High-Dimensional Multivariate Realized Volatility Estimation[*]

Tim Bollerslev[†], Nour Meddahi[‡], and Serge Nyawa[§]

October 22, 2018

Abstract

We provide a new factor-based estimator of the realized covolatility matrix, applicable in situations when the number of assets is large and the high-frequency data are contaminated with microstructure noises. Our estimator relies on the assumption of a factor structure for the noise component, separate from the latent systematic risk factors that characterize the cross-sectional variation in the frictionless returns. The new estimator provides theoretically more efficient and finite-sample more accurate estimates of large-scale integrated covolatility and correlation matrices than other recently developed realized estimation procedures. These theoretical and simulation-based findings are further corroborated by an empirical application related to portfolio allocation and risk minimization involving several hundred individual stocks.

Keywords: Realized covolatility matrix; high-dimensional estimation; high-frequency data; microstructure noise; robust measures.

JEL Classification: C13, C32, C58.

[*] We have benefited from comments by Jia Li, Andrew Patton, Kevin Sheppard, George Tauchen, an Associate Editor and two anonymous referees, as well as seminar and conference participants at Toulouse School of Economics, Duke University, the 2016 European Meeting of the Econometric Society, the 2016 Meeting of the International association of Applied Econometrics, and the 2016 Financial Econometrics Conference in Toulouse. The second and third authors acknowledge financial support of the grant ERC POEMH. The second author thanks support from SCOR, Amundi and the Jean-Jacques Laffont Chair in Digital Economy.

[†] Duke University, 213 Social Sciences Building, Box 90097, Durham NC 27708-0097, USA; boller@duke.edu.

[‡] Toulouse School of Economics, 21 Allée de Brienne, 31015 Toulouse, France; nour.meddahi@tse-fr.eu.

[§] Toulouse Business School, s.nyawa@tbs-education.fr.

1 Introduction

We contribute to the literature on the estimation of large-dimensional integrated covolatility matrices from high-frequency intraday data. The covolatility matrix plays a crucial role in many financial applications including risk management, portfolio allocation, hedging and asset pricing, and as such, accurate and well conditioned estimates of the integrated covolatility matrix, its inverse, and the correlation matrix are of great practical import.

Our new covolatility estimator is specifically designed to work in situations when the the number of assets is large and the high-frequency data used in the estimation might be contaminated with microstructure noises. It relies on the assumption of a factor structure for characterizing the microstructure noise component, separate from the factor structure that characterizes the latent genuine returns. The efficiency of the new estimator compares favorably to other recently developed procedures. These theoretical results, derived under the assumption of increasingly finer sampled intraday returns and an increasing number of assets, carry over to more accurate estimates of large-scale integrated covolatility and correlation matrices in empirically realistic situations with hundreds of assets and finitely sampled intraday returns. On applying the new estimator in the construction of minimum variance portfolios with a sample comprised of almost four-hundred individual stocks, it also results in systematically lower ex-post risks than other competing realized covolatility estimation procedures.

To more formally set out the ideas, let $X_t^* = \left(X_{1t}^*, ..., X_{pt}^*\right)'$ denotes the latent p-dimensional frictionless vector log-price process of interest. Importantly, we allow for p to be "large" and possibly in excess of the number of intraday price observations. Consistent with the lack of arbitrage, we will further assume that X_t follows a continuous Itô semimartigale process,

$$dX_t^* \; = \; \mu_t dt \; + \; \sigma_t dB_t, \qquad 0 \leq t \leq 1, \tag{1}$$

where the unit time-interval corresponds to a day, $B_t = \left(B_t^{(1)}, ..., B_t^{(p)}\right)'$ is a p-dimensional vector of standard independent Brownian motions, and $\mu_t = \left(\mu_t^{(1)}, ..., \mu_t^{(p)}\right)'$ and $\sigma_t = \left(\sigma_t^{(1)}, ..., \sigma_t^{(p)}\right)'$ denote a p-dimensional predictable locally bounded drift process and a càdlàg $p \times p$ spot covolatility process, respectively. The object of interest is the $p \times p$ integrated covolatility matrix,[1]

$$ICV \; = \; \int_0^1 \sigma_s \sigma_s^{'} ds. \tag{2}$$

This ex-post measure of the true daily covariation is, of course, latent. By the theory of

[1] Following the literature, we will also interchangeably refer to this as the integrated covariance, integrated volatility, or integrated covariation matrix.

quadratic variation, it may be consistently estimated by the summation of increasingly finer sampled cross-products of the high-frequency frictionless vector return process,

$$RCV \;=\; \sum_{t_i}(X^*_{t_{i+1}} - X^*_{t_i})(X^*_{t_{i+1}} - X^*_{t_i})', \tag{3}$$

where $0 \leq t_i \leq 1$ refer to the within day sampling times, $t_i - t_{i-1} \to 0$. In practice, of course, the X^*_t process is not directly observable. Instead, the actually observed price process, is subject to "noise" stemming from a host of market microstructure complications, including bid-ask spreads, non-trading, price discreteness, trades occurring on different markets or networks, rounding errors, among others (see, e.g., Hansen and Lunde (2006) and Diebold and Strasser (2013)),

$$X_t = X^*_t + u_t. \tag{4}$$

This in turn renders the estimator for ICV based on RCV with the actually observed X_t price process in place of X^*_t inconsistent.

Several competing estimators that remain consistent in the presence of market microstructure noise have been proposed in the univariate case ($p = 1$), including the sub-sampling and averaging approach of Zhang, Mykland, and Ait-Sahalia (2005), the realized kernel of Barndorff-Nielsen, Hansen, and Shephard (2008), and the pre-averaging (henceforth PRV) approach of Jacod, Li, Mykland, Podolskijc, and Vetter (2009). These estimators are naturally extended to the multivariate case ($p > 1$), provided that the observation times of all the assets are synchronous, and the number of assets is smaller than the number of intraday observations. In practice, of course, prices are generally not recorded at the same time for all assets, which can cause naive estimators of the covolatility matrix that pretend the data are synchronous to be seriously biased.[2]

One solution to the non-synchronicity problem is provided by Hayashi and Yoshida (2005), who propose including all overlapping (in time) intraday returns based on the actually observed price series in the calculation of RCV. However, the estimator of Hayashi and Yoshida (2005) doesn't deal with the microstructure noise that plagues the use of high-frequency data more generally. The multivariate realized kernel estimator of Barndorff-Nielsen, Hansen, Lunde, and Shephard (2011) (henceforth $MRker$[3]) simultaneously guarantee consistency,

[2] This effect was first noted empirically for sample correlation matrices by Epps (1979), and it is now commonly referred to as the Epps-effect.

[3] The realized kernel estimator is defined by:

$$K(Y) = \sum_{h=-n}^{n} k(\tfrac{h}{H+1})\Gamma_h,$$

$$\Gamma_h = \sum_{j=h+1}^{n} y_j y^{'}_{j-h}, \text{ for } h > 0; \qquad \Gamma_h = \Gamma^{'}_{-h}, \text{ for } h < 0,$$

positive semi-definiteness, robustness to microstructure noise, while also accounting for non-synchroneity of observations. The non-synchronicity issue, in particular, is resolved using so-called refresh-time sampling. The modulated realized covariance estimator (henceforth MRC) of Christensen, Kinnebrock, and Podolskij (2010), based on a multivariate extension of the univariate pre-averaging approach, also works in the presence of market microstructure noise. However, the MRC estimator assumes synchronous data, and it is not guaranteed to be positive semi-definite. Christensen, Kinnebrock, and Podolskij (2010) introduced the adjusted modulated realized covariance (henceforth MRC^{δ}) and the pre-averaged Hayashi-Yoshida estimator, in order to ensure the positive semi-definiteness, the noise-robustness and to resolve the non-synchronous data problem.

The covolatility estimators discussed above were explicitly designed for situations in which the number of assets is small relative to the number of intraday return observations, or the sample size available for the estimation. Of course, in many practical portfolio allocation, risk measurement and management decisions, the number of assets is often of the same order of magnitude or even larger than the sample size, entailing a curse of dimensionality type problem for any direct estimation of ICV matrix.[4] Two main approaches has emerged in the literature for dealing with this problem: (i) sparsity or decay assumptions pertaining directly to the different entries in the covolatility matrix; and (ii) the use of factor structures.

Estimators that rely on sparsity and decay assumptions include Zhang (2011) and Zheng and Li (2011). These estimators typically postulate that the covolatility matrix is comprised of only a small number of non-zero block diagonal matrices, or that the absolute magnitude of the elements in the matrix somehow decay away from the diagonal.[5] The blocking and regularization approach of Hautsch and Podolskij (2013), in which assets with similar observation frequency are grouped together in order to reduce the data loss stemming from the use of refresh-time sampling, also implicitly builds on similar ideas. As does the composite realized kernel estimator (henceforth $\hat{\Sigma}_{comp}$) of Lunde, Shephard, and Sheppard (2016), in which bivariate realized kernel estimators for all pairs of assets is combined and regularized in the construction of an estimation for the full high-dimensional covolatility matrix for all assets.

The use of factor structures that underly the second approach for high-dimensional re-

where n is the number of synchronized returns per asset, Γ_h is the h^{th} realized auto-covariance; $y_j = Y_j - Y_{j-1}$ for $j = 1, 2, ..., n$; with $Y_0 = \frac{1}{m}\sum_{j=1}^{m} Y(\tau_{p,j})$; $Y_n = \frac{1}{m}\sum_{j=1}^{m} Y(\tau_{p,p-m+j})$; $Y_j = Y(\tau_{p,j+m})$ for $j = 1, ..., n-1$; $\{\tau_{p,j}\}$ is the series of refresh time ; and k is a non-stochastic weighting function. The rate of convergence of this estimator is $n^{-1/5}$.

[4] This mirrors the problem in parametric $GARCH$ and stochastic volatility models, for which the dimensionality of parameter space in unrestricted versions of the models grow at the rate of p^4; see, e.g., Andersen, Bollerslev, Christoffersen, and Diebold (2006).

[5] The decay assumption is often somewhat arbitrary, since there is not a natural ordering of the assets.

alized covolatility matrix estimation, is, of course, omnipresent in finance (see, e.g., Ross (1976), Chen, Roll, and Ross (1986), Sharpe (1994), and Ledoit and Wolf (2003)). The use of this approach in the context of high-frequency data realized covolatility estimation was pioneered by Fan, Fan, and Lv (2008). It has the obvious advantages that it guarantees a positive semi-definite and, under weak conditions, invertible estimate of the covolatility matrix. Fan, Fan, and Lv (2008) further examine how the dimensionality of the problem favorably impact the accuracy of the estimator compared to other procedures. Other related factor-based approaches include Tao, Wang, and Chen (2011) and Bannouh, Martens, Oomen, and van Dijk (2012), who rely on mixtures of high-frequency intraday data and daily date for estimating the covolatility matrix implied by a factor structure; Fan, Liao, and Mincheva (2011) through their approximate factor models[6] for the estimation of high-dimensional covariance matrix; Fan, Liao, and Mincheva (2013) who introduce the Principal Orthogonal Complement Thresholding Estimator (Henceforth, POET) [7]; and the principal component analysis for the estimation of high dimensional factor models recently explored by Ait-Sahalia and Xiu (2017) and Dai, Lu, and Xiu (2018).[8]

Building on these ideas, we propose a new high dimensional covolatility matrix estimator under the assumption that the true dynamics of the returns may be described by a latent factor model. In contrast to the factor-based estimators discussed above, we explicitly allow for the possibility of market microstructure noise in the actually observed price series. Motivated by Hasbrouck and Seppi (2001), we assume that the cross-sectional dependencies in the market microstructure noise component may be described by its own factor model, resulting in two separately identified factor structures: a latent component of order $O_p(\sqrt{\Delta})$ accounting for the genuine cross-sectional dependencies in the returns, which becomes increasingly less important for discretely sampled observations over diminishing time-intervals of length Δ, and another component of order $O_p(1)$ for describing the noise, which remains invariant to the sampling frequency. Exploiting these differences in the orders of magnitude, and appropriately combining noise-robust $MRker$ and PRV-based estimates of the rotated return factors and their integrated volatilities, along with the corresponding loadings and integrated idiosyncratic volatility components, in turn allows for consistent noise-robust estimation of the full covolatility matrix in large dimensions.

The rest of the paper is organized as follow. Section 2 presents the theoretical setup and formally defines the new estimator. Section 3 derives the convergence rate of the new and

[6] They assume observable factors and allow the presence of the cross-sectional correlation in a sparse error covariance matrix.

[7] They assume a sparse error covariance matrix in an approximate factor model, and allow for the presence of some cross-sectional correlation, after taking out common but unobservable factors.

