(continued)

Table 8 (continued)

$\mathrm{RV}_{10\mathrm{min}}$	SPY	CL	NG	GC	SV	HG	LC	LH	KC	SB	CT	CN	SY	WC
S&P 500	**103.38**	83.04	1.75	–35.00	–43.92	67.72	18.68	17.29	13.72	1.82	14.03	11.27	13.61	20.55
Light crude	*0.45*	**329.37**	59.02	3.89	37.04	127.49	17.70	46.91	68.14	12.95	7.74	50.17	29.61	36.87
Natural gas	*0.01*	*0.12*	**771.89**	14.50	45.49	41.19	–2.69	8.37	0.61	–54.02	22.02	–18.33	4.83	–0.14
Gold	*–0.32*	*0.02*	*0.05*	**117.30**	193.10	14.79	–16.56	–28.31	–4.43	7.41	–6.04	2.63	11.57	2.25
Silver	*–0.19*	*0.09*	*0.07*	*0.79*	**510.20**	84.45	–18.71	–84.96	34.49	30.12	–0.59	1.41	21.38	14.52
Copper	*0.42*	*0.44*	*0.09*	*0.09*	*0.23*	**255.99**	16.10	–0.08	58.08	18.61	–18.35	39.15	40.07	63.66
Live cattle	*0.13*	*0.07*	*–0.01*	*–0.11*	*–0.06*	*0.07*	**206.36**	182.41	13.66	–7.21	26.27	21.69	24.43	18.48
Lean hogs	*0.06*	*0.09*	*0.01*	*–0.09*	*–0.13*	*–0.00*	*0.43*	**870.97**	–8.22	–37.37	1.44	38.99	50.73	23.20
Coffee “C”	*0.05*	*0.14*	*0.00*	*–0.02*	*0.06*	*0.14*	*0.04*	*–0.01*	**689.21**	75.36	9.40	96.86	77.15	83.21
Sugar #1	*0.01*	*0.03*	*–0.09*	*0.03*	*0.06*	*0.05*	*–0.02*	*–0.06*	*0.13*	**502.96**	17.81	92.54	76.18	91.50
Cotton #2	*0.07*	*0.02*	*0.04*	*–0.03*	*–0.00*	*–0.05*	*0.09*	*0.00*	*0.02*	*0.04*	**442.55**	–12.93	19.53	0.36
Corn	*0.04*	*0.09*	*–0.02*	*0.01*	*0.00*	*0.08*	*0.05*	*0.04*	*0.12*	*0.13*	*–0.02*	**999.24**	356.24	647.75
Soybeans	*0.05*	*0.06*	*0.01*	*0.04*	*0.04*	*0.10*	*0.07*	*0.07*	*0.11*	*0.13*	*0.04*	*0.44*	**653.10**	265.91
Wheat	*0.08*	*0.08*	*–0.00*	*0.01*	*0.02*	*0.15*	*0.05*	*0.03*	*0.12*	*0.16*	*0.00*	*0.78*	*0.40*	**683.20**

Variances and covariances are annualized and further scaled by 10^4. Correlations are in the lower triangle in italic font

Table 9 The table presents two Markov estimates of the variance-covariance for the 14 assets

