12-2018

Forecasting large covariance matrix with high-frequency data: A factor correlation matrix approach

Yingjie DONG
University of International Business and Economics

Yiu Kuen TSE
Singapore Management University, yktse@smu.edu.sg

Citation

DONG, Yingjie and TSE, Yiu Kuen. Forecasting large covariance matrix with high-frequency data: A factor correlation matrix approach. (2018). 1-21.
Available at: https://ink.library.smu.edu.sg/soe_research/2270

Forecasting Large Covariance Matrix with High-Frequency Data: A Factor Correlation Matrix Approach

Yingjie Dong
Business School, University of International Business and Economics, Beijing

Yiu-Kuen Tse
School of Economics, Singapore Management University, Singapore

October 2018

Abstract: We propose a factor correlation matrix approach to forecast large covariance matrix of asset returns using high-frequency data. We apply shrinkage method to estimate large correlation matrix and adopt principal component method to model the underlying latent factors. A vector autoregressive model is used to forecast the latent factors and hence the large correlation matrix. The realized variances are separately forecasted using the Heterogeneous Autoregressive model. The forecasted variances and correlations are then combined to forecast large covariance matrix. We conduct Monte Carlo studies to compare the finite sample performance of several methods of forecasting large covariance matrix. Our proposed method is found to perform better in reporting smaller forecast errors. Empirical application to a portfolio of 100 NYSE and NASDAQ stocks shows that our method provides lower out-of-sample realized variance in selecting global minimum variance portfolio. It also provides higher information ratio for Markowitz portfolios.

JEL Codes: C410, G120

Keywords: Large correlation matrix, Nonlinear shrinkage, Dimension reduction, Eigenanalysis, Factor model, High-Frequency data

Acknowledgment: Yingjie Dong acknowledges partial support provided by the Project 71803021 of NSFC, and the Fundamental Research Funds for the Central Universities in University of International Business and Economics (UIBE) (CXTD8-01).

Corresponding Author: Yiu-Kuen Tse, School of Economics, Singapore Management University, Singapore 178903, email: yktse@smu.edu.sg.

1 Introduction

Modeling time varying covariance matrix of asset returns plays a crucial role in modern financial risk management and asset allocation. Multivariate GARCH (MGARCH) models, which are derived from the ARCH/GARCH family, are useful tools to deal with this problem. MGARCH models include the constant-correlation MGARCH (CC-GARCH) model of Bollerslev (1990), the BEKK model of Engle and Kroner (1995), the Dynamic Conditional Correlation (DCC) model of Engle (2002), and the Time-Varying Correlation model of Tse and Tsui (2002). MGARCH models, however, are usually applied to low dimension portfolios and problems arise when the number of assets are large. These include biases in large covariance matrix estimates, as well as computational feasibility of the model. Aielli (2013) proposes a consistent corrected DCC (cDCC) method for high dimension portfolios. Pakel *et al.* (2014) propose a composite quasi-likelihood estimate to tackle the computational issue. The recent work by Engle *et al.* (2017) applies the nonlinear shrinkage method to the DCC model to improve the estimation results.

MGARCH models are typically applied to daily data. With the availability of high-frequency intraday data, researchers can model and forecast the variance and covariance of asset returns using tick-by-tick transaction or quotation data. A naive estimator for high-frequency data can be obtained by calculating each diagonal/off-diagonal element of the covariance matrix using the realized variance/covariance estimates. Johnstone (2001) and Johnstone and Lu (2009), among others, point out that as the size of the portfolio covariance matrix goes to infinity, this naive estimator is inconsistent and the eigenvalues and eigenvectors of the estimated covariance matrix may deviate substantially from the true values. To solve this problem, banding and thresholding techniques are proposed to yield consistent large covariance matrices. The works of Bickel and Levina (2008a), Bickel and Levina (2008b), Wang and Zou (2010), and Cai and Liu (2011), among others, address this issue. Aït-Sahalia and Xiu (2017) use principle component method to estimate large covariance matrices. Ledoit and Wolf (2003), (2004) and (2017) propose to calculate large covariance matrix using the shrinkage method. Compared against the MGARCH family of models, these methods can deal with the curse of dimensionality quite successfully. However, they do not assume any underlying dynamic structure of the covariance matrices and hence may have drawbacks for forecasting.

The main focus of this paper is to propose a method to estimate and forecast large covariance matrix. First, we modify the latent factor model of Tao *et al.* (2011) and apply it to correlation matrix. We assume that the dynamic high-dimension correlation matrix is driven by a low-dimension latent process, and this latent component can be estimated via principal component analysis. We model the dynamic structure of the latent correlation factors by fitting a vector autoregressive (VAR) model. This captures the short-memory dynamics of the latent factors. Forecasts for these factors are then used to generate forecasts for the full correlation matrix. Second, we forecast the volatility of individual asset returns using the Heterogeneous Autoregressive (HAR) model of Corsi (2009). This model captures the long-memory properties of realized variances.[1] Finally, we combine the realized volatility forecasts with the large correlation matrix forecasts to obtain large covariance matrix forecasts. This method enables us to model the dynamics of the large covariance matrix by focusing on a reduced number of latent factors. It also utilizes rich information of high-frequency intraday transaction data in calculating large correlation matrix.[2]

Our method differs from that of Tao *et al.* (2011) in two aspects. First, Tao *et al.* (2011) model the covariance matrix process by assuming a short-memory dynamic structure of the vectorized factor covariance matrices. Instead, we model the correlation matrix process and the univariate volatility processes separately. We assume a short-memory structure for the vectorized latent factors and a long-memory structure for the volatility processes. Second, to obtain raw large covariance matrix for the eigen-analysis, Tao *et al.* (2011) use a truncation method on elements of the realized covariance matrix. Instead, we calculate the raw large correlation matrix by regulating the eigenvalues of the matrix using the nonlinear shrinkage method of Ledoit and Wolf (2017).[3]

We perform an empirical comparison of our method against the following methods: the factor covariance matrix method of Tao *et al.* (2011), the cDCC method of Aielli (2013), and the DCC-

[1] See Ding *et al.* (1993), Bollerslev and Mikkelsen (1996) and Baillie (1996), among others, for discussion of long-memory properties of volatility.

[2] See Andersen *et al.* (2013) for discussion of the advantages of high-frequency volatility estimates over traditional ARCH/GARCH family estimates.