[8] They rely on the pre-averaging method with refresh time to solve the microstructure problems, while using three different specifications of factor models, and their corresponding estimators, respectively, to battle against the curse of dimensionality.

other competing estimators. This section also presents the results from a set of finite-sample simulations involving both synchronous and asynchronous high-frequency prices. Section 4 presents the results from an empirical application involving a large cross-section of individual stocks. Section 5 concludes. The details of the proofs and other more specific materials are deferred to Appendixes.

2 Theoretical setup

2.1 The benchmark model

We assume that the continuous Itô semimartingale process X_t in (1) follows a factor model of the form,

$$dX_t^* = bdF_t + dE_t, \tag{5}$$

where $b = (b_{ik})_{1 \leq i \leq p, 1 \leq k \leq K}$ denotes the $p \times K$ matrix of factor loadings, $F_t = (F_{1t}, ..., F_{Kt})'$ refers to the latent factor vector, K is assumed to be asymptotically finite and known, and $E_t = (E_{1t}, ..., E_{pt})'$ denotes the vector of idiosyncratic errors. The use of factor models in asset pricing finance is, of course, quite standard and traces back to the seminal work by Ross (1976) and Chamberlain and Rothschild (1983). The factor F_t is supposed to represent general influences which tend to affect all assets. Following standard assumptions in the literature, we assume that factor loadings b are time invariant and do not depend on t.

We further assume that the F_t and E_t vectors and the individually components therein are uncorrelated and driven by their own standard Brownian motions,

$$dF_{kt} = \sigma_{fkt} dB_{kt}^F,$$
$$dE_{it} = \sigma_{\varepsilon it} dB_{it}^E.$$

Integrating both sides of the resulting latent factor price process above over a time interval of length Δ, it readily follows that

$$\textstyle\int_{t-\Delta}^t dX_s^* = b \cdot \int_{t-\Delta}^t \sigma_{fs} dB_s^F + \int_{t-\Delta}^t \sigma_{\varepsilon s} dB_s^E.$$

Defining the corresponding returns, factors, and errors over the time-interval Δ,

$$r_t^* \equiv r_{t,\Delta}^* \equiv \textstyle\int_{t-\Delta}^t dX_s^*$$
$$f_t \equiv f_{t,\Delta} \equiv \textstyle\int_{t-\Delta}^t \sigma_{fs} dB_s^F$$
$$\varepsilon_t \equiv \varepsilon_{t,\Delta} \equiv \textstyle\int_{t-\Delta}^t \sigma_{\varepsilon s} dB_s^E$$

allows for following standard discrete-time factor representation,

$$r_t^* = bf_t + \varepsilon_t \tag{6}$$

where $r_t^* = (r_{1t}^*, ..., r_{pt}^*)'$, $f_t = (f_{1t}, ..., f_{Kt})'$, and $\varepsilon_t = (\varepsilon_{1t}, ..., \varepsilon_{pt})'$, respectively.

We make the additional assumptions directly pertaining to this representation, where $I_{t-\Delta}$ refers to information set available at time $t - \Delta$.

Assumption 1 $\forall t$, $\forall i, j, k, k' \in \{1, ..., p\}$, $i \neq j$, $k \neq k'$:

- $Cov\left(f_{kt}, \varepsilon_{it} | I_{t-\Delta}\right) = 0$;
- $Cov\left(f_{kt}, f_{k't} | I_{t-\Delta}\right) = 0$;
- $Cov(\varepsilon_{it}, \varepsilon_{jt} | I_{t-\Delta}) = 0$;
- $E\left(\varepsilon_{it} | I_{t-\Delta}\right) = 0$.

The latent X_{it}^* prices for each of the p individual assets are not directly observable. Instead, the actually observed prices are contaminated with market microstructure noise,

$$X_{it} = X_{it}^* + u_{it}. \tag{7}$$

We assume that this noise component has its own separate factor representation,

$$u_{it} = c_i g_t + \eta_{it}, \tag{8}$$

where the $K' \times 1$ g_t vector accounts for the cross-sectional dependence in the noise, and the $1 \times K'$ c_i vector denotes the corresponding factor loadings. We make the following additional assumptions about this structure.

Assumption 2 $\forall t$, $\forall i, k, k' \in \{1, ..., p\}$, $k \neq k'$:

- $Cov\left(g_{kt}, f_{k't} | I_{t-\Delta}\right) = 0$;
- $Cov\left(g_{kt}, \varepsilon_{it} | I_{t-\Delta}\right) = 0$;
- $Cov\left(\eta_{it}, f_{kt} | I_{t-\Delta}\right) = 0$, $Cov\left(\eta_{it}, g_{kt} | I_{t-\Delta}\right) = 0$, $Cov\left(\eta_{it}, \varepsilon_{it} | I_{t-\Delta}\right) = 0$;
- $Var(\eta_{it}) = \sigma_{\eta i}^2$, $\forall i \in \{1, ..., p\}$;
- $Var(g_{kt}) = \sigma_{g_k}^2$;
- g_{kt}, η_{it} are independent across assets and time.

Two main types of factors models are presented in the existing literature: strict factor models and approximate factor models. The main difference between these models is the assumption on the covariance matrix of idiosyncratic components. In a strict factor model, this matrix is assumed to be diagonal, while, its terms can be weakly correlated in an approximated factor model. For an identification purpose, the following assumptions are widely made:

- *Pervasiveness*: the factors influence a large number of assets which means that the loading vectors b are bounded and $\|\frac{1}{p}b'b - D\| \longrightarrow 0$ as $p \longrightarrow \infty$, where D is a $K \times K$ positive definite matrix;

- *Factors*: the fourth moment of factors exists and the covariance function of factors converges to a definite positive matrix as $1/\Delta \longrightarrow \infty$;

- Approximate factor models may exhibit both temporal and cross-sectional dependencies, as well as heteroskedastic error terms.

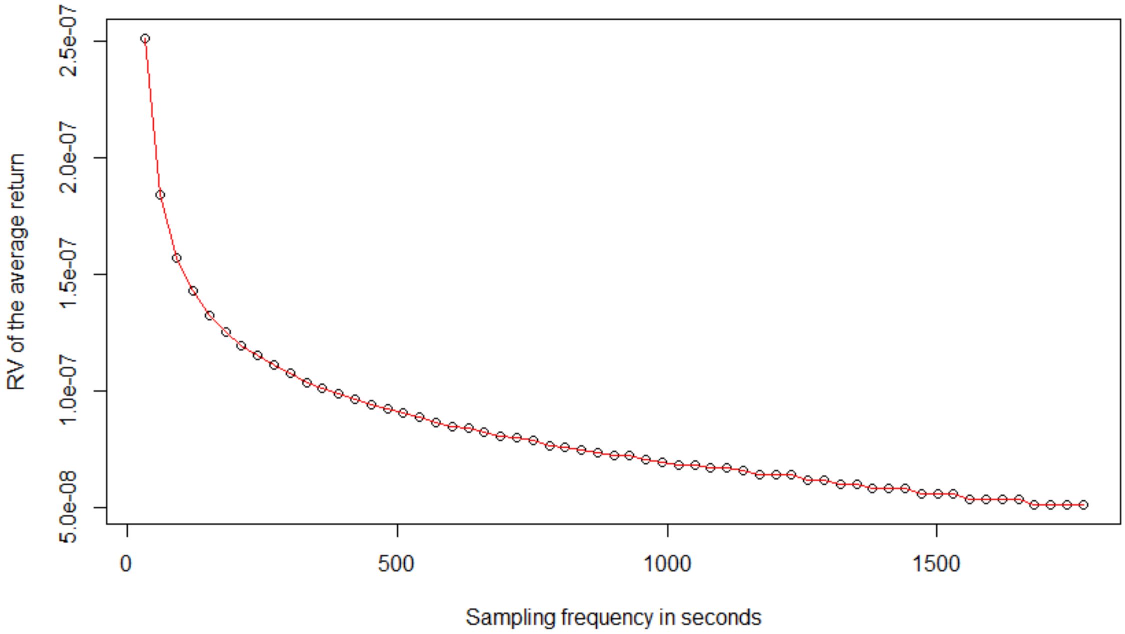
Our model is a strict factor model with some normalization assumptions: i) the pervasiveness assumption holds with $D = I_p$; ii) the fourth moments of factors exist and the covariance function of the factors converges to a diagonal matrix without loss of generality, as $1/\Delta$ goes to infinity; iii) we rule out the existence of time and cross-section dependence and heteroscedasticity of idiosyncratic terms which is left for future research.

As discussed further below, the assumption of a separate factor representation for the microstructure noise makes it possible to disentangle the estimation of the covolatility matrix into two parts: a traditional factor-based approach for the estimation of the latent component of order $O_p(\sqrt{\Delta})$ associated with the traditional factor structure in the returns, and a separate estimation of the factor noise components of order $O_p(1)$.

The use of a factor structure for the microstruture noise is directly motivated by Hasbrouck and Seppi (2001), who document strong commonalities in various liquidity proxies such as the bid-ask spread. To further corroborate the dominance of common factors in the noise, we run two empirical exercises.

Firstly, we construct the signature plot of the cross-sectional average return, computed from a sample of 384 individual stocks analyzed in the empirical section below. Under a cross-sectional uncorrelation of microstructure noise, the noise component is supposed to vanish by the law of large numbers. As a consequence, the resulting signature plot is supposed to be flat. However, as presented in figure 1, we obtain a strictly decreasing curve. This is an evidence that the cross-sectional average return still contain a microstructure term. Thus, microstructure noises must be cross-sectionally correlated and common factors may capture this cross-sectional correlation.

Figure 1. Signature plot of the cross-sectional average return



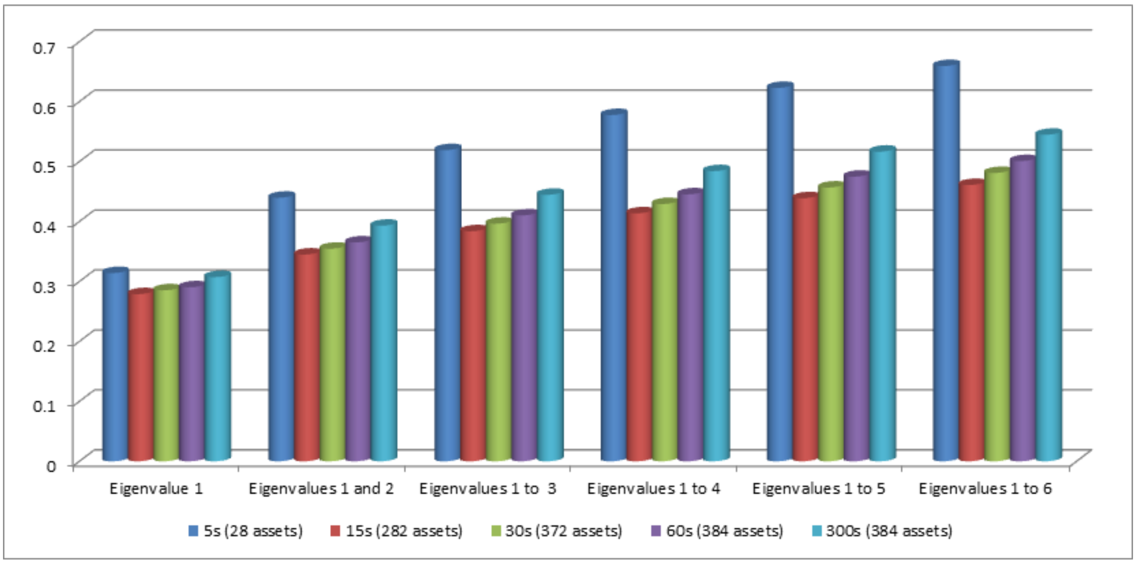
Note: This figure plots the signature plot of the cross-sectional average return, computed from a sample of 384 individual stocks analyzed in the empirical section below.

Secondly, we estimated the covariance matrix for the market microstructure noise for the same sample. Decomposing the resulting covariance matrix estimates for each day in the sample, strongly supports the idea that the cross-sectional dependencies may be adequately captured by a few factors. Further details concerning these results are provided in Appendix A.3.

Figure 2 depicts the average shares of the total variability in the observed returns which can be explained by the first six factors. The analysis is done for various frequencies: 5, 15, 30, 60 and 300 seconds. It is well-known that the variance of the market microstructure is better estimated at the highest frequency. Thus, the higher the sampling frequency, the more accurate is the estimation of the shares of the total variability of microstructural noise that can be explained by factors. However, when one increases the frequency, one has less assets. Estimations based on 15, 30 and 60 seconds are robust and corroborate the factor structure of the noise. At the 300 seconds frequency, the observed factor structure concerns latent returns. Clearly, Figure 2 supports the factor structure of the noise, especially at the 5-seconds frequency, even if the number of assets is relatively small.[9]

[9] At the 5 seconds frequency, the number of stocks involved drops drastically (only 28 assets remain in the sample, in contrast to the other cases involving more than 282 assets). In general, the factors are better approximated the larger the number of stocks. Correspondingly, the cases of 60s, 30s, and 15s sampling provide more reliable information about the factor structure of the microstructure noise. Note that the ratios aren't necessarily monotonically decreasing with the sampling frequency, as the factors driving the

Figure 2. Ratio of largest eigenvalues relative to the total variation



Note: This figure plots the average shares of the total variability in the microstructure noises which can be explained by the first six factors. The analysis is done for various frequencies: 5, 15, 30, 60 and 300 seconds.