MC[1]	SPY	CL	NG	GC	SV	HG	LC	LH	KC	SB	CT	CN	SY	WC
S&P 500	**106.38**	69.59	–1.52	–12.53	–9.17	49.64	5.09	–3.36	10.86	–1.08	5.40	11.11	12.02	8.20
Light crude	*0.37*	**326.26**	14.89	9.34	28.47	82.09	6.53	2.76	16.30	–5.62	5.13	2.92	16.26	14.55
Natural gas	*–0.00*	*0.03*	**932.63**	7.18	18.48	10.08	12.13	–5.79	22.71	–0.61	–9.77	–22.36	14.87	1.44
Gold	*–0.10*	*0.04*	*0.02*	**149.93**	219.11	26.36	–0.83	1.71	1.00	–2.37	–1.74	–0.26	2.41	–0.02
Silver	*–0.04*	*0.07*	*0.03*	*0.76*	**557.24**	73.28	0.12	–5.14	6.68	–13.18	1.91	4.63	5.79	–1.43
Copper	*0.30*	*0.28*	*0.02*	*0.13*	*0.19*	**259.93**	3.97	–6.91	18.82	–1.84	9.30	16.38	9.38	10.22
Live cattle	*0.03*	*0.02*	*0.03*	*–0.00*	*0.00*	*0.02*	**213.33**	128.46	11.34	–1.90	16.85	–2.27	1.36	5.90
Lean hogs	*–0.01*	*0.01*	*–0.01*	*0.00*	*–0.01*	*–0.01*	*0.31*	**830.83**	9.27	–5.11	9.64	3.63	11.97	6.54
Coffee "C"	*0.04*	*0.04*	*0.03*	*0.00*	*0.01*	*0.05*	*0.03*	*0.01*	**636.57**	36.08	7.91	2.97	7.70	8.80
Sugar #1	*–0.00*	*–0.01*	*–0.00*	*–0.01*	*–0.02*	*–0.00*	*–0.01*	*–0.01*	*0.06*	**528.18**	–1.88	33.55	15.23	28.24
Cotton #2	*0.02*	*0.01*	*–0.01*	*–0.01*	*0.00*	*0.03*	*0.05*	*0.02*	*0.01*	*–0.00*	**490.22**	12.67	11.50	13.48
Corn	*0.04*	*0.01*	*–0.03*	*–0.00*	*0.01*	*0.04*	*–0.01*	*0.00*	*0.00*	*0.05*	*0.02*	**835.96**	299.02	523.38
Soybeans	*0.05*	*0.04*	*0.02*	*0.01*	*0.01*	*0.02*	*0.00*	*0.02*	*0.01*	*0.03*	*0.02*	*0.43*	**578.95**	220.34
Wheat	*0.03*	*0.03*	*0.00*	*–0.00*	*–0.00*	*0.02*	*0.02*	*0.01*	*0.01*	*0.05*	*0.02*	*0.70*	*0.35*	**666.82**

(continued)

Table 9 (continued)

MC^2	SPY	CL	NG	GC	SV	HG	LC	LH	KC	SB	CT	CN	SY	WC
S&P 500	**116.84**	80.35	–12.55	–12.92	–7.44	76.88	8.40	–5.69	13.76	1.67	42.68	–9.93	10.89	0.47
Light crude	*0.38*	**391.86**	7.67	–1.45	39.30	125.36	6.38	–11.56	29.37	–7.56	89.63	–25.61	22.06	–3.37
Natural gas	*–0.04*	*0.01*	**921.74**	1.64	–44.28	84.25	–32.41	–10.85	74.22	–53.72	–56.59	34.81	–12.94	–22.93
Gold	*–0.11*	*–0.01*	*0.00*	**116.87**	149.72	15.23	–2.08	25.73	–27.00	–4.72	–37.49	0.15	5.05	–3.15
Silver	*–0.04*	*0.10*	*–0.07*	*0.71*	**380.17**	23.70	4.60	16.03	–59.90	23.58	–60.49	–36.39	–1.79	–38.08
Copper	*0.37*	*0.33*	*0.15*	*0.07*	*0.06*	**365.15**	–29.91	17.98	24.53	0.16	80.92	–7.86	–34.35	–16.89
Live cattle	*0.07*	*0.03*	*–0.09*	*–0.02*	*0.02*	*–0.13*	**142.36**	18.71	–20.08	4.14	–15.27	9.25	–19.75	31.61
Lean hogs	*–0.02*	*–0.02*	*–0.01*	*0.10*	*0.03*	*0.04*	*0.06*	**606.71**	69.63	7.01	426.46	–27.19	19.11	35.20
Coffee “C”	*0.06*	*0.07*	*0.12*	*–0.12*	*–0.15*	*0.06*	*–0.08*	*0.14*	**421.94**	23.83	5.64	–15.59	54.31	–39.85
Sugar #1	*0.01*	*–0.02*	*–0.10*	*–0.02*	*0.07*	*0.00*	*0.02*	*0.02*	*0.06*	**345.21**	95.41	11.69	–28.41	29.00
Cotton #2	*0.17*	*0.19*	*–0.08*	*–0.15*	*–0.13*	*0.18*	*–0.05*	*0.74*	*0.01*	*0.22*	**544.65**	83.64	19.19	68.24
Corn	*–0.04*	*–0.06*	*0.05*	*0.00*	*–0.09*	*–0.02*	*0.04*	*–0.05*	*–0.03*	*0.03*	*0.16*	**475.69**	236.91	320.47
Soybeans	*0.05*	*0.06*	*–0.02*	*0.02*	*–0.00*	*–0.09*	*–0.09*	*0.04*	*0.14*	*–0.08*	*0.04*	*0.57*	**365.44**	162.80
Wheat	*0.00*	*–0.01*	*–0.04*	*–0.01*	*–0.10*	*–0.04*	*0.13*	*0.07*	*–0.10*	*0.08*	*0.14*	*0.73*	*0.42*	**407.53**