[3] We thank Ledoit and Wolf for providing the codes (www.econ.uzh.ch/en/people/faculty/wolf/publications.html).

shrinkage method of Engle *et al.* (2017). Our method performs the best in reporting smaller forecast errors in our Monte Carlo simulation study. Also, it has better performance in terms of out-of-sample portfolio allocation for constructing both the global minimum variance (GMV) portfolio and the Markowitz portfolio with momentum signal.

The plan of the rest of this paper is as follows. In Section 2, we describe the construction of our factor correlation matrix approach. Some Monte Carlo results for the performance of our estimates are reported in Section 3. Section 4 describes an empirical investigation of the performance of different large covariance matrix forecasts in terms of out-of-sample asset allocation. Some concluding remarks are given in Section 5. A summary of the implementation procedure of our method is described in the Appendix.

2 Forecasting Large Covariance Matrix

2.1 Model Set-up

Let $\mathbf{X}_t = (X_{1t}, \cdots, X_{dt})'$ be an Itô process given by

$$d\mathbf{X}_t = \boldsymbol{\mu}_t dt + \boldsymbol{\sigma}_t' d\mathbf{B}_t, \qquad t = 1, \cdots, T, \tag{1}$$

where the stochastic processes $\mathbf{X}_t, \mathbf{B}_t, \boldsymbol{\mu}_t,$ and $\boldsymbol{\sigma}_t$ are defined on the filtered probability space denoted by $(\Omega, \mathcal{F}, \{\mathcal{F}_t, t \in [0,T]\},\ P)$. $\mathbf{B}_t$ is a d-dimensional standard Brownian motion with respect to $\mathcal{F}_t$, $\boldsymbol{\mu}_t$ is a d-dimensional drift vector, $\boldsymbol{\sigma}_t$ is a $d \times d$ matrix, and $\boldsymbol{\mu}_t$ and $\boldsymbol{\sigma}_t$ are assumed to be predictable processes with respect to the filtration $\mathcal{F}_t$. We assume d to be large, typically in the hundreds.

The *integrated covariance matrix* of $\mathbf{X}_t$ for the tth period (from time $t-1$ to time t) is defined as the $d \times d$ matrix

$$\Sigma_t = \int_{t-1}^{t} \boldsymbol{\sigma}_s' \boldsymbol{\sigma}_s \, ds, \qquad t = 1, \cdots, T, \tag{2}$$

and the *integrated correlation matrix* for the tth period is the $d \times d$ matrix

$$\Gamma_t = \widetilde{\Sigma}_t^{-\frac{1}{2}} \Sigma_t \widetilde{\Sigma}_t^{-\frac{1}{2}}, \quad t = 1, \cdots, T, \tag{3}$$

where $\widetilde{\Sigma}_t$ is obtained by replacing off-diagonal elements of Σ_t by zero.

We denote t_{il} as the lth trading time stamp of asset i and n_{it} as the total number of observed transactions in period t for asset i, where $i = 1, \cdots, d$ and $l = 1, \cdots, n_{it},$. At time stamp t_{il}, we observe the trading price $Y_{t_{il}}$, which is the contaminated price of the efficient price $X_{t_{il}}$ due to market microstructure noise. Thus,

$$Y_{t_{il}} = X_{t_{il}} + \epsilon_{t_{il}}, \quad i = 1, \cdots, d, \qquad l = 1, \cdots, n_{it}, \tag{4}$$

where $\epsilon_{t_{il}}$ are assumed to be iid microstructure noise with mean zero and (time invariant) variance η_i at the lth time stamp for stock i. We also assume that $\epsilon_{t_{il}}$ and $X_{t_{il}}$ are independent. Our objective is to estimate and forecast the integrated covariance matrix of $\mathbf{X}_t$ using high-frequency data.

Given an arbitrary positive definite matrix V, we define the correlation matrix transformation (CMT) of V, denoted by V^*, by

$$V^* = \widetilde{V}^{-\frac{1}{2}} V \widetilde{V}^{-\frac{1}{2}}, \tag{5}$$

where $\widetilde{V}$ is V with the off-diagonal elements replaced by zero. Note that V^* is a positive definite matrix with its diagonal elements being unity, and is thus a well-defined correlation matrix.[4]

2.2 Estimation of Large Correlation Matrix using High-Frequency Data

We adopt the matrix factor model of Tao *et al.* (2011) for high-frequency covariance matrix estimation and apply it to large correlation matrix. Specifically, we assume

$$\Gamma_t = \mathbf{A} \Gamma_t^f \mathbf{A}' + \Gamma_0, \tag{6}$$

where Γ_t^f, $t = 1, \cdots, T$, are $r \times r$ ($r \ll d$) positive definite matrices treated as a dynamical factor correlation process, $\mathbf{A}$ is a $d \times r$ factor loading matrix with $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$, and Γ_0 is a $d \times d$ positive definite time invariant matrix. Thus, to capture the dynamics of the $d \times d$ correlation matrices Γ_t, we control the parametric dimension by modeling the $r \times r$ latent factor matrices Γ_t^f.[5]

[4] Note that $\widetilde{V}^{-\frac{1}{2}}$ is the diagonal matrix with diagonal elements being the reciprocal of the square-root of the diagonal elements of V. Γ_t of equation (3) is the CMT of Σ_t in equation (2).

[5] Note that Γ_t^f need not be a well-defined correlation matrix. We assume, however, this is a latent factor matrix generating the large correlation matrix Γ_t.

We first estimate the covariance matrix Σ_t, from which the correlation matrix Γ_t can be calculated using the CMT. To estimate Σ_t, we adopt the nonlinear shrinkage method proposed by Ledoit and Wolf (2017).[6] The corresponding estimate of the correlation matrix will then be denoted by $\widehat{\Gamma}_t$.