2.2 Estimation methodology

The general setup and assumptions outlined in the previous section implies that the integrated covolatility matrix of interest may be succinctly expressed as,

$$\Sigma = bDiag\left[\int_0^1 \sigma^2_{f1u}du, ..., \int_0^1 \sigma^2_{fKu}du\right] b' + Diag\left[\int_0^1 \sigma^2_{\varepsilon 1u}du, ..., \int_0^1 \sigma^2_{\varepsilon pu}du\right]. \quad (9)$$

We rely on traditional factor analysis together with the pre-averaging approach for conveniently estimating the different components of Σ. As usual, the factors and the factor loadings are only determined up to a rotation.[10] Correspondingly, our estimation strategy is comprised of four separate steps for estimating:

- The rotated factors $\tilde{f}$.
- The integrated volatilities of $\tilde{f}$.
- The rotated loadings $\tilde{b}$.
- The integrated volatility of the idiosyncratic component.

fundamental prices start to play a role.

[10] Let H denote a $K \times K$ orthogonal H matrix such that $H'H = I_K$. The Σ matrix defined by the rotated factors $\tilde{f}_t = Hf_t$ and rotated factor loadings $\tilde{b} = bH'$, is then identical to the matrix in (9).

We will discuss each of these four steps in turn. We will begin by assuming that all of the high-frequency returns used in the estimation span the same time-interval of length Δ, with $\Delta \to 0$ corresponding to continuous-time case. However, we will also subsequently consider the empirically more realistic case with unevenly spaced non-synchronous discrete-time observations.

2.2.1 Estimation of $\tilde{f}$

Following the Principal Component Analysis (henceforth PCA) of Connor and Korajczyk (1988), $f_{j\Delta}$ is chosen to minimize the scaled sum of squared values of the idiosyncratic component,

$$\begin{cases} \underset{f_{j\Delta},b}{Min} \ \frac{1}{p} \sum\limits_{j=1}^{\lfloor 1/\Delta \rfloor} (r^*_{j\Delta} - bf_{j\Delta})'(r^*_{j\Delta} - bf_{j\Delta}) \\ s.t \ \frac{1}{p} b'b = I_K \end{cases}$$

It follows readily from the solution to this optimization problem that

$$\hat{f}_{k\Delta} = \tfrac{1}{p} W' r^*_{k\Delta}, \quad \forall k = 1, ..., \lfloor 1/\Delta \rfloor,$$

where W denotes the matrix of ordered eigenvectors of $\sum\limits_{j=1}^{\lfloor 1/\Delta \rfloor} \left[r^*_{j\Delta} r^{*'}_{j\Delta} \right]$. Taking $\Delta \to 0$, we obtain the continuous time expression,

$$\hat{f}_t = \frac{1}{p} W' r^*_t, \tag{10}$$

in which the columns of W correspond to the ordered eigenvectors of Σ.

The estimator defined by equation (10) is not feasible because r^*_t and Σ are latent. In order to obtain a feasible estimator, we need consistent estimates of the ordered eigenvectors W of Σ. Let $\hat{W}$ denote the matrix of K ordered eigenvectors of an estimator $\hat{\Sigma}$ of Σ that is robust to microstructure noise. The simulation results in Appendix A.5 shows that *MRker* provides a good candidate.[11] Hence, we propose as feasible estimator:

$$\hat{f}_t = \frac{1}{p} \hat{W}' r_t, \tag{11}$$

where r_t is the $p \times 1$ vector of observed returns, $\hat{W} = \left(\underline{\hat{W}}_1, ..., \underline{\hat{W}}_K \right)$ is a consistent estimator of the $p \times K$ matrix W of ordered eigenvectors of Σ provided by *MRker*.

[11] This approach mirrors the "Linear Shrinkage" estimator of the covariance matrix of Ledoit and Wolf (2003). In order to improve the covariance matrix estimator in large dimensions, a "Linear Shrinkage" estimator is obtained from the spectral decomposition of the sample covariance matrix by keeping the eigenvectors, while transforming the eigenvalues.

We need to verify that the resulting $\hat{f}$ consistently estimates a rotation $\tilde{f}$ of f plus a microstruture noise component. To do so, we express $\hat{f}$ as a function of the true factor f, the idiosyncratic component ϵ_t, and the factor representation of the microstructure noise component u_t

$$\hat{f}_t \quad = \quad \tfrac{1}{p}\hat{W}'bf_t + \tfrac{1}{p}\hat{W}'\epsilon_t + \tfrac{1}{p}\hat{W}'c(g_t - g_{t-\Delta}) + \tfrac{1}{p}\hat{W}'(\eta_t - \eta_{t-\Delta})$$

The consistency result in the estimation of a rotation $\tilde{f}$ of f contaminated by a microstructure noise component is given in the following theorem inspired by the paper of Stock and Watson (2002).

Lemma 2.1 *There exists an orthogonal matrix S such that $S\hat{f}$ consistently estimates f up to a microstruture noise component, so that for $\Delta \to 0$ and $p \to \infty$:*

- $\frac{1}{p}S\hat{W}'bf_t \xrightarrow{p} f_t$.
- $\frac{1}{p}S\hat{W}'\epsilon_t \xrightarrow{p} 0$.
- $\frac{1}{p}S\hat{W}'(\eta_t - \eta_{t-\Delta}) \xrightarrow{p} 0$.

Proof: See the Supplementary Appendix (section 1).

2.2.2 Estimation of $\int_0^1 \sigma^2_{\tilde{f}ku} du$

Consider the following decomposition of $\hat{f}_t$,

$$\hat{f}_{kt} = \tfrac{1}{p}W_k' r_t^* + \tfrac{1}{p}W_k'(u_t - u_{t-\Delta}) + \tfrac{1}{p}W_k^{\epsilon'} r_t^* + \tfrac{1}{p}W_k^{\epsilon'}(u_t - u_{t-\Delta}),$$

where $W_k^{\epsilon'}$ is the error term in the estimation of W. We assume that $\frac{1}{p}W_k^{\epsilon'} r_t^*$ and $\frac{1}{p}W_k^{\epsilon'}(u_t - u_{t-\Delta})$ are of orders smaller than $max(n,p)^{(-1/2)}$.[12] Since $\frac{1}{p}W_k'\epsilon_t = O_p(n^{-1/2}p^{-1/2})$ and $\frac{1}{p}W_k'(\eta_t - \eta_{t-\Delta}) = O_p(p^{-1/2})$, it follows that

$$\hat{f}_{kt} \quad = \quad \tilde{f}_{kt} + \tfrac{1}{p}W_k'c(g_t - g_{t-\Delta}) + O_p(p^{-1/2})$$

For n and p sufficiently large,

$$\hat{f}_{kt} \quad \approx \quad \tilde{f}_{kt} + \tfrac{1}{p}W_k'c(g_t - g_{t-\Delta})$$

[12] The intuition is that p and n are sufficiently large such that the error components $\frac{1}{p}W_k^{\epsilon'} r_t^*$ and $\frac{1}{p}W_k^{\epsilon'}(u_t - u_{t-\Delta})$ are dominated by their latent counterparts, $\frac{1}{p}W_k' r_t^*$ and $\frac{1}{p}W_k'(u_t - u_{t-\Delta})$ respectively . These two latent components are respectively of orders $n^{-1/2}$ and $p^{-1/2}$. The simulation results presented in the Appendix A.5 show that errors in the estimation of W are very small and decreases with p and n.

Note that $\hat{f}$ is effectively a rotation of the latent factor f contaminated by microstructure noises. Hence, by the literature on the estimation of integrated volatility using data contaminated by microstructure noise, $\int_0^1 \sigma^2_{\tilde{f}ku} du$ can be estimated by,

$$\widehat{\int_0^1 \sigma^2_{\tilde{f}ku} du} = PRV(\hat{f}_k), \tag{12}$$

where the PRV estimator is defined in Appendix A.1.

2.2.3 Estimation of $\tilde{b}_{ik}$

Since the factors are pairwise independent and also independent of the idiosyncratic component, it follows that the integrated covolatility matrix for r_i^* and $\tilde{f}_k$ equals $\tilde{b}_{ik}.IV(\tilde{f}_k)$. Thus, $\tilde{b}_{ik} = ICV(r_i^*, \tilde{f}_k)/IV(\tilde{f}_k)$, so that an estimate for $\tilde{b}_{ik}$ is naturally obtained by,

$$\hat{b}_{ik} = \frac{MRC(r_i, \hat{f}_k)}{PRV(\hat{f}_k)}. \tag{13}$$

with the MRC estimator formally defined in Appendix A.1.

2.2.4 Estimation of $\int_0^1 \sigma^2_{\varepsilon iu} du$

Define $\hat{\epsilon}_{it} = r_{it} - \sum_{k=1}^{K} \hat{b}_{ik} \cdot \hat{f}_{kt}$. It is easy to show that

$$\hat{\epsilon}_{it} = \epsilon_{it} + (u_t - u_{t-\Delta}) - \sum_{k=1}^{K} \tilde{b}_{ik}\tilde{f}^{\epsilon}_{kt} - \sum_{k=1}^{K} \tilde{b}^{\epsilon}_{ik}\tilde{f}_{kt} - \sum_{k=1}^{K} \tilde{b}^{\epsilon}_{ik}\tilde{f}^{\epsilon}_{kt} - \tfrac{1}{p}\sum_{k=1}^{K}\sum_{l=1}^{K} \tilde{b}_{ik}W_l'c(g_t - g_{t-\Delta}) - \tfrac{1}{p}\sum_{k=1}^{K}\sum_{l=1}^{K} \tilde{b}^{\epsilon}_{ik}W_l'c(g_t - g_{t-\Delta})$$

where $\tilde{f}^{\epsilon}_{kt}$ and $\tilde{b}^{\epsilon}_{ik}$ denote the estimation errors in the estimation of $\tilde{f}_{kt} + \frac{1}{p}\sum_{k=1}^{K} W_k'c(g_t - g_{t-\Delta})$ and $\tilde{b}_{ik}$, respectively. Since $\tilde{f}^{\epsilon}_{kt} = O_p(p^{-1/2})$ and $\tilde{b}^{\epsilon}_{ik} = O_p(n^{-1/4})$, let's assume that n and p are both sufficiently large such that $\sum\limits_{k=1}^{K} \tilde{b}_{ik}\tilde{f}^{\epsilon}_{kt}$, $\sum\limits_{k=1}^{K} \tilde{b}^{\epsilon}_{ik}\tilde{f}_{kt}$, $\sum\limits_{k=1}^{K} \tilde{b}^{\epsilon}_{ik}\tilde{f}^{\epsilon}_{kt}$ and $\frac{1}{p}\sum_{k=1}^{K}\sum_{l=1}^{K} \tilde{b}^{\epsilon}_{ik}W_l'c(g_t - g_{t-\Delta})$ can be neglected. Then,

$$\hat{\epsilon}_{it} \quad \approx \quad \epsilon_{it} + (u_t - u_{t-\Delta}) - \tfrac{1}{p}\sum_{k=1}^{K}\sum_{l=1}^{K} \tilde{b}_{ik}W_l'c(g_t - g_{t-\Delta}),$$

it follows that $\hat{\epsilon}_{it}$ equals the idiosyncratic component ϵ_{it} contaminated with microstruture noise. Thus, $\int_0^1 \sigma^2_{\varepsilon iu} du$ may be consistently estimated by,

$$\widehat{\int_0^1 \sigma^2_{\varepsilon iu} du} = PRV(\hat{\epsilon}_i). \tag{14}$$

