axioms

MDPI

Article

Forecasting High-Dimensional Covariance Matrices Using High-Dimensional Principal Component Analysis

Hideto Shigemoto [1] **and Takayuki Morimoto** [2,*]

[1] Graduate School of Science and Technology, Kwansei Gakuin University, 2-1 Gakuen, Sanda 669-1337, Japan
[2] School of Science, Kwansei Gakuin University, 2-1 Gakuen, Sanda 669-1337, Japan
* Correspondence: morimot@kwansei.ac.jp

Abstract: We modify the recently proposed forecasting model of high-dimensional covariance matrices (HDCM) of asset returns using high-dimensional principal component analysis (PCA). It is well-known that when the sample size is smaller than the dimension, eigenvalues estimated by classical PCA have a bias. In particular, a very small number of eigenvalues are extremely large and they are called spiked eigenvalues. High-dimensional PCA gives eigenvalues which correct the biases of the spiked eigenvalues. This situation also happens in the financial field, especially in situations where high-frequency and high-dimensional data are handled. The research aims to estimate the HDCM of asset returns using high-dimensional PCA for the realized covariance matrix using the Nikkei 225 data, it estimates 5- and 10-min intraday asset-returns intervals. We construct time-series models for eigenvalues which are estimated by each PCA, and forecast HDCM. Our simulation analysis shows that the high-dimensional PCA has better estimation performance than classical PCA for the estimating integrated covariance matrix. In our empirical analysis, we show that we will be able to improve the forecasting performance using the high-dimensional PCA and make a portfolio with smaller variance.

Keywords: covariance forecasting; high-dimensional covariance; high-frequency data; principal component analysis; time series

check for updates

Citation: Shigemoto, H.; Morimoto, T. Forecasting High-Dimensional Covariance Matrices Using High-Dimensional Principal Component Analysis. *Axioms* **2022**, *11*, 692. https://doi.org/10.3390/axioms11120692

Academic Editor: Nikita Moiseev

Received: 21 October 2022
Accepted: 29 November 2022
Published: 3 December 2022

1. Introduction

Modeling and forecasting covariance matrices of asset returns have an essential role in portfolio allocations and risk management. For estimating and forecasting covariance matrix, a lot of papers are published on both low- and high-frequency data. Concerning the low-frequency data, the multivariate GARCH models [1], for example, BEKK-GARCH [2] and DCC-GARCH [3,4], are usually used to estimate and forecast the covariance matrix as latent. On the other hand, the availability of high-frequency data recently enabled the direct estimation of the covariance matrix, for example, the realized covariance matrix estimator [5], and the multivariate realized kernel estimator [6]. Additionally, some forecasting models such as the multivariate HAR [7], conditional autoregressive Wishart (CAW) [8], and realized DCC [9] models, use these covariance estimators to forecast them. However, when the dimensions increase, these covariance estimators and forecasting models have less accurate performance and suffer from an increase in the number of estimated parameters because of various reasons, such as the curse of dimensionality.

To solve these problems, the DCC-NL model which can overcome the curse of dimensionality using nonlinear shrinkage estimation is proposed [10]. To analyze the conditional high-dimensional covariance matrix (HDCM), recent studies using some multivariate GARCH models use the DCC-NL model instead of Tse and Tsui's and Engle's DCC-GARCH models [11–14]. Then, to solve the curse of dimensionality, many studies assume that the covariance matrix process or the price process follows a factor structure. Wang and Zou [15] propose a covariance estimator assuming that the integrated covariance matrix is sparse.

Considering a sparse covariance matrix allows only important elements to remain and also reduces the number of elements to be estimated. In addition, Tao et al. [16] introduce a covariance estimator which uses the matrix factor structure for an HDCM. We can obtain not only a consistent estimator of an HDCM but also a forecasted value using the vector autoregressive (VAR) model for a low-dimensional factor covariance matrix. Kim et al. [17] propose a threshold covariance estimator to regularize some realized covariance measures under the same assumption as [15]. Shen et al. [18] apply the method proposed by [16] to a realized covariance matrix and consider the CAW model instead of the VAR model for the factors. However, these studies assume sparsity in the integrated covariance matrix itself, which represents the target to be estimated. If there are some common factors across asset returns, the assumption that the integrated covariance is sparse becomes unrealistic because there are correlations among all pairs of assets through the common factors [19–22].