To calculate the time invariant matrices $\mathbf{A}$ and Γ_0 in (6), we use the method of Tao *et al.* (2011) for covariance matrices and apply it to our model. Thus, we define

$$\widehat{\mathbf{S}} = \frac{1}{n}\sum_{t=1}^{n}(\widehat{\Gamma}_t - \widehat{\Gamma})^2, \tag{7}$$

where $\widehat{\Gamma} = \frac{1}{n}\sum_{t=1}^{n}\widehat{\Gamma}_t$. We use the r orthonormal eigenvectors corresponding to the r largest eigenvalues of $\widehat{\mathbf{S}}$ as the columns of the factor loading matrix $\mathbf{A}$, and denote this estimate by $\widehat{\mathbf{A}}$. The estimated factor matrix is then computed as

$$\widehat{\Gamma}_t^f = \widehat{\mathbf{A}}'\widehat{\Gamma}_t\widehat{\mathbf{A}}, \tag{8}$$

and the estimate of Γ_0 is

$$\widehat{\Gamma}_0 = \widehat{\Gamma} - \widehat{\mathbf{A}}\widehat{\mathbf{A}}'\widehat{\Gamma}\widehat{\mathbf{A}}\widehat{\mathbf{A}}'. \tag{9}$$

2.3 Forecasting Factor Correlation Matrix and Large Correlation Matrix

We use the Vector Autoregressive (VAR) Model to capture the short-run dynamics of the latent factors. For a $r \times r$ matrix Γ, let vech(Γ) denote the vector obtained by stacking together all elements on and below the diagonal of Γ. The VAR model for Γ_t^f is given by

$$\text{vech}(\Gamma_t^f) = \boldsymbol{\alpha}_0 + \sum_{j=1}^{q}\boldsymbol{\alpha}_j\text{vech}(\Gamma_{t-j}^f) + \boldsymbol{e}_t, \tag{10}$$

where $\boldsymbol{\alpha}_0$ is a $\tilde{r} \times 1$ vector with $\tilde{r} = r(r+1)/2$, and $\boldsymbol{\alpha}_j$, for $j = 1, \cdots, q$, are $\tilde{r} \times \tilde{r}$ square matrices. $\boldsymbol{e}_t$ is a $\tilde{r} \times 1$ vector white noise process with zero mean and finite fourth moments. Empirically, we fit equation (10) using $\widehat{\Gamma}_t^f$ as observed values of Γ_k^f, for $k = 1, \cdots, t-1$, to obtain the estimated coefficients $\widehat{\boldsymbol{\alpha}}_j$ for $j = 0, 1, \cdots, q$, and then $\widehat{\boldsymbol{\alpha}}_j$ are used to compute the out-of-sample forecasted latent factors matrix for the tth period.

[6] Engle, Ledoit and Wolf (2017) show that the nonlinear shrinkage method has superior performance when applied to the Dynamic Conditional Correlation (DCC) Model. An alternative method is the threshold multi-scale realized volatility matrix (TMSRVM) estimator proposed by Tao *et al.* (2013). This method will also be considered in our empirical application.

We denote the forecast of Γ_t^f conditional upon information up to time $t-1$ using the estimated VAR model by $\check{\Gamma}_t^f$. Then, the forecast of the $d \times d$ large correlation matrix $\check{\Gamma}_t$ is computed as

$$\check{\Gamma}_t = \widehat{\mathbf{A}}\check{\Gamma}_t^f\widehat{\mathbf{A}}' + \widehat{\Gamma}_0. \tag{11}$$

Note that $\check{\Gamma}_t$ may not be a well defined correlation matrix (positive definite matrix with unit diagonal elements). To resolve this problem we apply the CMT on $\check{\Gamma}_t$ to obtain $\check{\Gamma}_t^*$ as the forecasted correlation matrix. On the other hand, if $\check{\Gamma}_t$ is not positive definite, we project the matrix onto the space of positive definite matrices using the method of Fan *et al.* (2012).

2.4 Forecasting Realized Variance and Large Covariance Matrix

We further forecast the variance of individual assets separately using the Heterogenous Autoregressive (HAR) model of realized volatility proposed by Corsi (2009). We estimate the HAR equation as follows

$$RV_{i,t} = \omega_i + \alpha_i RV_{i,t-1} + \beta_i RV_{i,t-1}^w + \gamma_i RV_{i,t-1}^m, \qquad i = 1, \cdots, d, \tag{12}$$

where $RV_{i,t}$ is the calculated realized variance of asset i in period t, $RV_{i,t-1}^w = \frac{1}{5}\sum_{s=1}^{5} RV_{i,t-s}$, $RV_{i,t-1}^m = \frac{1}{22}\sum_{s=1}^{22} RV_{i,t-s}$. To compute $RV_{i,t}$, we use the subsampling method of Zhang *et al.* (2005) at 3-min intervals. The estimated models in equation (12) are used to forecast the realized variances. These forecasts are then collected to form the matrix $\check{\mathbf{D}}_t$, which is a $d \times d$ diagonal matrix with its ith diagonal element being the forecasted realized variance.

Finally, we compute the forecasted large covariance matrix as

$$\check{\Sigma}_t = \check{\mathbf{D}}_t^{\frac{1}{2}}\check{\Gamma}_t^*\check{\mathbf{D}}_t^{\frac{1}{2}}. \tag{13}$$

We call this forecast procedure M1, which is summarized in the Appendix.

3 Monte Carlo Simulation

We conduct a Monte Carlo study to investigate the finite sample performance of our proposed factor correlation matrix method.

3.1 Simulation Model Set-up

The following price generation process is assumed in our Monte Carlo experiment:

$$d\ \log \mathbf{X}_t = \varsigma_t' d\mathbf{W}_t, \tag{14}$$

$$\Sigma_t = \mathbf{D}_t^{\frac{1}{2}} \Gamma_t \mathbf{D}_t^{\frac{1}{2}}, \quad \text{where } \Sigma_t = \varsigma_t' \varsigma_t \tag{15}$$

$$d\sigma_{it}^2 = \kappa(\alpha_i - \sigma_{it}^2)dt + \gamma\sigma_{it}dB_{it}, \quad i = 1, \cdots, d, \tag{16}$$

where B_{it} is a 1-dimensional standard Brownian motion, $\mathbf{D}_t$ is a $d \times d$ diagonal matrix with its ith diagonal element being σ_{it}^2, and $\mathbf{W}_t$ is a d-dimensional standard Brownian motion which is uncorrelated with B_{it}, for $i = 1, \cdots, d$. We assume $\kappa = 5$, $\gamma = 0.4$. At each simulation run α_i are randomly drawn from the uniform distribution in the interval $[0.1, 0.2]$.