Variances and covariances are annualized and further scaled by 10^4. The 1-composite estimator is computed with $k = 5$ and the 2-composite estimator is computed with $k = 3$. Correlations are in the lower triangle in italic font

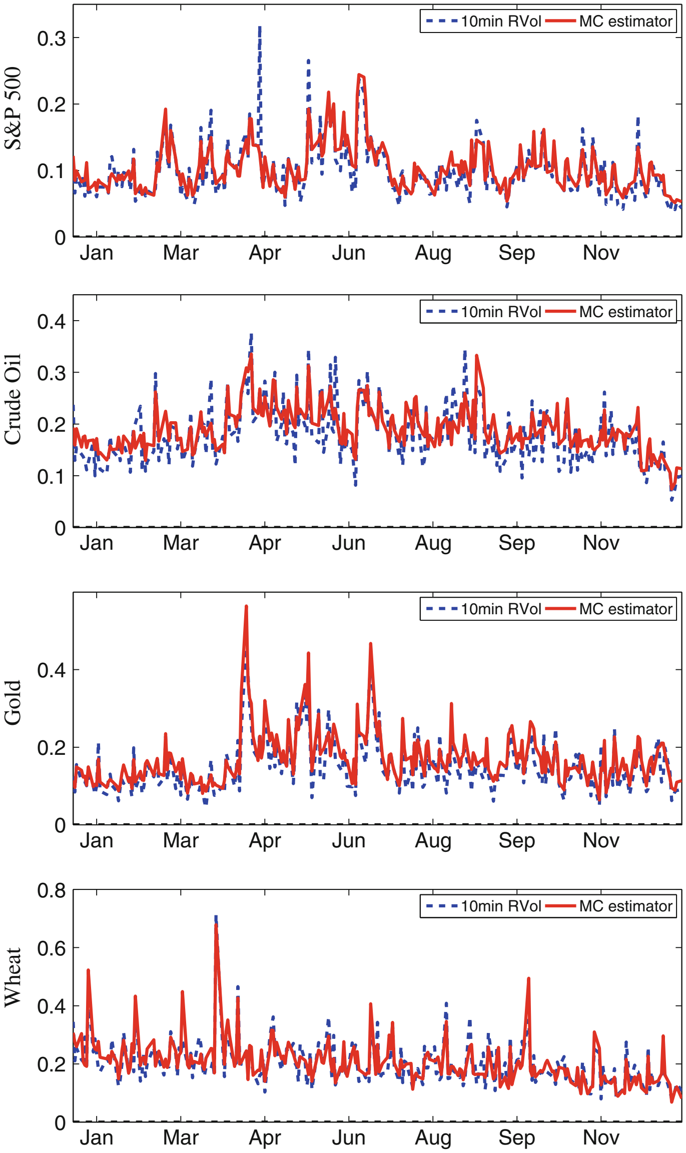


Fig. 3 Realized volatility based on 10 min returns against volatility computed with MC estimator (order $k = 5$) in 2013. Estimated values are annualized