2.2.5 Putting the pieces together

Our covolatility matrix estimator is defined by plugging the different estimators discussed above into the expression for $\widehat{\Sigma}$ in equation (9),

$$\begin{aligned}
\widehat{\Sigma} &= \begin{pmatrix} \hat{b}_{11} & \cdots & \hat{b}_{1K} \\ \vdots & & \vdots \\ \hat{b}_{p1} & \cdots & \hat{b}_{p1} \end{pmatrix} \begin{pmatrix} \int_0^1 \widehat{\sigma^2_{f1u}} du & & \\ & \ddots & \\ & & \int_0^1 \widehat{\sigma^2_{fKu}} du \end{pmatrix} \begin{pmatrix} \hat{b}_{11} & \cdots & \hat{b}_{p1} \\ \vdots & & \vdots \\ \hat{b}_{1K} & \cdots & \hat{b}_{pK} \end{pmatrix} \\
&+ \begin{pmatrix} \int_0^1 \widehat{\sigma^2_{\varepsilon 1u}} du & & \\ & \ddots & \\ & & \int_0^1 \widehat{\sigma^2_{\varepsilon pu}} du \end{pmatrix} \\
&= \begin{pmatrix} \frac{MRC(r_1,\hat{f}_1)}{PRV(\hat{f}_1)} & \cdots & \frac{MRC(r_1,\hat{f}_K)}{PRV(\hat{f}_K)} \\ \vdots & & \vdots \\ \frac{MRC(r_p,\hat{f}_1)}{PRV(\hat{f}_1)} & \cdots & \frac{MRC(r_p,\hat{f}_K)}{PRV(\hat{f}_K)} \end{pmatrix} \begin{pmatrix} PRV(\hat{f}_1) & & \\ & \ddots & \\ & & PRV(\hat{f}_K) \end{pmatrix} \\
&\begin{pmatrix} \frac{MRC(r_1,\hat{f}_1)}{PRV(\hat{f}_1)} & \cdots & \frac{MRC(r_p,\hat{f}_1)}{PRV(\hat{f}_1)} \\ \vdots & & \vdots \\ \frac{MRC(r_1,\hat{f}_K)}{PRV(\hat{f}_K)} & \cdots & \frac{MRC(r_p,\hat{f}_K)}{PRV(\hat{f}_K)} \end{pmatrix} + \begin{pmatrix} PRV(\hat{\epsilon}_1) & & \\ & \ddots & \\ & & PRV(\hat{\epsilon}_p) \end{pmatrix}.
\end{aligned}$$

Or, more succinctly,

$$\widehat{\Sigma}_{ij} = \sum_{k=1}^{K} \frac{MRC(r_i, \hat{f}_k).MRC(r_j, \hat{f}_k)}{PRV(\hat{f}_k)}; \qquad \widehat{\Sigma}_{ii} = \sum_{k=1}^{K} \frac{MRC(r_i, \hat{f}_k)^2}{PRV(\hat{f}_k)} + PRV(\hat{\epsilon}_i), \tag{15}$$

for $i, j = 1, ..., p$[13].

Remark: Our estimator is constructed using the pre-averaging estimator PRV and the modulated realized covariance estimator MRC. Since those two estimators have been adapted in the literature to account for serially correlated microstructure noises (see, e.g., Jacod, Li, Mykland, Podolskijc, and Vetter (2009) and Hautsch and Podolskij (2013)), our estimator can easily be adapted into this specific setting. Our setup can also be easily adapted to account for semi-martingale processes with jumps. Tools used in this paper for the estimation strategy ($MRKer$, MRC and PRV) have extensions to the case of semi-martingale processes with jumps. Additionally, as in Pelger (2018), the model can also be split into two sub-models: i) a factor representation for small movement of returns; ii) and a factor representation for big movements using a threshold to identify jumps. Only the first model can be used for the estimation of integrated volatility. Moreover, our model

[13] Due to the factor structure of our estimator $\widehat{\Sigma} = \hat{b}\widehat{\Sigma}_f \hat{b}' + \widehat{\Sigma}_\varepsilon$ and since $\widehat{\Sigma}_f$ and $\widehat{\Sigma}_\varepsilon$ are diagonal matrices with positive elements, the positive semi-definiteness is guaranteed. It can be easily shown that: $\forall X,\ X'\widehat{\Sigma}X \geq 0$.

may also be extended to an approximate factor structure. In that case, the factors may be extracted using the procedure in Bai and Ng (2002); the loadings and idiosyncratic terms will be estimated using the same procedure discussed in section 2. Estimation of the additional parameters describing the covolatility between the idiosyncratic terms, may be handled using $MRC(\hat{\varepsilon}_i, \hat{\varepsilon}_j)$. The convergence rate of our estimator under the Frobenius norm will not be affected, since estimation errors generated by $MRC(\hat{\varepsilon}_i, \hat{\varepsilon}_j)$, $\forall i \neq j$, are $O_p(\sqrt{p(p-1)}n^{-1/4})$.

3 Comparing different estimators

3.1 Convergence rates

Our new estimator defined in (15) consistently estimates Σ for $\Delta \to 0$ and $p \to \infty$. It is instructive to more formally assess how the values of $n = 1/\Delta$ and p impact the estimation errors. The following lemma provides the specific convergence rates for the integrated volatilities, the loadings of the rotated factors, and the integrated covolatility matrix of the idiosyncratic errors, where $\|.\|_F$ denotes the Frobenius norm.[14]

Lemma 3.1 *Under Assumptions 1-2, for $n \to \infty$ and $p \to \infty$:*

- $\left|\hat{\Sigma}^{\tilde{f}}_{kk} - \Sigma^{\tilde{f}}_{kk}\right| = O_p\left(n^{-1/4}\right)$.
- $\left\|\hat{b}_k - b_k\right\|_F = \left\|\hat{b}_k - b_k\right\|_2 = O_p\left(p^{1/2}n^{-1/4}\right)$.
- $\left\|\hat{\Sigma}^{\epsilon} - \Sigma^{\epsilon}\right\|_F = O_p(p^{1/2}n^{-1/4})$.

Proof: See the Supplementary Appendix (section 1).

Appropriately combining these convergence rates for the individual components, it is possible to deduce the overall rate of convergence of $\hat{\Sigma}$. In order to compare this rate to other competing large dimensional realized covolatility estimators, the following Theorem provides the convergence rate for $\hat{\Sigma}$ along with the rates for the adjusted modulated realized covariance estimator MRC^{δ} of Christensen, Kinnebrock, and Podolskij (2010), the multidimensional kernel estimator *MRker* of Barndorff-Nielsen, Hansen, Lunde, and Shephard (2011), and the composite realized kernel $\hat{\Sigma}_{comp}$ of Lunde, Shephard, and Sheppard (2016).

Theorem 3.1 *Under Asumptions 1-2, for $n \to \infty$ and $p \to \infty$:*

- $\left\|\hat{\Sigma} - \Sigma\right\|_F = O_p(pn^{-1/4})$.
- $\left\|MRC^{\delta} - \Sigma\right\|_F = O_p(pn^{-1/5})$.

[14] The Frobenius norm for the matrix $A = (a_{ij})_{1\leq i,j\leq p}$ is formally defined by $\|A\|_F = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p}|a_{ij}|^2}$.

- $\|MRker - \Sigma\|_F = O_p(pn^{-1/5})$.
- $\left\|\hat{\Sigma}_{comp} - \Sigma\right\|_F = O_p(\sqrt{p(p-1)}n^{-1/5})$.

Proof: See the Supplementary Appendix (section 1).

The results in Theorem 3.1 suggest that under the Frobenius norm, the dimensionality of the covolatility matrix reduces the speed of convergence for the new $\hat{\Sigma}$ estimator by an order of p. Of course, this is also the case for all of the other estimators. Meanwhile, the speed of convergence of $\hat{\Sigma}$ exceeds that of MRC^{δ}, $MRker$ or $\hat{\Sigma}_{comp}$.

The next theorem derives the convergence rate of $\hat{\Sigma}^{-1}$.

Theorem 3.2 *Under Asumptions 1-2, for $n \to \infty$ and $p \to \infty$:*

$$\left\|\hat{\Sigma}^{-1} - \Sigma^{-1}\right\|_F = O_p(p^2 n^{-1/4})$$

Proof: See the Supplementary Appendix (section 1).

The simulation results discussed in the next section confirm that this superior asymptotic performance carries over to empirically realistic finite-sample settings.

3.2 Finite-sample simulations: synchronous prices

We simulate artificial high-frequency prices from a K-factor(s) continuous-time stochastic volatility model in which the actually observed prices are contaminated by noise. While K is allowed to vary from 1 to 5, we only report in this section results for the case $K = 2$. Other simulation results are provided in the Supplement Appendix. We add as competitors, two PCA-based estimators of the covolatility matrix, namely: the *POET* estimator of Fan, Liao, and Mincheva (2013) and the PCA-based estimator of Dai, Lu, and Xiu (2018)(henceforth, *PCA-PRV*). Specific details concerning the simulation design are provided in Appendix A.4.

We begin by simulating frictionless price vectors of length $p = 50$, $p = 100$, $p = 300$ and $p = 500$ based on the true covolatility matrix Σ. We then generate noisy prices by adding market microstructure noise to the vectors of frictionless prices. Each path of the noisy price vector is comprised of $n + 1$ observations. We start by assuming that all of the prices are synchronously recorded, with one observation every five minutes and a trading day of 6.5 hours, resulting in 79 prices per day.[15] We also have simulation results for other sampling frequencies such as: one observation every minute and one observation every 30 seconds (cf. appendix). We consider three different levels of noise in the simulation setup, corresponding

[15] This closely mirrors Lunde, Shephard, and Sheppard (2016), who report around 100 observations on average per day after the synchronization of 473 liquid stocks.

to three values of the signal-to-noise ratio parameter ξ^2: 0.001, 0.005, and 0.01. Due to a space constraint, we only report the results for $K = 2$, $\xi^2 = 0.005$ and 79 prices per day. Results of other cases are reported in the Supplement Appendix.

We evaluate the performance of the same four estimators of Σ analyzed in Theorem 3.1 by computing the errors relative to the true integrated covolatility matrix (columns labeled *Covariance* in the tables), the integrated correlation matrix (columns labeled *Correlation*), and the inverse of the integrated covariance matrix (columns labeled *Inverse*). We rely on the scaled Frobenius norm for assessing the difference between the estimates and the true matrices.[16]

Tables 1 presents the average values based on $1,000$ Monte Carlo replications, with the standard deviations across the simulations reported in parentheses. The new $\hat{\Sigma}$ estimator systematically outperforms all of the five alternative estimators $\hat{\Sigma}_{comp}$, $MRker$, MRC^{δ}, $PCA-PRV$ and $POET$, in terms of most accurately estimating the true covolatility matrix. This holds true across all of the different noise levels and the four values of p. As a whole, the estimation errors systematically increase with the dimensionality of the matrix and the magnitude of the market microstructure noise. These results, of course, are consistent with the theoretical predictions from Theorem 3.1. Looking at columns five and six, which report the separate (unscaled) norms for estimating the diagonal and the off-diagonal elements in Σ, it does not appear that the more accurate estimates afforded by the new $\hat{\Sigma}$ estimator come solely from one or the other. Interestingly, the $\hat{\Sigma}_{comp}$ estimator of Lunde, Shephard, and Sheppard (2016) appears to perform especially poorly for estimating the diagonal variance elements.

This superior performance of the $\hat{\Sigma}$ estimator carries over to the estimation of the correlation matrix implied by the true covolatility matrix. It also holds true for estimating Σ^{-1} for low noise levels[17]. However, $\hat{\Sigma}_{comp}^{-1}$ performs slightly better than $\hat{\Sigma}^{-1}$ for estimating Σ^{-1} for higher levels of market microstructure noise. Also, whereas $\hat{\Sigma}_{comp}$ and $\hat{\Sigma}$ are both guaranteed to be positive semi-definite, the inverse of both $MRker$ and MRC^{δ} fails to exist when $p > n$, and $MRker^{-1}$ and $\left(MRC^{\delta}\right)^{-1}$ generally also perform very poorly for estimating the inverse when $p = 50$ and close to $n = 78$.

Remark: We confirmed the good finite sample properties of our estimator under autocorrelated microstructure noise. In this specific case, the higher order dependence is considered by assuming that the factors in microstructure noise are the sum of an *iid* process and an *AR(1)* as in Aït-Sahalia, Mykland, and Zhang (2011). Table 7 and Table 8 of the

[16] The scaled Frobenius norm is defined by diving the usual Frobenius norm with $\sqrt{p}$. As discussed in Hautsch, Kyj, and Oomen (2012), this scaling allows for a more meaningful comparison across different values of p.

[17] In some of the simulated samples, both the new estimator and all of the competitors analyzed in the simulations are nearly singular, making their inverses numerically unstable. Importantly, however, this problem did not occur in any of the actual empirical applications discussed in the next section.

Table 1. Covolatility estimators, synchronous prices.