We assume that the correlation matrix Γ_t follows a factor model as in (6) and the number of factors r is 3. We generate the diagonal elements of Γ_t^f from three AR(1) processes with mean, AR coefficients and noise variance being (12, 0.55, 3), (5, 0.4, 1.2), and (3, 0.25, 0.7), respectively. Moreover, we assume the off-diagonal elements of Γ_t^f to be equal to 0. For $\mathbf{S}$ and Γ_0 in (6), we use empirically calculated values based on tick-by-tick transactions of 100 largest capitalization stocks from the NYSE and NASDAQ (as of 2015) in the period 2004 through 2016. We then simulate Γ_t from the factor model, with the loading matrix $\mathbf{A}$ being the eigenvectors corresponding to the three largest eigenvalues of $\widehat{\mathbf{S}}$. We update each individual price process every 10 sec and update the correlation matrix Γ_t daily.

We generate simulated transactions with initial value of $X_0 = \log(60)$ and σ_0 being randomly drawn from the uniform distribution in the interval $[0.1, 0.2]$. We add iid microstructure noise to the simulated price process with noise-to-signal ratio being 0.005%.[7] We let $d = 100$ and repeat the simulation procedure 100 times.

3.2 Monte Carlo Simulation Results

We calculate the forecasted covariance matrices using our proposed factor correlation matrix method M1. For comparison, we also vary M1 by modeling the factor covariance matrix pro-

[7] See Dong and Tse (2017b) for empirical estimates of the market microstructure noise variance.

cess instead of the factor correlation matrix process, and call this method M2. Similar to M1, M2 uses nonlinear shrinkage method to estimate the raw covariance matrix and a VAR model to capture the dynamics of the latent covariances (not correlations). For M1 and M2, we use 1-min returns for the nonlinear shrinkage estimate of Ledoit and Wolf (2017) to calculate the raw covariance estimate. M2 differs from M1 in that the HAR forecast for the realized variance of individual assets is not performed and the dynamic covariances are directly modelled using the VAR model. For further comparison, we also include the method of Tao *et al.* (2011) in our MC simulation and denote this estimate as M3. For M3, we calculate the $d \times d$ realized covariance matrices $\widehat{\Sigma}_t$ using the threshold multi-scale realized volatility matrix (TMSRVM) estimator of Tao *et al.* (2013) and treat them as raw estimates of Γ_t. We use the threshold values 0.95 and 0.98 for TMSRVM and intraday returns are also sampled at 1-min frequency. We then compute the forecasted covariance matrices by enforcing the factor covariance matrix method. Thus, M3 is the same as M2 except for the method in estimating the raw large covariance matrix. We fit a VAR(1) model for the estimated vectorized factor correlation/covariance matrices for M1, M2 and M3.

Since the true number of factors is 3, we select the number of factors r to be 2, 3 or 4 for M1, M2 and M3 in our computation. Finally, we also include the DCC model with nonlinear shrinkage of Engle, Ledoit, and Wolf (2017) for comparison. This method uses daily return data and is denoted as the DCC-shrinkage method.

We compare all estimates by investigating their performance in calculating the $d \times d$ covariance matrices in terms of the Frobenius norm errors and spectral norm errors, as well as errors of the estimated inverse covariance matrices. In Table 1 we report the out-of-sample norm errors of the forecasted covariance matrices as well as the norm errors of the inverse covariance matrices.

From Table 1 we can see that our proposed forecast M1 performs the best by reporting smaller Frobenius norm errors and spectral norm errors, both for the forecasted covariance matrices and the forecasted inverse of covariance matrices. M1, M2 and M3 produce similar results whether r is 2, 3 or 4. Comparing M1 and M2, we can see that the use of the factor correlation matrix model together with long-memory forecasts of realized variances outperforms the factor covariance matrix approach. Comparing M2 and M3, we can see that the use of nonlinear shrinkage method rather than the TMSRVM method produces better results. As expected, methods using high-frequency data performs better than the DCC-Shrinkage method using daily data only.

We also do some robustness checks by varying some settings of our Monte Carlo simulation studies. For M1 and M2, we implement the nonlinear shrinkage method using returns at 30-sec and 90-sec sampling frequencies. Results are quite similar to those of 1-min frequency. For M1, M2 and M3, we also fit a VAR(2) model for the estimated vectorized factor correlation/covariance matrices. Results are also very similar.

4 Empirical Comparison of Portfolio Selection

We compare the performance of various forecasts of variance matrix based on out-of-sample asset allocation. We select 100 largest market capitalization stocks (as of 2015) that are listed in NYSE or NASDAQ, with at least 200 trading days in any calendar year between 2004 and 2016 (3171 trading days). Tick-by-tick millisecond data are compiled and downloaded from the WRDS Daily TAQ (DTAQ) database.

4.1 Portfolio Selection Problems

We compare the performance of various variance forecast methods based on selection for the global minimum variance (GMV) portfolio and the Markowitz portfolio with momentum signal.

For the GMV portfolio, we choose the portfolio weights to minimize the portfolio variance by solving the following minimization problem

$$\min_{\boldsymbol{w}} \ \boldsymbol{w}'\Sigma_t\boldsymbol{w}, \quad \text{subject to } \boldsymbol{w}'\mathbf{1} = 1, \tag{17}$$

where $\boldsymbol{w}$ is the vector of portfolio weights, $\mathbf{1}$ is the vector of ones, and Σ_t is the portfolio covariance matrix at t. The analytical solution of this portfolio is

$$\boldsymbol{w} = \frac{\Sigma_t^{-1}\mathbf{1}}{\mathbf{1}'\Sigma_t^{-1}\mathbf{1}}. \tag{18}$$

We also investigate the problem of choosing portfolio weights such that the portfolio variance is minimized given a specific expected rate of return r_p. Thus, we solve the following minimization problem

$$\min_{\boldsymbol{w}} \ \boldsymbol{w}'\Sigma_t\boldsymbol{w}, \quad \text{subject to } \boldsymbol{w}'\mathbf{1} = 1 \text{ and } \boldsymbol{w}'\boldsymbol{\mu} = r_p, \tag{19}$$

where $\boldsymbol{\mu}$ is the expected rate of return of the constituents stocks. The analytical solution of this problem is

$$\boldsymbol{w} = \Sigma_t^{-1}[\boldsymbol{\mu}, \mathbf{1}] \begin{bmatrix} a & b \\ b & c \end{bmatrix}^{-1} \begin{bmatrix} r_p \\ 1 \end{bmatrix}, \tag{20}$$

where $a = \boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}$, $b = \boldsymbol{\mu}'\Sigma^{-1}\mathbf{1}$, and $c = \mathbf{1}'\Sigma^{-1}\mathbf{1}$.