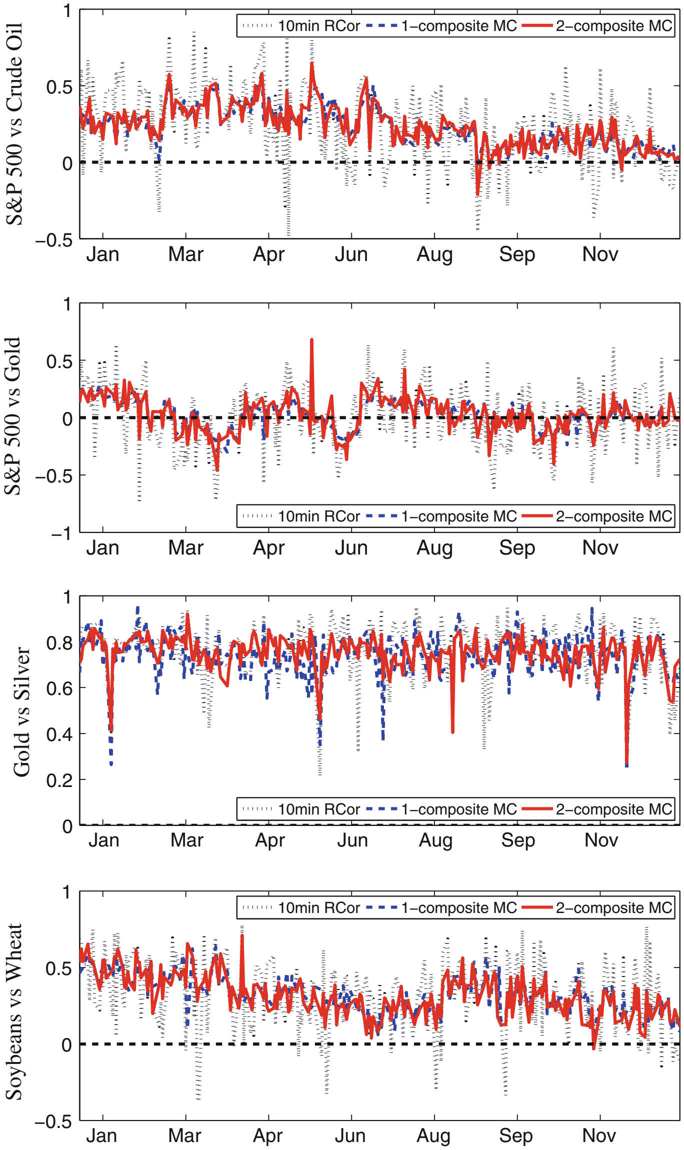


Fig. 4 Realized correlation based on 10 min returns against correlation computed with MC estimator (order $k = 5$) in 2013

degree of time-variation. In the case of SPY/Gold we observe a less stable correlation that changes sign several times during the year. The general patterns are successfully captured by the composite Markov estimators, and while the realized correlation is in agreement about the general trends, it exhibits far more day-to-day variation which suggests that it is less accurate. The smooth and persistent behavior of the Markov estimators may be attributed to these estimators being more accurate.

8 Conclusion

In this paper we have proposed a multivariate volatility estimator that is based on the theory of finite Markov chains. The Markov chain estimator takes advantage of the fact that high-frequency prices are confined to a grid. This is the first robust multivariate estimator for which standard errors are readily available. Previous estimators include the multivariate realized kernel estimator, whose standard error also requires an estimate of the long-run variance of the noise, which is difficult to estimate because the noise is, in practice, small, serially dependent, and endogenous. The multivariate kernel estimator (MRK) converges at rate $n^{1/5}$. In contrast, the Markov estimator converges at rate $n^{1/2}$ owing to the specification assumed for the high-frequency data. These rates are, however, not directly comparable for practical situation, as the order of the Markov chain may be required to increase with n, in order to accommodate inhomogeneity resulting from time-varying volatility. Our simulation design suggests that the Markov estimator and the MRK performs similarly in practice, so the major advantage of the Markov estimator is the readily available standard error.