Signal-to-Noise ratio $\xi^2 = 0.005$, $K = 2$					
Number of assets: p=50					
	Covariance	Correlation	Inverse	Diag	Off-Diag
$\hat{\Sigma}$	2.492	1.299	4.567	21.09	377.6
	(0.729)	*(0.316)*	*(0.360)*		
MRker	2.645	1.472	5667	23.88	412.6
	(0.714)	*(0.170)*	*(93231)*		
MRC^{δ}	2.607	1.499	1050	22.09	385.3
	(0.605)	*(0.170)*	*(4936)*		
$\hat{\Sigma}_{comp}$	2.625	1.431	4.120	40.92	392.6
	(0.733)	*(0.172)*	*(0.694)*		
$PCA-PRV$	2.587	1.454	7.164	22.09	383.3
	(0.623)	*(0.173)*	*(10.32)*		
$POET$	5.663	2.922	402.6	209.0	1449
	(0.382)	*(0.229)*	*(22.68)*		
Number of assets: p=100					
$\hat{\Sigma}$	3.554	1.792	4.734	41.93	1500
	(1.261)	*(0.394)*	*(27.54)*		
MRker	3.865	2.124	NA	41.63	1701
	(0.927)	*(0.238)*	*NA*		
MRC^{δ}	3.811	2.161	NA	39.152	1589
	(0.771)	*(0.229)*	*NA*		
$\hat{\Sigma}_{comp}$	3.809	2.061	5.008	63.44	1639
	(0.942)	*(0.242)*	*(0.833)*		
$PCA-PRV$	3.732	2.067	6.038	39.15	1536
	(0.800)	*(0.236)*	*(10.29)*		
$POET$	7.653	4.371	596.0	364.6	5648
	(0.516)	*(0.334)*	*(130.2)*		
Number of assets: p=300					
$\hat{\Sigma}$	5.642	3.035	9.304	137.0	12669
	(2.120)	*(0.724)*	*(0.247)*		
MRker	6.313	3.707	NA	110.6	13623
	(1.546)	*(0.413)*	*NA*		
MRC^{δ}	6.204	3.761	NA	102.1	12685
	(1.250)	*(0.398)*	*NA*		
$\hat{\Sigma}_{comp}$	6.251	3.649	6.821	146.0	13365
	(1.557)	*(0.417)*	*(1.260)*		
$PCA-PRV$	5.991	3.508	5.586	102.123	11884
	(1.300)	*(0.415)*	*(1.877)*		
$POET$	12.17	7.681	NA	981.0	44653
	(0.853)	*(0.559)*	*NA*		
Number of assets: p=500					
$\hat{\Sigma}$	6.940	3.856	14.87	218.2	31870
	(2.678)	*(0.824)*	*(61.76)*		
MRker	7.937	4.765	NA	174.6	36191
	(1.905)	*(0.490)*	*NA*		
MRC^{δ}	7.915	4.871	NA	165.1	33994
	(1.417)	*(0.471)*	*NA*		
$\hat{\Sigma}_{comp}$	7.878	4.716	18.82	221.2	35703
	(1.911)	*(0.494)*	*(1.960)*		
$PCA-PRV$	7.598	4.601	15.68	165.1	31669
	(1.471)	*(0.477)*	*(11.87)*		
$POET$	14.94	10.08	NA	1498	111142
	(0.977)	*(0.717)*	*NA*		

Note: This table presents simulation results based on $1,000$ Monte Carlo replications for the simulation design described in Appendix A.4. We compute the Scaled Frobenius norm of estimation errors of: the integrated covolatility matrix, the integrated correlation matrix, the inverse of the integrated covolatility matrix, diagonal elements of the integrated covolatility matrix and off-diagonal elements.

Supplementary Appendix provides such simulation results.

3.3 Finite-sample simulations: asynchronous prices

The simulation results discussed above were based on synchronous prices. This section evaluates the performance of the same six estimators in the more realistic situation when the prices for different assets are not necessarily recorded at the same time and therefore first have to be synchronized.[18]

To accommodate this feature within the simulations, we augment the previously discussed factor setup by dividing the assets into three separate groups of differing observation frequencies. For assets in the first group, an observation is available on average every 30 seconds, in the second group every 90 seconds, and in the final third group every 150 seconds. All of the observation times for each of the individual assets within each of the three groups are drawn from Poisson distributions.

The results from these augmented simulations are reported in Table 2. To conserve space we only report the results for the case corresponding to $\xi^2 = 0.01$. As expected, all of the estimators perform worse in an absolute sense compared to the situation with synchronously observed prices in Table 1[19]. However, the relative performance of the different estimators is entirely in line with the previously discussed results in Table 2, underscoring the superior overall performance of the new $\widehat{\Sigma}$ estimator. The empirical application discussed in the next section also further corroborates this.

4 Empirical Application

Our empirical application is based on a large cross-section of individual stocks. It closely follows Lunde, Shephard, and Sheppard (2016) in assessing the performance of the different covolatility estimators by comparing the resulting risk minimizing portfolios.

[18] This issue is especially acute for the $MRker$ and MRC^{δ} estimators, which require that the synchronization process is applied to full p-dimensional price vector. By comparison, the computation of $\widehat{\Sigma}$ only needs for the prices to be synchronized on a pairwise basis, in turn resulting less of a loss of observations.

[19] The estimation error increases when incorporating the asynchronous sampling times because of the loss of data during the synchronization process. The error size is still acceptable. This is a finite sample property. Asymptotically, the effect of asynchronous price data on the covolatility matrix is small. The consistency of $\hat{\Sigma}$ is the consequence of the consistency of MRC under asynchronous sampling times. The theoretical assumptions about the irregularity and asynchronicity of the sampling times are the same as in Christensen, Kinnebrock, and Podolskij (2010).

Table 2. Covolatility estimators, asynchronous prices.

Signal-to-Noise ratio $\xi^2 = 0.01$, $K = 2$					
Number of assets: p=50					
	Covariance	Correlation	Inverse	Diag	Off-Diag
$\hat{\Sigma}$	4.914	2.374	3.788	49.88	1182.1
	(0.470)	*(0.173)*	*(7.871)*		
MRker	5.207	2.732	4960	44.54	1332
	(0.515)	*(0.205)*	*(13222)*		
MRC^{δ}	5.180	2.689	1594	43.08	1316
	(0.497)	*(0.200)*	*(6580)*		
$\hat{\Sigma}_{comp}$	5.047	2.646	4.430	42.70	1271
	(0.482)	*(0.179)*	*(0.258)*		
$PCA-PRV$	5.155	2.617	7.187	43.08	1292
	(0.498)	*(0.206)*	*(10.49)*		
$POET$	6.233	3.312	385.8	168.5	1841
	(0.512)	*(0.189)*	*(415.6)*		
Number of assets: p=100					
$\hat{\Sigma}$	5.430	3.141	4.041	94.92	2878
	(0.438)	*(0.209)*	*(12.03)*		
MRker	5.768	3.702	NA	71.747	3294.052
	(0.411)	*(0.182)*	*NA*		
MRC^{δ}	5.757	3.655	NA	66.73	3283
	(0.410)	*(0.178)*	*NA*		
$\hat{\Sigma}_{comp}$	5.619	3.565	4.622	60.35	3155
	(0.403)	*(0.153)*	*(0.325)*		
$PCA-PRV$	5.657	3.485	8.110	66.73	3150
	(0.422)	*(0.203)*	*(41.32)*		
$POET$	6.306	4.386	319.9	248.7	3834
	(0.428)	*(0.176)*	*(4848)*		
Number of assets: p=300					
$\hat{\Sigma}$	10.35	5.426	7.315	473.2	32356
	(0.825)	*(0.327)*	*(30.27)*		
MRker	11.15	6.595	NA	295.0	37950
	(0.809)	*(0.307)*	*NA*		
MRC^{δ}	11.11	6.493	NA	281.7	37699
	(0.797)	*(0.353)*	*NA*		
$\hat{\Sigma}_{comp}$	10.97	6.126	8.568	512.4	36548
	(0.912)	*(0.374)*	*(3.578)*		
$PCA-PRV$	10.87	6.094	8.225	281.7	35987
	(0.811)	*(0.346)*	*(8.580)*		
$POET$	12.57	7.468	NA	1000	46936
	(0.892)	*(0.321)*	*NA*		
Number of assets: p=500					
$\hat{\Sigma}$	12.46	6.842	9.780	357.2	78228
	(1.554)	*(0.351)*	*(80.61)*		
MRker	13.61	8.443	NA	416.6	93282
	(0.923)	*(0.335)*	*NA*		
MRC^{δ}	13.58	8.265	NA	396.6	92472
	(0.914)	*(0.381)*	*NA*		
$\hat{\Sigma}_{comp}$	13.45	8.159	10.85	754.2	89258
	(0.932)	*(0.298)*	*(5.365)*		
$PCA-PRV$	13.22	7.742	8.570	396.6	87690
	(0.930)	*(0.372)*	*(10.67)*		
$POET$	15.09	9.419	NA	1461	114468
	(1.039)	*(0.398)*	*NA*		

Note: This table presents simulation results based on $1,000$ Monte Carlo replications for the simulation design described in Appendix A.4. We compute the Scaled Frobenius norm of estimation errors of: the integrated covolatility matrix, the integrated correlation matrix, the inverse of the integrated covolatility matrix, diagonal elements of the integrated covolatility matrix and off-diagonal elements.

4.1 Data

We rely on intraday data from the TAQ database. Our original sample is comprised of all of the stocks included in the S&P 500 during the period spanning January 2007 to December 2011. Following Lunde, Shephard, and Sheppard (2016), we remove stocks that trade less than 195 times during a given day. We further clean the data following the procedures advocated in Barndorff-Nielsen, Hansen, Lunde, and Shephard (2011). All-in-all, this leaves us with a total of 384 stocks.

4.2 Risk minimization

Our comparison of the different covolatility estimators rely on their ability to minimize portfolio risks. Specifically, let $\hat{\Omega}_t$ denote a covolatilty estimate for day t. We will assume that $\hat{\Omega}_t$ follows a random walk, and use it as the forecast for the day $t+1$ covolatility matrix. Correspondingly, the portfolio weights $\hat{w}_{t+1}$ that minimize the day $t+1$ risk, subject to a cross exposure constraint, may be found by solving:

$$\begin{cases} Min \quad w_{t+1}^{'}\hat{\Omega}_t w_{t+1} \\ s.t. \quad w_{t+1}^{'}1 = 1 \quad and \quad \sum_{i=1}^{p} |w_{i,t+1}| \leq 1+2s. \end{cases} \tag{16}$$

The gross exposure parameter s represents the share of the stocks in the portfolio that can be held short.[20] Setting $s = 0$ restricts the portfolio to long positions only, while higher values of s allow for increasingly larger short positions. We will consider values of s ranging from 0 to 1. The gross exposure constraint also ensures that the optimization problem has a unique solution, even if $\hat{\Omega}_t$ is not positive semi-definite.[21] It also serves to moderate the impact of estimation errors in the covolatility matrices used in place of $\hat{\Omega}_t$ more generally (see, e.g., the discussion Fan, Li, and Yu (2012)).

We evaluate the performance of the different covolatility estimators, by calculating,

$$\hat{w}_{t+1}^{'} RCov_{t+1} \hat{w}_{t+1} \tag{17}$$

where $RCov_{t+1}$ denotes the day $t+1$ realized covariance matrix constructed from five-minute returns. This approach closely mirrors that of Lunde, Shephard, and Sheppard (2016). In addition to the results for the six specific covolatility estimators discussed above, we also report the results for a naive equally weighted portfolio $\hat{w}_{t+1} = \frac{1}{p}I_p$, as recently advocated by DeMiguel, Garlappi, and Uppal (2009).

[20] The classical Markowitz portfolio problem corresponds to $s = \infty$.

[21] This is especially useful for the $MRker$ and MRC^{δ} estimators, which are not guaranteed to be positive semi-definite.

Consistent with the simulation results for the asynchronous price series discussion above, we rely the refresh-time sampling approach of Barndorff-Nielsen, Hansen, Lunde, and Shephard (2011) to synchronize the data used in the actual implementation of the estimators.[22] The practical implementation of the new $\widehat{\Sigma}$ estimator further requires a choice for the number of systematic risk factors, K. We use the information criteria IC advocated by Bai and Ng (2002) for choosing the value of K that minimizes[23],

$$IC = log\left[\frac{1}{p}\sum_{j=1}^{\lfloor 1/\Delta \rfloor}(r^*_{j\Delta} - bf_{j\Delta})'(r^*_{j\Delta} - bf_{j\Delta})\right] + K \times g(p, \lfloor 1/\Delta \rfloor), \tag{18}$$

with the penalty function define by $g(p, \lfloor 1/\Delta \rfloor) = \frac{p+\lfloor 1/\Delta \rfloor}{p\lfloor 1/\Delta \rfloor} \times log\left[\frac{p\lfloor 1/\Delta \rfloor}{p+\lfloor 1/\Delta \rfloor}\right]$. In order to reduce the impact of market microstructure noise, IC is applied in the dataset sampled at the 5-minutes frequency. The number of factors chosen by this criteria range between one and four for each of the different days, with an average value of 3.277 over the full sample.