Empirically, we replace Σ_t in equations (18) and (20) with the forecasted covariance matrices. We follow Engle *et al.* (2017) and treat the momentum factor of Jegadeesh and Titman (1993) as the required portfolio return r_p. We construct portfolios based on the calculated out-of-sample optimal weights, and then evaluate different methods by comparing the corresponding portfolio's realized variance and information ratio. The latter is defined as the portfolio return divided by the portfolio volatility and is particularly relevant as a performance measure for the Markowitz portfolio with momentum signal.[8]

4.2 Epps Effect and Sampling Frequency

Due to transaction asynchronicity, selection of sampling frequency is an important issue in high-frequency data analysis of multiple stocks. The well known Epps effect due to Epps (1979) highlights the problem that stock return correlation tends to go to zero when the tick data are sampled at higher frequencies. For illustration, we calculate the daily realized correlation of XOM and IBM in 2016. Figure 1 reports the average daily correlation when transactions are sampled at frequencies from 1 min to 30 min. We observe that the mean realized correlation increases as the sampled transactions become more sparse. The estimates tend to be stable after 15-min frequency. We check this phenomenon using other Dow Jones Industry Average (DJIA) stocks and obtain similar results. To mitigate the Epps effect, we sample transactions at 15-min frequency in our study.[9]

[8] Note that focusing on the out-of-sample standard deviation is now inappropriate due to estimation error in the momentum signal.

[9] Aït-Sahalia and Xiu (2017) suggest sampling transactions at frequencies between 15 min and 30 min. In contrast, Tao *et al.* (2011) use 5-min returns. We sample intraday transactions based on the Calendar Time Sampling scheme. For comparison of schemes of Calender Time, Tick Time and Business Time, see Oomen (2006) and Dong and Tse (2017a). We also add that we incorporate the close-to-open overnight returns in our sampled data.

4.3 Out-of-Sample Comparison of Portfolio Selection

We compare the performance of different covariance estimates in terms of their ability to select portfolios with lowest variance for the GMV portfolio and higher information ratio for the Markowitz portfolio with momentum signal. We calculate the optimal portfolio weights based on (18) and (20) using the out-of-sample forecasted covariance matrices. We then construct optimal portfolios of the next period based on the calculated optimal weights. To avoid an excessive amount of turnover and thus transaction costs, we update all portfolios at biweekly frequency, that is, every 10 consecutive trading days.[10] To calculate the volatility of the constructed portfolio, we use the RV method using portfolio intraday returns at 15-min frequency.

To select the number of factors r in equation (6), we calculate the shrinked biweekly correlation matrices $\widehat{\Gamma}_t$, for $t = 1, \cdots, 317$, and plot 100 sorted eigenvalues of $\widehat{\mathbf{S}}$ in Figure 2. We observe that the largest eigenvalue is substantially larger than others. We also calculate the eigenvalues for the shrinked biweekly covariance matrices and report the results in Figure 3. Similar observation is obtained, except that the magnitude of the eigenvalues declines more gradually.[11] Thus, to fit the factor correlation/covariance matrices we let the number of factors r be 3, 4 and 5 in our empirical implementation. The number of coefficients of the VAR model increases quickly as the lag parameter q or the number of factors r increases. Thus, we fit the diagonal-VAR(q) models for the vectorized factor matrices, with $q = 1$. We fit the cDCC model of Aielli (2013) and the shrinkage DCC method of Engle *et al.* (2017) using the biweekly close-to-close returns. For the cDCC model, we compute the DCC coefficients using the bivariate composite quasi-likelihood method of Pakel *et al.* (2014) based on contiguous pairs.[12]

We report the calculated mean portfolio realized variance and information ratio in Table 2, for both the GMV portfolio and the Markowitz portfolio with momentum signal problem. We observe that empirically M1 performs the best in reporting smaller mean portfolio realized variance and

[10] As there are 3171 trading days in our sample, we have a total of 317 periods. We start to calculate the out-of-sample portfolio weights at $t = 251$. To calculate the forecasted biweekly variance, we use a model similar to HAR and select daily RV, weekly RV and monthly RV as explanatory variables.

[11] Results for the TMSRVM covariance matrices are similar.

[12] We also fit these models using daily close-to-close returns. But poorer results are obtained.

larger portfolio information ratio. The cDCC estimates have rather poor performance. The factor correlation/covariance matrix models are robust with respect to the choice of r. The results further confirm our finding in the MC simulation study of the advantage of using the factor correlation matrix model set-up, as well as the use high-frequency data. We also achieve better results by using the nonlinear shrinkage estimate for the covariance matrices. Interestingly, although the DCC-Shrinkage estimate of Engle *et al.* (2017) does not utilize high-frequency data, it performs quite well compared against M2 and M3 for the Markowitz portfolio with momentum signal problem. This may further suggests the good performance of the nonlinear shrinkage estimate of Ledoit and Wolf (2017).[13]

5 Conclusions

We have proposed a factor correlation matrix approach to model and forecast large covariance matrices using high-frequency data. The dynamical structure of the large correlation matrices is assumed to be driven by a low-dimension latent process. We compute the low-dimension latent process using the principle component analysis on the shrinked correlation matrices and model the vectorized components using a short-memory VAR model. In contrast, the realized variance of individual assets is forecasted using the HAR model. We then forecast the large covariance matrix by combining the short-memory estimated correlation matrix and the long-memory realized volatilities. Our Monte Carlo simulation and empirical studies show that our method performs the best among alternative methods in the literature.

References

[1] Aielli, Gian Piero. “Dynamic conditional correlation: on properties and estimation.” Journal of Business & Economic Statistics 31, no. 3 (2013): 282-299.

[2] Aït-Sahalia, Yacine, and Jean Jacod. High-frequency financial econometrics. Princeton University Press, 2014.

[13] We perform some additional robustness checks for our empirical findings. We fit the diagonal-VAR(q) models for the vectorized factor matrices with $q = 2$. Results are similar to the case for $q = 1$.

[3] Aït-Sahalia, Yacine, and Dacheng Xiu. "Using principal component analysis to estimate a high dimensional factor model with high-frequency data." Journal of Econometrics (2017).

[4] Andersen, Torben G., Tim Bollerslev, and Francis X. Diebold. "Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility." The Review of Economics and Statistics 89, no. 4 (2007): 701-720.