The estimator performs well in a simulation design, and is relatively insensitive to the choice of the order of the Markov chain, k, which is the tuning parameter that must be chosen in practice.

A potential limitation of the estimator is the high-dimensional objects that the estimator is computed from. For the full estimator the dimension can be as large as $(S^d)^k$, where S is the number of primitive states for the individual series, d is the dimension of the process and k is the order of the Markov chain. The dimension will typically be much smaller in practice because many states are not observed in a given sample, and the transition matrix will be very sparse, because most transitions between states are unobserved. So there is a need to further analyze the finite sample properties of the full Markov estimator, and to characterize its limitations.

The two composite Markov estimators alleviate the challenges with high dimensional objects, but may require a projection to guarantee a positive semidefinite estimate. For this purpose we have proposed a novel projection that makes use of the standard errors of the elements of the matrix being projected. Since these are readily available for the 2-composite estimator it is appealing to incorporate these, so that a projection leaves accurately estimated elements relatively unchanged.

The empirical analysis of commodity prices illustrated the three Markov estimators, and benchmarked them against conventional realized variances. The estimates were largely in agreement, but the Markov estimators fare particularly well with regards to estimating correlations. While the time series of daily correlation estimates based on the realized variance were somewhat erratic, those of the Markov estimators were more stable.

Appendix: Details on the Simulation Design with Stochastic Volatility

For comparison with the simulation design with constant volatility we seek to have the integrated variance be 1, in expectation. This is achieved as follows. Note that

$$\mathrm{E}((\mathrm{d}\log Y_{i,t})^2) = \mathrm{E}\left[\exp\left\{2\left(\beta_0 + \beta_1\tau_{i,t}\right)\right\}\Delta\right],$$

and since we (approximately) have $\tau_{i,t} \sim N(0, a^2)$ with

$$a^2 = \frac{\frac{1-\exp(2\alpha\Delta)}{-2\alpha}}{1-\exp(\alpha\Delta)^2} = \frac{1}{-2\alpha},$$

it follows that

$$2(\beta_0 + \beta_1\tau_{i,t}) \sim N(2\beta_0, 4\beta_1^2\frac{1}{-2\alpha}).$$

Hence

$$\mathrm{E}\left[\exp\left\{2\left(\beta_0 + \beta_1\tau_{i,t}\right)\right\}\right] = \exp\left(2\beta_0 + \beta_1^2\frac{1}{-\alpha}\right),$$

which will be equal to 1 if we set $\beta_0 = \beta_1^2/(2\alpha)$.

References

1. Ait-Sahalia, Y., Fan, J., Xiu, D.: High-frequency covariance estimates with noisy and asynchronous financial data. J. Am. Stat. Assoc. **105**(492), 1504–1517 (2010)
2. Andersen, T., Dobrev, D., Schaumburg, E.: Duration-based volatility estimation. working paper (2008)
3. Andersen, T.G., Bollerslev, T.: Answering the skeptics: yes, standard volatility models do provide accurate forecasts. Int. Econ. Rev. **39**(4), 885–905 (1998)
4. Andersen, T.G., Bollerslev, T., Diebold, F.X., Ebens, H.: The distribution of realized stock return volatility. J. Finan. Econ. **61**(1), 43–76 (2001)
5. Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P.: The distribution of exchange rate volatility. J. Am. Stat. Assoc. **96**(453), 42–55 (2001). Correction published in 2003, vol. 98, p. 501
6. Bandi, F.M., Russell, J.R.: Market microstructure noise, integrated variance estimators, and the accuracy of asymptotic approximations. working paper (2006)
7. Bandi, F.M., Russell, J.R.: Microstructure noise, realized variance, and optimal sampling. Rev. Econ. Stud. **75**, 339–369 (2008)

8. Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N.: Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise. Econometrica **76**, 1481–536 (2008)
9. Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N.: Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. J. Econometrics **162**, 149–169 (2011)
10. Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N.: Subsampling realised kernels. J. Econometrics **160**, 204–219 (2011)
11. Barndorff-Nielsen, O.E., Shephard, N.: Econometric analysis of realised volatility and its use in estimating stochastic volatility models. J. R. Stat. Soc. B **64**, 253–280 (2002)
12. Barndorff-Nielsen, O.E., Shephard, N.: Econometric analysis of realised covariation: high frequency based covariance, regression and correlation in financial economics. Econometrica **72**, 885–925 (2004)
13. Christensen, K., Oomen, R.C., Podolskij, M.: Realised quantile-based estimation of the integrated variance. Working paper (2008)
14. Christensen, K., Podolskij, M.: Realized range-based estimation of integrated variance. J. Econometrics **141**, 323–349 (2007)
15. Christoffersen, P., Lunde, A., Olesen, K.V.: Factor structure in commodity futures return and volatility. Working paper (2014)
16. Delattre, S., Jacod, J.: A central limit theorem for normalized functions of the increments of a diffusion process, in the presence of round-off errors. Bernoulli **3**, 1–28 (1997)
17. Grant, M.C., Boyd, S.P., Ye, Y.: CVX: Matlab Software for Disciplined Convex Programming (2014)
18. Hansen, P.R.: A martingale decomposition of discrete Markov chains. Working paper (2014)
19. Hansen, P.R., Horel, G.: Quadratic variation by Markov chains. Working paper (2009)
20. Hansen, P.R., Horel, G.: Limit theory for the long-run variance of finite Markov chains. Working paper (2014)
21. Hansen, P.R., Large, J., Lunde, A.: Moving average-based estimators of integrated variance. Econometric Rev. **27**, 79–111 (2008)
22. Hansen, P.R., Lunde, A.: Realized variance and market microstructure noise. J. Bus. Econ. Stat. **24**, 127–218 (2006). The 2005 Invited Address with Comments and Rejoinder
23. Hautsch, N., Kyj, L.M., Oomen, R.: A blocking and regularization approach to high dimensional realized covariance estimation. J. Appl. Econometrics **27**, 625–645 (2012)
24. Horel, G.: Estimating integrated volatility with markov chains. Ph.D. thesis, Stanford University (2007)
25. Jacod, J., Li, Y., Mykland, P.A., Podolskij, M., Vetter, M.: Microstructure noise in the continuous case: the pre-averaging approach. Stochast. Process. Their Appl. **119**, 2249–2276 (2009)
26. Kemeny, J., Snell, J.: Finite markov chains. Springer, New York (1976)
27. Ledoit, O., Santa-Clara, P., Wolf, M.: Flexible multivariate GARCH modeling with an application to international stock markets. Rev. Econ. Stat. **85**, 735–747 (2003)
28. Li, Y., Mykland, P.A.: Determining the volatility of a price process in the presence of rounding errors. Working paper (2006)
29. Lunde, A., Shephard, N., Sheppard, K.: Econometric analysis of vast covariance matrices using composite realized kernels. Working paper (2014)
30. Maheu, J.M., McCurdy, T.H.: Nonlinear features of realized FX volatility. Rev. Econ. Stat. **84**, 668–681 (2002)
31. Renò, R.: A closer look at the epps effect. Int. J. Theor. Appl. Finance **6**, 87–102 (2003)
32. Zhang, L.: Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach. Bernoulli **12**, 1019–1043 (2006)
33. Zhang, L., Mykland, P.A., Aït-Sahalia, Y.: A tale of two time scales: determining integrated volatility with noisy high frequency data. J. Am. Stat. Assoc. **100**, 1394–1411 (2005)
34. Zhou, B.: High-frequency data and volatility in foreign exchange rates. J. Bus. Econ. Stat. **14**(1), 45–52 (1996)
35. Zhou, B.: Parametric and nonparametric volatility measurement. In: Dunis, C.L., Zhou, B. (eds.) Nonlinear Modelling of High Frequency Financial Time Series, Chap. 6, pp. 109–123. Wiley, New York (1998)