Table 3. Minimum variance portfolios

	s=0	s=0.01	s=0.05	s=0.1	s=0.15	s=0.20	s=0.25	s=0.5	s=1
$\hat{\Sigma}$	0.334	0.298	0.287	0.261	0.256	0.252	0.245	0.24	0.241
$\hat{\Sigma}_{comp}$	0.409	0.343	0.31	0.32	0.308	0.303	0.301	0.325	0.326
MRker	0.399	0.351	0.335	0.313	0.305	0.302	0.278	0.263	0.258
MRC^{δ}	0.412	0.368	0.352	0.334	0.331	0.323	0.362	0.343	0.319
EqualWeight	0.636	0.636	0.636	0.636	0.636	0.636	0.636	0.636	0.636
$PCA - PRV$	0.395	0.355	0.339	0.318	0.31	0.302	0.319	0.317	0.327
$POET$	0.401	0.338	0.311	0.287	0.277	0.266	0.289	0.278	0.286

Note: This table presents the ex-post variation of competing covolatility estimators, for different gross exposure levels, as described in equations (16) and (17).

Looking across the different rows of the table 3, the portfolios constructed based on the new $\hat{\Sigma}$ estimator systematically result in the lowest ex-post variation. This dominance holds true for all of the different values of the gross exposure constraint s. Meanwhile, the portfolios that rule out short positions reported in the first column ($s = 0$) unambiguously perform the worst. The differences observed across the other values of s are generally small and not

[22] Applying the synchronization to all of the stocks results in an average of 104.4 intraday observations.

[23] Since the number of stocks p and the intraday observations n diverge, we implement the Bai and Ng (2002) estimator of K using intraday observations sampled at 5 minutes frequency. There is an underlining assumption that the number of factors is asymptotically bounded by a fix positive number $kmax$.

always monotonic. All of the realized volatility-based portfolios also convincingly beat the $\frac{1}{p}$ naively diversified portfolios. In contrast to the simulation-based comparisons discussed above, where the $\hat{\Sigma}_{comp}$ systematically outperformed $MRker$ and MRC^{δ} that is not the case here.

5 Conclusion

We provide a new realized covolatility estimator that is guaranteed to be positive semidefinite in large dimensions and also works in the presence of market microstructure noise. The estimator relies on two separate factor structures: one of order $O_p(\sqrt{\Delta})$ for describing the cross-sectional variation in the systematic risks, and another of order $O_p(1)$ for describing the noise. The practical implementation of the estimator relies on traditional factor analysis together with already existing procedures for consistently and robustly estimating the different components of the covolatility matrix.

The convergence rate of the new estimator compares favorably to other recently developed procedures, including the adjusted modulated realized covariance estimator MRC^{δ} of Christensen, Kinnebrock, and Podolskij (2010), the multivariate kernel estimator $MRker$ of Barndorff-Nielsen, Hansen, Lunde, and Shephard (2011), and the composite realized kernel $\hat{\Sigma}_{comp}$ of Lunde, Shephard, and Sheppard (2016). Simulations confirm that the theoretical results derived under the assumption of synchronous prices observed over increasingly finer time intervals carry over to empirically realistic settings with a finite number of asynchronous intraday observations. Applying the new estimator in the construction of ex-ante minimum variance portfolios from a set comprised of several hundred individual equities also produces the lowest ex-post variation compared to other practically feasible competing covolatility estimators.

References

Aït-Sahalia, Y., Mykland, P. A., Zhang, L., 2011. Ultra high frequency volatility estimation with dependent microstructure noise. Journal of Econometrics 160, 160–175.

Ait-Sahalia, Y., Xiu, D., 2017. Using principal component analysis to estimate a high dimensional factor model with high-frequency data. Journal of Econometrics 201, 384–399.

Andersen, T., Bollerslev, T., 1998. Answering the skeptics: Yes standard volatility models do provide accurate forecasts. International Economic Review 39, 885–905.

Andersen, T., Bollerslev, T., Christoffersen, P., Diebold, F., 2006. Volatility and correlation forecasting. Handbook of Economic Forecasting (eds. G. Elliott, C.W.J. Granger and A. Timmermann).

Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221.

Bannouh, K., Martens, M., Oomen, R., van Dijk, D., 2012. Realized mixed frequency factor models for vast dimensional covariance estimation. ERIM Report Series 017.

Barndorff-Nielsen, O., Hansen, P.R.and Lunde, A., Shephard, N., 2008. Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise. Econometrica 76, 1481–1536.

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

Barndorff-Nielsen, O., Shephard, N., 2002. Econometric analysis of realised volatility and its use in estimating stochastic volatility models. Journal of the Royal Statistical Society, Series B 64, 253–280.

Chamberlain, G., Rothschild, M., 1983. Arbitrage, factor structure, and mean-variance analysis on large asset markets. Econometrica 51, 1281–1304.

Chen, N.-F., Roll, R., Ross, S. A., 1986. Economic forces and the stock market. Journal of Business 59, 383–403.

Christensen, K., Kinnebrock, S., Podolskij, M., 2010. Preaveraging estimators of the ex-post covariance matrix in noisy diffusion models with nonsynchronous data. Journal of Econometrics 159, 116–133.

Connor, G., Korajczyk, R., 1988. Risk and return in an equilibrium apt: Application of a new test methodology. Journal of Financial Economics 21, 255–289.

Dai, C., Lu, K., Xiu, D., 2018. Knowing factors or factor loadings, or neither? evaluating estimators of large covariance matrices with noisy and asynchronous data. Journal of Econometrics, forthcoming.

DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? Review of Financial Studies 22, 1916–1953.

Diebold, F., Strasser, G., 2013. On the correlation structure of microstructure noise: A financial economic approach. Review of Economic Studies 80, 1304–1337.

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

Fan, J., Fan, Y., Lv, J., 2008. High dimensional covariance matrix estimation using a factor model. Journal of Econometrics 147, 186–197.

Fan, J., Li, Y., Yu, K., 2012. Vast volatility matrix estimation using high-frequency data for portfolio selection. Journal of the American Statistical Association 107:497, 412–428.

Fan, J., Liao, Y., Mincheva, M., 2011. High-dimensional covariance matrix estimation in approximate factor models. Annals of Statistics 39, 3320–3356.

Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society B 75, 603–680.

Hansen, P., Lunde, A., 2006. Realized variance and market microstruture noise. Journal of Business and Economic Statistics 24, 127–161.

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

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

Hautsch, N., Podolskij, M., 2013. Pre-averaging based estimation of quadratic variation in the presence of noise and jumps: Theory, implementation, and empirical evidence. Journal of Business and Economic Statistics 31, 165–183.

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

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

Ledoit, O., Wolf, 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance 23, 603–621.

Lunde, A., Shephard, N., Sheppard, K., 2016. Econometric analysis of vast covariance matrices using composite realized kernels and their application to portfolio choice. Journal of Business and Economic Statistics 34, 504–518.

Pelger, M., 2018. Large-dimensional factor modeling based on high-frequency observations. Journal of Econometrics, forthcoming.

Ross, S. A., 1976. The arbitrage theory of capital asset pricing. Journal of Economic Theory 13, 341–360.

Sharpe, W. F., 1994. The Sharpe ratio. Journal of Portfolio Management 23, 49–58.

Stock, J. H., Watson, M. W., 2002. Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 97, 1167–1179.

Tao, M., Wang, Y., Chen, X., 2011. Fast convergence rates in estimating large volatility matrices using high-frequency financial data. Journal of the American Statistical Association 106.

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

Zhang, L., Mykland, P., 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.

Zheng, X., Li, Y., 2011. On the estimation of integrated covariance matrices of high dimensional diffusion processes. Annals of Statistics 39, 3121–3151.

Appendix

A.1 Alternative estimators

- The pre-averaging estimator is defined by:

$$PRV(r) = \frac{\sqrt{\Delta_n}}{\theta\psi_2} \sum_{i=0}^{\lfloor 1/\Delta_n \rfloor - k_n + 1} (\overline{Y}_i^n)^2 - \frac{\psi_1 \Delta_n}{2\theta^2 \psi_2} \sum_{i=1}^{\lfloor 1/\Delta_n \rfloor} r_i^2, \tag{19}$$

  where n is the number of observed returns; Δ_n is the time interval between two observations; $r_i = Y_{i\Delta_n} - Y_{(i-1)\Delta_n}$ is the i^{th} return computed from the observed price series Y; $\overline{Y}_i^n = \sum\limits_{j=1}^{k_n - 1} g(j/n) r_{i+j}$ is the i^{th} pre-averaging return and θ is a setting parameter to choose optimally such that $k_n\sqrt{\Delta_n} = \theta + o(\Delta_n^{1/4})$. Also $\phi_1(s) = \int\limits_s^1 g'(u)g'(u-s)du$, $\phi_2(s) = \int\limits_s^1 g(u)g(u-s)du$, and $\psi_{\text{i}} = \phi_i(0)$. The most important result of the pre-averaging approach is resumed in the asymptotic behavior established in Jacod, Li, Mykland, Podolskijc, and Vetter (2009).

$$\Delta_n^{-1/4}(PRV(r) - IV) \to N(0;\Gamma), \tag{20}$$

  with $\Gamma = \int\limits_0^1 \frac{4}{\psi_2^2}\left(\Phi_{22}\theta\sigma_t^4 + 2\Phi_{12}\frac{\sigma_t^2 V_\epsilon}{\theta} + \Phi_{11}\frac{V_\epsilon^2}{\theta^3}\right)dt$, V_ϵ is the noise variance, IV the true integrated volatility and $\Phi_{ij} = \int\limits_s^1 \phi_i(s)\phi_j(s)ds$.

- The realized kernel is defined by:

$$K(Y) = \sum_{h=-n}^{n} k(\frac{h}{H+1})\Gamma_h, \tag{21}$$

$$\Gamma_h = \sum_{j=h+1}^{n} y_j y_{j-h}^{'}, \text{ for } h > 0; \qquad \Gamma_h = \Gamma_{-h}^{'}, \text{ for } h < 0,$$

  where n is the number of synchronized returns per asset, Γ_h is the h^{th} realized autocovariance; $y_j = Y_j - Y_{j-1}$ for $j = 1, 2, ..., n$; with $Y_0 = \frac{1}{m}\sum\limits_{j=1}^{m} Y(\tau_{p,j})$; $Y_n = \frac{1}{m}\sum\limits_{j=1}^{m} Y(\tau_{p,p-m+j})$; $Y_j = Y(\tau_{p,j+m})$ for $j = 1, ..., n-1$; $\{\tau_{p,j}\}$ is the series of refresh time ; and k is a non-stochastic weighting function. The rate of convergence of this estimator is $n^{-1/5}$.

- The modulated realized covariance estimator is defined by:

$$MRC\,[Y]_n = \frac{n}{(n-k_n+2)}\frac{1}{\psi_2 k_n}\sum_{i=0}^{n-k_n+1} \bar{Y}_i^n \left(\bar{Y}_i^n\right)^{'} - \frac{\psi_1^{k_n}}{2n\theta^2\psi_2^{k_n}}\sum_{i=1}^{n}(r_i)(r_i)^{'}, \tag{22}$$

where Y is the observed price vector, n is the number of observed returns per asset, $\bar{Y}_i$ the i^{th} averaged return vector, r_i the i^{th} usual return vector defined as in (4), g a weighting function, $\psi_1^{k_n} = k_n \sum_{i=1}^{k_n-1} \left(g(\frac{i}{k_n}) - g(\frac{i-1}{k_n})\right)^2$, $\psi_2^{k_n} = \frac{1}{k_n} \sum_{i=1}^{k_n-1} g^2(\frac{i}{k_n})$, $k_n - 1$ the number of returns in each average, such that $\frac{k_n}{n^{1/2}} = \theta + o(n^{-1/4})$ and θ is a setting parameter. When the assets are not observed at the same time, the non-synchronicity issue is resolved using the refresh time method of Barndorff-Nielsen, Hansen, Lunde, and Shephard (2011).

- The adjusted modulated realized covariance estimator is defined by:

$$MRC\left[Y\right]_n^{\delta} = \frac{n}{(n-k_n+2)} \frac{1}{\psi_2 k_n} \sum_{i=0}^{k_n} \bar{Y}_i^n \left(\bar{Y}_i^n\right)', \tag{23}$$

where θ is such that $\frac{k_n}{n^{1/2+\delta}} = \theta + o(n^{-1/4+\delta/2})$. This estimator is consistent, with a sub-optimal rate of convergence of $n^{-1/5}$, and is positive semi-definite.