[5] Andersen, Torben, Tim Bollerselv, Peter F Christoffersen and Francis X. Diebold. 2013. "Financial Risk Measurement for Financial Risk Management." In Handbook of the Economics of Finance, edited by G.Constantinides, M. Harris and R. Stulz, 1127-1220. North Holland.

[6] Baillie, Richard T. "Long memory processes and fractional integration in econometrics." Journal of Econometrics 73, no. 1 (1996): 5-59.

[7] Bickel, Peter J., and Elizaveta Levina. "Regularized estimation of large covariance matrices." The Annals of Statistics (2008a): 199-227.

[8] Bickel, Peter J., and Elizaveta Levina. "Covariance regularization by thresholding." The Annals of Statistics (2008b): 2577-2604.

[9] Bollerslev, Tim. "Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model." The Review of Economics and Statistics (1990): 498-505.

[10] Bollerslev, Tim, and Hans Ole Mikkelsen. "Modeling and pricing long memory in stock market volatility." Journal of Econometrics 73, no. 1 (1996): 151-184.

[11] Cai, Tony, and Weidong Liu. "Adaptive thresholding for sparse covariance matrix estimation." Journal of the American Statistical Association 106, no. 494 (2011): 672-684.

[12] Corsi, Fulvio. "A simple approximate long-memory model of realized volatility." Journal of Financial Econometrics 7, no. 2 (2009): 174-196.

[13] Ding, Zhuanxin, Clive WJ Granger, and Robert F. Engle. "A long memory property of stock market returns and a new model." Journal of Empirical Finance 1, no. 1 (1993): 83-106.

[14] Dong, Yingjie, and Yiu-Kuen Tse. "Business Time Sampling Scheme with Applications to Testing Semi-Martingale Hypothesis and Estimating Integrated Volatility." Econometrics 5, no. 4 (2017a): 51.

[15] Dong, Yingjie, and Yiu-Kuen Tse. "On estimating market microstructure noise variance." Economics Letters 150 (2017b): 59-62.

[16] Engle, Robert. "Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models." Journal of Business & Economic Statistics 20, no. 3 (2002): 339-350.

[17] Engle, Robert, and Bryan Kelly. "Dynamic equicorrelation." Journal of Business & Economic Statistics 30, no. 2 (2012): 212-228.

[18] Engle, Robert F., and Kenneth F. Kroner. "Multivariate simultaneous generalized ARCH." Econometric Theory 11, no. 1 (1995): 122-150.

[19] Engle, Robert F., Olivier Ledoit, and Michael Wolf. "Large dynamic covariance matrices." Journal of Business & Economic Statistics (2017): 1-13.

[20] Engle, Robert F., Olivier Ledoit, and Michael Wolf. "Large dynamic covariance matrices." Journal of Business & Economic Statistics (2017): 1-13.

[21] Epps, Thomas W. "Comovements in stock prices in the very short run." Journal of the American Statistical Association 74, no. 366a (1979): 291-298.

[22] Fan, Jianqing, Yingying Li, and Ke Yu. "Vast volatility matrix estimation using high-frequency data for portfolio selection." Journal of the American Statistical Association 107, no. 497 (2012): 412-428.

[23] Jegadeesh, Narasimhan, and Sheridan Titman. "Returns to buying winners and selling losers: Implications for stock market efficiency." The Journal of Finance 48.1 (1993): 65-91.

[24] Johnstone, Iain M. "On the distribution of the largest eigenvalue in principal components analysis." Annals of Statistics (2001): 295-327.

[25] Johnstone, Iain M., and Arthur Yu Lu. "On consistency and sparsity for principal components analysis in high dimensions." Journal of the American Statistical Association 104, no. 486 (2009): 682-693.

[26] Ledoit, Olivier, and Michael Wolf. "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection." Journal of Empirical Finance 10.5 (2003): 603-621.

[27] Ledoit, Olivier, and Michael Wolf. "A well-conditioned estimator for large-dimensional covariance matrices." Journal of Multivariate Analysis 88.2 (2004): 365-411.

[28] Ledoit, Olivier, and Michael Wolf. "Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks." The Review of Financial Studies 30.12 (2017): 4349-4388.

[29] Oomen, Roel C. A. "Properties of realized variance under alternative sampling schemes." Journal of Business & Economic Statistics 24, no. 2 (2006): 219-237.

[30] Pakel, Cavit, Neil Shephard, Kevin Sheppard, and Robert F. Engle. "Fitting vast dimensional time-varying covariance models." Manuscript NYU and revision of SSRN 1354497 (2014).

[31] Tao, Minjing, Yazhen Wang, and Harrison H. Zhou. "Optimal sparse volatility matrix estimation for high-dimensional Itô processes with measurement errors." The Annals of Statistics 41, no. 4 (2013): 1816-1864.

[32] Tao, Minjing, Yazhen Wang, Qiwei Yao, and Jian Zou. "Large volatility matrix inference via combining low-frequency and high-frequency approaches." Journal of the American Statistical Association 106, no. 495 (2011): 1025-1040.

[33] Tse, Yiu K., and Albert K. C. Tsui. "A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations." Journal of Business & Economic Statistics 20, no. 3 (2002): 351-362.

[34] Wang, Yazhen, and Jian Zou. "Vast volatility matrix estimation for high-frequency financial data." The Annals of Statistics 38, no. 2 (2010): 943-978.

[35] Zhang, Lan, Per A. Mykland, and Yacine Aït-Sahalia. "A tale of two time scales: Determining integrated volatility with noisy high-frequency data." Journal of the American Statistical Association 100, no. 472 (2005): 1394-1411.

Appendix

Suggested Procedure of Forecasting Large Covariance Matrix

We suggest a method of forecasting large covariance matrix of asset returns using high-frequency data. This method assumes long-memory property of realized volatility and short-memory property of correlation matrix. We call this method M1, and the forecasting steps are as follows.

(1) Compute the $d \times d$ covariance matrix of asset returns in the period from time $t-1$ to time t using 15-min intraday returns and close-to-open returns (overnight jumps) using the nonlinear shrinkage method of Ledoit and Wolf (2017). Denote this matrix by $\widehat{\Sigma}_t$ and compute the correlation matrix in this period, denoted by $\widehat{\Gamma}_t$, by applying the CMT to $\widehat{\Sigma}_t$.

(2) Compute

$$\widehat{\Gamma} = \frac{1}{n}\sum_{t=1}^{n}\widehat{\Gamma}_t,$$

and

$$\widehat{\mathbf{S}} = \frac{1}{n}\sum_{t=1}^{n}(\widehat{\Gamma}_t - \widehat{\Gamma})^2.$$

(3) Specify the number of latent factors r ($r \ll d$) of the correlation matrix. Compute the $d \times r$ matrix $\widehat{\mathbf{A}}$ as the r orthonormal eigenvectors of $\widehat{\mathbf{S}}$ corresponding to the r largest eigenvalues. Also, compute

$$\widehat{\Gamma}_0 = \widehat{\Gamma} - \widehat{\mathbf{A}}\widehat{\mathbf{A}}'\widehat{\Gamma}\widehat{\mathbf{A}}\widehat{\mathbf{A}}'. \tag{21}$$

and calculate the estimated latent factor matrix for period t as

$$\widehat{\Gamma}_t^f = \widehat{\mathbf{A}}'\widehat{\Gamma}_t\widehat{\mathbf{A}}, \quad t = 1, \cdots, T. \tag{22}$$

(4) Estimate the parameters $\boldsymbol{\alpha}_j$, for $j = 0, 1, \cdots, q$, in the VAR model

$$\text{vech}(\widehat{\Gamma}_t^f) = \boldsymbol{\alpha}_0 + \sum_{j=1}^{q}\boldsymbol{\alpha}_j\text{vech}(\widehat{\Gamma}_{t-j}^f) + \boldsymbol{e}_t, \tag{23}$$

where $\boldsymbol{\alpha}_0$ is a $\tilde{r} \times 1$ vector with $\tilde{r} = r(r+1)/2$, and $\boldsymbol{\alpha}_j$, for $j = 1, \cdots, q$, are $\tilde{r} \times \tilde{r}$ square matrices. Use the estimated coefficients $\widehat{\boldsymbol{\alpha}}_j$ to compute the forecasted $r \times r$ latent factor matrix conditional on information up to time $t-1$ and denote it by $\check{\Gamma}^f_t$.

(5) Forecast the $d \times d$ correlation matrix in period t conditional on information up to time $t-1$ using the factor model as

$$\check{\Gamma}_t = \widehat{\mathbf{A}} \check{\Gamma}^f_t \widehat{\mathbf{A}}' + \widehat{\Gamma}_0.$$

To ensure a well defined correlation matrix, apply the CMT to $\check{\Gamma}_t$ and estimate the correlation matrix of the d asset returns by

$$\check{\Gamma}^*_t = \widetilde{\check{\Gamma}}_t^{-\frac{1}{2}} \check{\Gamma}_t \widetilde{\check{\Gamma}}_t^{-\frac{1}{2}},$$

where $\widetilde{\check{\Gamma}}_t$ is $\check{\Gamma}_t$ with off-diagonal elements replaced by zero. If $\check{\Gamma}_t$ is not positive-definite, project the matrix onto the space of positive definite matrices using the method of Fan *et al.* (2012). This procedure ensure that $\check{\Gamma}^*_t$ is a well-defined correlation matrix.

(6) Forecast the realized volatilities using the HAR method as in Section 2.4. Denote $\check{\mathbf{D}}_t$ as the $d \times d$ diagonal matrix with its ith element being the forecasted realized variance of asset i in period t. Finally, compute the forecasted $d \times d$ covariance matrix as

$$\check{\Sigma}_t = \check{\mathbf{D}}_t^{\frac{1}{2}} \check{\Gamma}^*_t \check{\mathbf{D}}_t^{\frac{1}{2}}.$$

Table 1. Norms of errors of estimated covariance matrices and their inverses

			Error of covariance matrix		Error of inverse covariance matrix	
Method		r	Frobenius Norm ($\times 10^{-3}$)	Spectral norm ($\times 10^{-3}$)	Frobenius norm	Spectral norm
M1		2	2.5444	2.0188	22367	11944
		3	2.5432	2.0183	22368	11945
		4	2.5433	2.0180	22371	11945
M2		2	3.7947	2.7714	25509	13986
		3	3.5904	2.6057	25558	14002
		4	3.4367	2.4834	25558	13998
M3	95%	2	13.5865	13.0353	4.57×10^{8}	4.57×10^{8}
		3	13.5864	13.0351	5.31×10^{8}	5.31×10^{8}
		4	13.5863	13.0349	14.29×10^{8}	14.29×10^{8}
	98%	2	14.5765	13.8933	4.19×10^{8}	4.19×10^{8}
		3	14.5763	13.8930	5.38×10^{8}	5.38×10^{8}
		4	14.5761	13.8928	10.57×10^{8}	10.57×10^{8}
DCC-Shrinkage			6.9558	5.0044	48907	16576

Notes: M1 uses the factor correlation matrix method, where the raw large correlation matrices are calculated using CMT on the covariance matrices computed using nonlinear shrinkage method of Ledoit and Wolf (2017) and the volatilities are calculated using the HAR method of Corsi (2009). M2 calculates covariance matrices using the factor covariance matrix method, where the raw large covariances matrices are calculated using the nonlinear shrinkage method. M3 is the same as M2, except that the covariance matrices are calculated using the TMSRVM method of Tao *et al.* (2011). r is the number of low-dimension factors in the factor model. The DCC-Shrinkage method is due to Engle, Ledoit, and Wolf (2017), which applies nonlinear shrinkage to the DCC model.

Table 2. Estimated realized variance of constructed portfolios

		GMV		Markowitz portfolio with a signal	
Method	r	Volatility (%)	Information ratio	Volatility (%)	Information ratio
M1	3	9.7373	1.4246	9.8131	1.5702
	4	9.6986	1.5344	9.7853	1.6455
	5	9.6904	1.5157	9.7846	1.6292
M2	3	9.8345	0.9158	9.9533	1.0761
	4	10.1826	0.7693	10.2871	0.9157
	5	10.1832	0.7730	10.2874	0.9195
M3	3	10.9701	1.1198	11.0449	1.1279
	4	10.9777	1.1083	11.0519	1.1179
	5	10.9741	1.1298	11.0484	1.1379
DCC-Shrinkage		11.0258	1.1662	11.0845	1.1918
cDCC		20.0295	0.8213	27.1846	0.8643

Notes: The figures are the mean realized daily volatility (annualized standard deviation) and information ratio of the constructed portfolios. Out-of-sample optimal portfolio weights are calculated for the global minimum variance (GMV) portfolio and the Markowitz portfolio with momentum signal. M1 uses the factor correlation matrix method, where the raw large correlation matrices are calculated using CMT on the covariance matrices computed using nonlinear shrinkage method of Ledoit and Wolf (2017) and the volatilities are calculated using the HAR method of Corsi (2009). M2 calculates covariance matrices using the factor covariance matrix method, where the raw large covariances matrices are calculated using the nonlinear shrinkage method. M3 is the same as M2, except that the covariance matrices are calculated using the TMSRVM method of Tao *et al.* (2011). r is the number of low-dimension factors in the factor model. The DCC-Shrinkage method is due to Engle, Ledoit, and Wolf (2017), which applies nonlinear shrinkage to the DCC model. The cDCC model is due to Aielli (2013).