A.2 Estimation of rotated factors, $\tilde{f}$

Consider the following least squared problem where $f_{j\Delta}$ is chosen to minimize the scaled sum of squared values of the idiosyncratic component:

$$\begin{cases} \underset{f_{j\Delta},b}{Min} \; \frac{1}{p} \sum\limits_{j=1}^{\lfloor 1/\Delta \rfloor} (r_{j\Delta}^* - bf_{j\Delta})'(r_{j\Delta}^* - bf_{j\Delta}) \\ s.t \;\; \frac{1}{p} b'b = I_K \end{cases}$$

This is equivalent to:

$$\begin{cases} \underset{f_{j\Delta},b}{Min} \; \frac{1}{p} \sum\limits_{j=1}^{\lfloor 1/\Delta \rfloor} (r_{j\Delta}^* - bf_{j\Delta})'(r_{j\Delta}^* - bf_{j\Delta}) \\ s.t \;\; \forall k = 1, ..., K, \frac{1}{p} \underline{b}_k' \underline{b}_k = 1 \\ \forall k = 1, ..., K, \forall l = k+1, ..., K, \underline{b}_k' \underline{b}_l = 0 \end{cases}$$

where $\underline{b}_k$ corresponds to the column k of b. The Lagrangian of this problem is defined by

$$L = \frac{1}{p} \sum_{j=1}^{\lfloor 1/\Delta \rfloor} (r_{j\Delta}^* - bf_{j\Delta})'(r_{j\Delta}^* - bf_{j\Delta}) - \sum_{k=1}^{K} \lambda_k (\underline{b}_k' \underline{b}_k - p) - \sum_{k=1}^{K} \sum_{l=k+1}^{K} \mu_{kl} \underline{b}_k' \underline{b}_l$$

By deriving this Lagrangian with respect to $f_{k\Delta}$, we obtain

$$\begin{aligned}
\frac{\partial L}{\partial f_{k\Delta}} &= \frac{\partial}{\partial f_{k\Delta}}\left[\frac{1}{p}\sum_{j=1}^{\lfloor 1/\Delta\rfloor}(r^*_{j\Delta}-bf_{j\Delta})'(r^*_{j\Delta}-bf_{j\Delta})\right]\\
&= \frac{\partial}{\partial f_{k\Delta}}\left[\frac{1}{p}(r^*_{k\Delta}-bf_{k\Delta})'(r^*_{k\Delta}-bf_{k\Delta})\right]\\
&= \frac{\partial}{\partial f_{k\Delta}}\left[\frac{1}{p}(r^{*'}_{k\Delta}r^*_{k\Delta}-r^{*'}_{k\Delta}bf_{k\Delta}-f'_{k\Delta}b'r^*_{k\Delta}+f'_{k\Delta}b'bf_{k\Delta})\right]\\
&= (-b'r^*_{k\Delta}-b'r^*_{k\Delta}+b'bf_{k\Delta}+b'bf_{k\Delta})\\
&= (-2b'r^*_{k\Delta}+2b'bf_{k\Delta})
\end{aligned}$$

$$\begin{aligned}
\frac{\partial L}{\partial f_{k\Delta}}=0 &\iff (-2b'r^*_{k\Delta}+2b'bf_{k\Delta})=0\\
&\iff b'bf_{k\Delta}=b'r^*_{k\Delta}\\
&\iff f_{k\Delta}=(b'b)^{-1}b'r^*_{k\Delta}\\
&\iff f_{k\Delta}=(pI_K)^{-1}b'r^*_{k\Delta}\\
&\iff f_{k\Delta}=\frac{1}{p}b'r^*_{k\Delta}
\end{aligned}$$

Hence,

$$f_{k\Delta}=\frac{1}{p}b'r^*_{k\Delta},\quad \forall k=1,...,\lfloor 1/\Delta\rfloor \tag{24}$$

We are going now to concentrate the objective function by replacing $f_{j\Delta}$ by its formula given by (17).

$$\begin{aligned}
\frac{1}{p}\sum_{j=1}^{\lfloor 1/\Delta\rfloor}(r^*_{j\Delta}-bf_{j\Delta})'(r^*_{j\Delta}-bf_{j\Delta}) &= \frac{1}{p}\sum_{j=1}^{\lfloor 1/\Delta\rfloor}(r^*_{j\Delta}-b.\tfrac{1}{p}b'r^*_{j\Delta})'(r^*_{j\Delta}-b.\tfrac{1}{p}b'r^*_{j\Delta})\\
&= \frac{1}{p}\sum_{j=1}^{\lfloor 1/\Delta\rfloor}r^{*'}_{j\Delta}(I_p-\tfrac{1}{p}bb')'(I_p-\tfrac{1}{p}bb')r^*_{j\Delta}\\
&= \frac{1}{p}\sum_{j=1}^{\lfloor 1/\Delta\rfloor}r^{*'}_{j\Delta}r^*_{j\Delta}-\frac{1}{p}\sum_{j=1}^{\lfloor 1/\Delta\rfloor}r^{*'}_{j\Delta}bb'r^*_{j\Delta}\\
&= \frac{1}{p}\sum_{j=1}^{\lfloor 1/\Delta\rfloor}r^{*'}_{j\Delta}r^*_{j\Delta}-\frac{1}{p}\sum_{j=1}^{\lfloor 1/\Delta\rfloor}\sum_{k=1}^{K}r^{*'}_{j\Delta}\underline{b}_k\underline{b}'_kr^*_{j\Delta}\\
&= \frac{1}{p}\sum_{j=1}^{\lfloor 1/\Delta\rfloor}r^{*'}_{j\Delta}r^*_{j\Delta}-\frac{1}{p}\sum_{k=1}^{K}\sum_{j=1}^{\lfloor 1/\Delta\rfloor}r^{*'}_{j\Delta}\underline{b}_k\underline{b}'_kr^*_{j\Delta}\\
&= \frac{1}{p}\sum_{j=1}^{\lfloor 1/\Delta\rfloor}r^{*'}_{j\Delta}r^*_{j\Delta}-\frac{1}{p}\sum_{k=1}^{K}\sum_{j=1}^{\lfloor 1/\Delta\rfloor}\left(r^{*'}_{j\Delta}\underline{b}_k\right)\left(\underline{b}'_kr^*_{j\Delta}\right)\\
&= \frac{1}{p}\sum_{j=1}^{\lfloor 1/\Delta\rfloor}r^{*'}_{j\Delta}r^*_{j\Delta}-\frac{1}{p}\sum_{k=1}^{K}\sum_{j=1}^{\lfloor 1/\Delta\rfloor}\left(\underline{b}'_kr^*_{j\Delta}\right)\left(r^{'*}_{j\Delta}\underline{b}_k\right)\\
&= \frac{1}{p}\sum_{j=1}^{\lfloor 1/\Delta\rfloor}r^{*'}_{j\Delta}r^*_{j\Delta}-\frac{1}{p}\sum_{k=1}^{K}\underline{b}'_k\left(\sum_{j=1}^{\lfloor 1/\Delta\rfloor}r^*_{j\Delta}r^{'*}_{j\Delta}\right)\underline{b}_k
\end{aligned}$$

From the last equality, we deduce that the optimal $b=(\underline{b}_1,...,\underline{b}_K)$ is the solution of the following problem

$$\begin{cases}
\underset{\underline{b}_1,\dots,\underline{b}_K}{Max}\ \ \frac{1}{p}\sum_{k=1}^{K}\underline{b}'_k\left(\sum_{j=1}^{\lfloor 1/\Delta\rfloor}r^*_{j\Delta}r^{'*}_{j\Delta}\right)\underline{b}_k\\
s.t\ \ \forall k=1,...,K,\frac{1}{p}\underline{b}'_k\underline{b}_k=1\\
\forall k=1,...,K,\forall l=k+1,...,K,\underline{b}'_k\underline{b}_l=0
\end{cases}$$

The problem above is equivalent to resolve K optimization problems defining by: $\forall k \in \{1, ..., K\}$:

$$\begin{cases} \underset{\underline{b}_k}{Max} \quad \frac{1}{p}\underline{b}_k' \left(\sum\limits_{j=1}^{\lfloor 1/\Delta \rfloor} r_{j\Delta}^* r_{j\Delta}^{'*} \right) \underline{b}_k \\ s.t \;\; \frac{1}{p}\underline{b}_k'\underline{b}_k = 1 \\ \forall l \neq k, \underline{b}_k'\underline{b}_l = 0 \end{cases} \tag{25}$$

The Lagrangian of the above problem has the following form

$$L \;\; = \;\; \frac{1}{p}\underline{b}_k' \left(\sum\limits_{j=1}^{\lfloor 1/\Delta \rfloor} r_{j\Delta}^* r_{j\Delta}^{'*} \right) \underline{b}_k - \lambda_k \left(\frac{1}{p}\underline{b}_k'\underline{b}_k - 1 \right) - \sum\limits_{l \neq k}^{K} \mu_{kl}\underline{b}_k'\underline{b}_l$$

By resolving for $\underline{b}_k$

$$\frac{\partial L}{\partial \underline{b}_k} \;\; = \;\; \frac{2}{p} \sum\limits_{j=1}^{\lfloor 1/\Delta \rfloor} \left[r_{j\Delta}^* r_{j\Delta}^{*'} \right] \underline{b}_k - \frac{2\lambda_k}{p}\underline{b}_k - \sum\limits_{l \neq k} \mu_{kl}\underline{b}_l$$

$$\frac{\partial L}{\partial b} = 0 \Longleftrightarrow \frac{2}{p} \sum\limits_{j=1}^{\lfloor 1/\Delta \rfloor} \left[r_{j\Delta}^* r_{j\Delta}^{*'} \right] \underline{b}_k - \frac{2\lambda_k}{p}\underline{b}_k - \sum\limits_{l \neq k} \mu_{kl}\underline{b}_l = 0$$

By a left multiplication by $\underline{b}_m'$ ($\forall m \neq k$)

$$\begin{aligned} & \frac{2}{p}\underline{b}_m' \sum\limits_{j=1}^{\lfloor 1/\Delta \rfloor} \left[r_{j\Delta}^* r_{j\Delta}^{*'} \right] \underline{b}_k - \frac{2\lambda_k}{p}\underline{b}_m'\underline{b}_k - \sum\limits_{l \neq k} \mu_{kl}\underline{b}_m'\underline{b}_l = 0 \\ \Leftrightarrow \;\; & \frac{2}{p} \sum\limits_{j=1}^{\lfloor 1/\Delta \rfloor} \underline{b}_m' \left[r_{j\Delta}^* r_{j\Delta}^{*'} \right] \underline{b}_k - \frac{2\lambda_k}{p}\underline{b}_m'\underline{b}_k - \mu_{km}\underline{b}_m'\underline{b}_m = 0 \\ \Leftrightarrow \;\; & \mu_{km} = 0 \end{aligned}$$

The third equation comes from the uncorrelation assumption of factors and the identification constraint on loadings. Hence, $\forall m \neq k$, $\mu_{km} = 0$. We deduce that

$$\frac{2}{p} \sum\limits_{j=1}^{\lfloor 1/\Delta \rfloor} \left[r_{j\Delta}^* r_{j\Delta}^{*'} \right] \underline{b}_k - \frac{2\lambda_k}{p}\underline{b}_k = 0$$

This is equivalent to

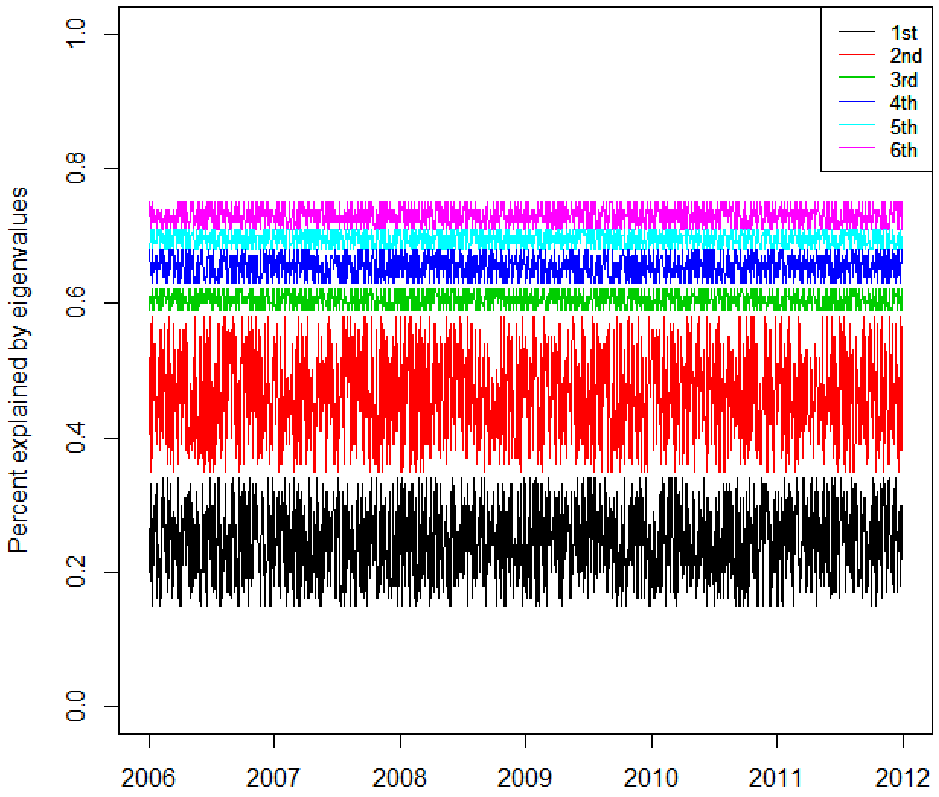
$$\sum\limits_{j=1}^{\lfloor 1/\Delta \rfloor} \left[r_{j\Delta}^* r_{j\Delta}^{*'} \right] \underline{b}_k - \lambda_k\underline{b}_k = 0$$

It follows that $\underline{b}_k$ is an eigenvector associated to the matrix $\sum\limits_{j=1}^{\lfloor 1/\Delta \rfloor} \left[r_{j\Delta}^* r_{j\Delta}^{*'} \right]$.

A.3 Factor structure in the noise

In order to underscore the empirical relevance of factor structures in the market microstructure noise component, we consider a sample of 384 stocks (as further described in Section 4) for all trading days from 2006 to 2011. For each trading days, we compute the realized covariance matrix and we divide it by $2n$, where n is the number of intraday transaction times after synchronization. By doing so, we get an estimator of the covolatility of the microstructure noise. The next step consists on a spectral decomposition of the obtained matrix. The following figure plots the ratio of the sum of the largest eigenvalues (the biggest eigenvalue, the first two biggest eigenvalues, the first three biggest eigenvalues, until the first six biggest eigenvalues) to the total sum of eigenvalues: these ratios can been interpreted as the part of the total variability explained by the considered factors (the first factor, the first two factors, until the first six factors).

Figure A.1. Ratio of largest eigenvalues relative to the total variation



Note: This figure plots the part of the total variability in microstructure noises explained by the considered factors (the first factor, the first two factors, until the first six factors).

Consistent with the idea of a factor structure in the market microstructure noise component, the figure shows that the four largest eigenvalues of the noise covolatility matrix explain more than 60% of the total variability for all of the trading days from 2006 to 2011.

A.4 Simulation design

Our simulation design replicates a two factor model in which the prices are observed with noise.

- The loading factors b is generated such that elements of the k^{th} column $\underline{b}_k$, for $k = 1, ..., K$, follow a normal law with mean 0 and standard deviation 1: $\underline{b}_{ik} \sim N(0, 1)$, $\forall$ $i = 1, ..., p$.
- The two factor components in the frictionless return representation are generated by the following model:[24]
  - Factor 1

    $$f_{1t} = \sigma_{f1t} dB_{1t}$$

    with B_{1t} a brownian motion and σ_{f1t} generated by a $GARCH$ diffusion model as in Andersen and Bollerslev (1998),

    $$d\sigma^2_{f1t} = \kappa_{f1}\left(\theta_{f1} - \sigma^2_{f1t}\right) dt + \lambda_{f1}\sigma^2_{f1t} dW_{1t}$$

    with $Corr(W_{1t}, B_{1t}) = -0.5$, $\kappa_{f1} = 0.035$, $\theta_{f1} = 0.636$, $\phi_{f1} = 0.296$, $\lambda_{f1} = \sqrt{2\kappa_{f1}\phi_{f1}}$, $\sigma_{f10} = \theta_{f1}$
  - Factor 2

    $$f_{2t} = \sigma_{f2t} dB_{2t}$$

    with B_{2t} a brownian motion and σ_{f2t} generated by a $GARCH$ diffusion model as in Andersen and Bollerslev (1998),

    $$d\sigma^2_{f2t} = \kappa_{f2}\left(\theta_{f2} - \sigma^2_{f2t}\right) dt + \lambda_{f2}\sigma^2_{f2t} dW_{2t}$$

    with $Corr(W_{2t}, B_{2t}) = -0.5$, $\kappa_{f2} = 0.035$, $\theta_{f2} = 0.3$, $\phi_{f2} = 0.296$, $\lambda_{f2} = \sqrt{2\kappa_{f2}\phi_{f2}}$, $\sigma_{f20} = \theta_{f2}$
- The idiosyncratic error term in the factor representation is assumed to satisfy

  $$\varepsilon_{it} = \sigma_{it} dW^{\varepsilon}_{it}$$

  with W^{ε}_{it} a brownian motion such that $W^{\varepsilon}_{it} \perp W_{1t}, W_{2t}$ and $W^{\varepsilon}_{it} \perp B_{1t}, B_{2t}$, with the spot volatility generated by three different representative models:

[24] Recall that f_{kt} is assumed to be the return of some portfolio

  - For $1 \leq i \leq p/3$, the volatility of the idiosyncratic component is generated by a Nelson GARCH diffusion limit model as in Barndorff-Nielsen and Shephard (2002):

$$d(\sigma_{it}^2) = (0.1 - \sigma_{it}^2)\, dt + 0.2\sigma_{it}^2 dB_{it}^{\varepsilon},$$

    with $Corr(W_{it}^{\varepsilon}, B_{it}^{\varepsilon}) = -0.3$ and $B_{it}^{\varepsilon} \perp W_{1t}, W_{2t}$ and $B_{it}^{\varepsilon} \perp B_{1t}, B_{2t}$;

  - For $p/3 < i \leq 2p/3$, the volatility process is assumed to follow a geometric Ornstein-Uhlenbeck (OU) model as in Barndorff-Nielsen and Shephard (2002):

$$dlog(\sigma_{it}^2) = -0.6\,(0.157 + log(\sigma_{it}^2))\, dt + 0.25 dB_{it}^{\varepsilon},$$

    with $Corr(W_{it}^{\varepsilon}, B_{it}^{\varepsilon}) = -0.3$ and $B_{it}^{\varepsilon} \perp W_t$ and $B_{it}^{\varepsilon} \perp B_t$;

  - For $2p/3 < i \leq p$, the volatility follows a $GARCH$ diffusion model as in Andersen and Bollerslev (1998):

$$d\sigma_{it}^2 = \kappa_{\varepsilon}\,(\theta_{\varepsilon} - \sigma_{it}^2)\, dt + \gamma_{\varepsilon}\sigma_{it} dB_{it}^{\varepsilon},$$

    with $Corr(W_{it}^{\varepsilon}, B_{it}^{\varepsilon}) = -0.3$ and $B_{it}^{\varepsilon} \perp W_t$ and $B_{it}^{\varepsilon} \perp B_t$; $\kappa_{\varepsilon} = 0.035$, $\theta_{\varepsilon} = 0.636$, $\gamma_{\varepsilon} = 0.296$, $\sigma_{i0} = \theta_{\varepsilon}$

- The slope in the factor representation of the microstructure noise is such that: $c_i \sim N(1,1)$, $\forall i = 1, ..., p$;

- As in Barndorff-Nielsen, Hansen, and Shephard (2008), the variance of the microstructure noise of the asset i satisfies the equality: $Var(u_i) = \xi^2 \sqrt{\frac{1}{n}\sum_{t=1}^{n} \sigma_{it}^4}$, with ξ^2 the noise-to-signal ratio which takes values in $\{0.001, 0.005, 0.01\}$ and σ_{it} the spot volatility of the true price process of asset i at time t.

- The variance of the idiosyncratic component η_{it} in the factor representation of the microstructure noise is assumed to have a fraction $1/n^{1.1}$ of the total variance $Var(u_i)$. Then, the variance of the factor term in this representation is given by: $\sigma_g^2 = \frac{\left(\overline{Var(u)} - \overline{\sigma_\eta^2}\right)}{\bar{C_p^2}}$, with $\bar{C_p^2} = \frac{1}{p}\sum_{i=1}^{p} c_i^2$.

- g_t and η_{it} are such that: $g_t \sim N(0, \sigma_g^2)$ and $\eta_{it} \sim N(0, \frac{1}{n^{1.1}} Var(u_i))$.

A.5 Estimation of W

In order to confirm that the eigenvectors of $MRker$ provide reliable estimates for W, we simulate daily efficient price vectors of dimension $p \in \{50, 100, 300\}$. We consider three different levels of microstructure noise: low, median and high with noise-to-signal ratio equal to 0.001, 0.01 and 0.1, respectively. Prices are generated by the same two factor simulation design describe in Appendix A.4. We compute the true covolatility matrix $MRker$ for each price path, and derive their spectral decompositions. The following figures illustrate the results for each of the different noise levels.

Figure A.2. Eigenvectors estimation using the multirealized kernel $MRker$: low noise

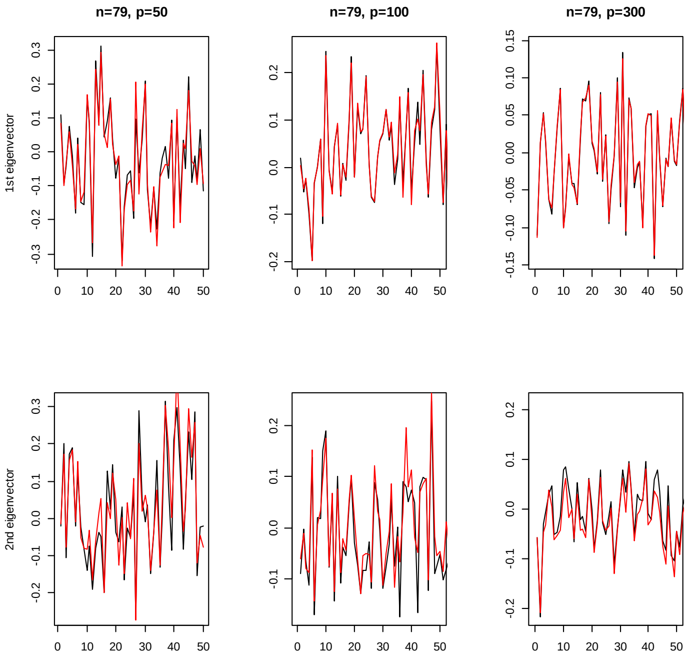


Figure A.3. Eigenvectors estimation using the multirealized kernel $MRker$: medium noise

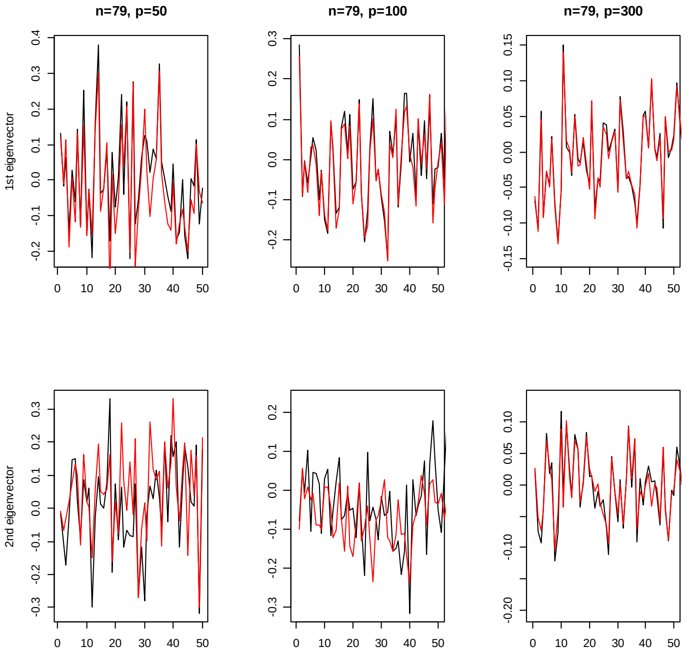
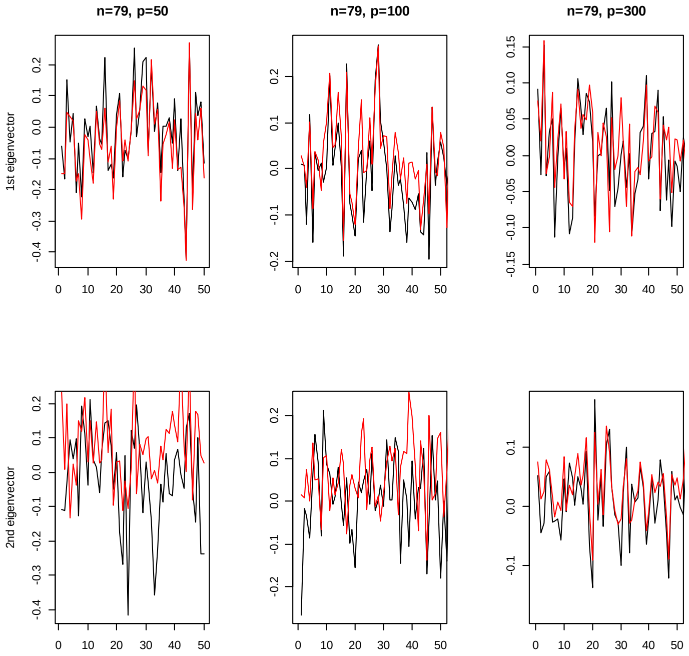


Figure A.4. Eigenvectors estimation using the multirealized kernel $MRker$: high noise



As is evident from the figures, the first two eigenvectors of the latent covolatility matrix are well estimated by the eigenvectors of the $MRker$ matrix. For low noise levels the two are almost indistinguishable, but there is also a close coherence for the high noise case.