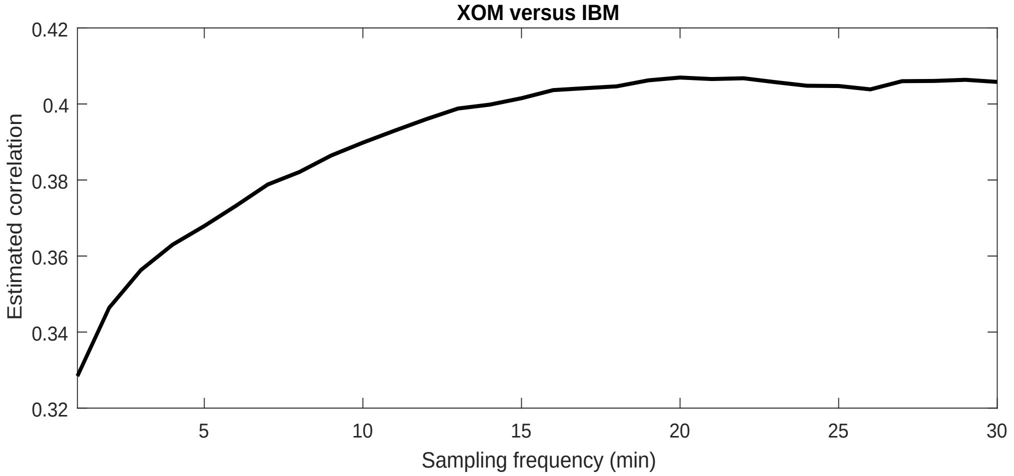

Figure 1: Averaged estimated daily realized correlation at different frequencies.

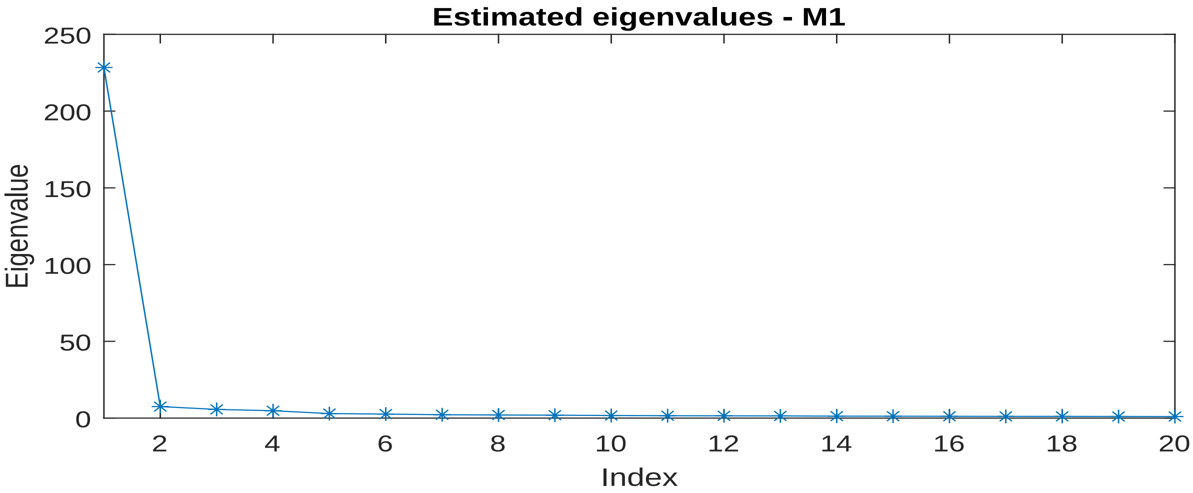

Figure 2: Estimated eigenvalues for the calculated correlation matrices.

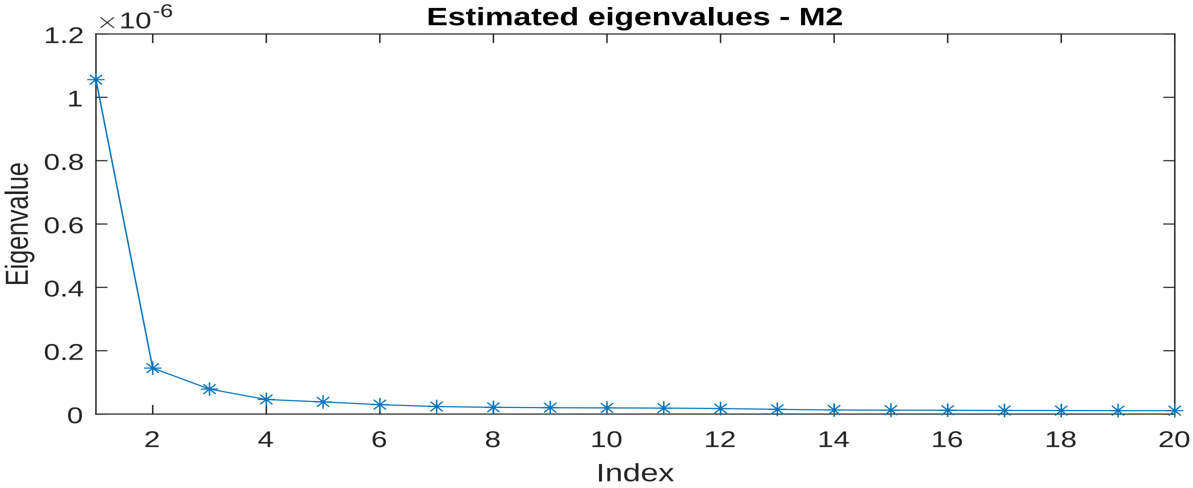

Figure 3: Estimated eigenvalues for the calculated covariance matrices.