Fan et al. [19] propose the principal orthogonal complement thresholding (POET) method which assumes sparsity, not for the covariance matrix itself, but for the covariance matrix of the residual process, and estimates the latent factor using principal component analysis (PCA) to solve some problems. For high-frequency data, Fan et al. [20] assume the observable factor structure inspired by [23], and propose the covariance estimator under the assumption that the covariance matrix of the residual process is sparse. To estimate the latent factor structure, Aït-Sahalia and Xiu [24] impose sparsity on the residual covariance matrix and apply POET to high-frequency data using PCA to estimate an HDCM. They show that even when the factor is latent, if the residual covariance matrix is sufficiently sparse, the factor part can be estimated by PCA on the consistent estimator of the integrated covariance matrix, like the realized covariance matrix. In addition, they show that their estimator is a consistent estimator even if the interval of intraday return is $\Delta \to 0$ and the dimension is $d \to \infty$. In addition, Dai et al. [25] also propose an estimation method of the sparse residual covariance matrix using thresholding, and a high-dimensional covariance estimator using the POET estimator. The difference between [24] and [25] is the sparse structure. While Aït-Sahalia and Xiu [24] assume the block-diagonalize structure instead of thresholding, Dai et al. [25] do not assume the block-diagonalize structure but set a more general assumption, and use soft-, hard-, and adaptive-lasso (AL) [26], and smoothly clipped absolute deviation (SCAD) [27] thresholding. For the sparse estimation of the residual covariance matrix, Cai and Liu [28] propose the adaptive and hard thresholding method, but this method cannot guarantee the positive definiteness under the finite sample [29], and also has less performance than [25]. Brownlees et al. [21] propose the realized network estimator using the graphical lasso to estimate the precision matrix. Jian et al. [29] build time-series models for estimated eigenvalues based on the estimator of [24], and forecast the HDCM. In addition, they propose the regularized method to guarantee the positive definiteness.

The classical PCA, which is used by these models, creates a bias under $d > M$; d is the dimension of a covariance matrix and M is the sample size [15,24,30,31]. Wang and Fan [31] characterize the asymptotic distribution of empirical eigenvalues under the i.i.d setting and $d > M$. They also propose the shrinkage POET (SPOET) method based on their asymptotic distribution. The SPOET method corrects the biases of eigenvalues estimated by classical PCA.

In this paper, we estimate the HDCM under the factor structure for the high-frequency data, and create the forecasting models using its eigenvalues. It is well-known that the realized covariance matrix is a consistent estimator of the integrated covariance matrix when the number of intraday observations M goes to ∞. However, in the empirical situation, we consider the microstructure noise, and often use the realized covariance matrix which is estimated using 5- or 10-min interval intraday returns. In this case, since the Japanese stock market opens from 9 a.m. to 3 p.m. with an hour break, the sample sizes are 60 and 30 per day. Under such a situation, although we want to consider a large portfolio including 100 or 200 stocks, the matrix dimension is larger than the sample size, $d > M$. Therefore, we apply spoet corresponding to $d > M$ to the realized covariance matrix, rather than the POET using PCA as considered in [24,25]. Additionally, we construct the forecasting models similar to [29], by deriving the eigenvalues of the realized covariance matrix estimated using SPOET.

There are two contributions to the literature. First, this paper shows through a simulation study that SPOET considered in the i.i.d. setting has excellent performance for estimating the integrated covariance matrix under the assumption of continuous Itô semi-martingale. Second, our empirical analysis shows that the forecasting models using SPOET are more accurate covariance matrix than the models using the POET. Hence, using our proposed models gives us a more accurate covariance estimator under the high-dimensional setting that results in bad performance and unreliable results. This point is the largest difference between [29] and this paper. Although Jian et al. [29] do not consider the relationship between the dimension of the covariance matrix and the sample size of intraday, we focus on the relationship and make these models forecast more accurately than their models.

The paper is organized as follows: Section 2 explains the factor model, the sparse estimations, and the principal component analysis to estimate the factor part. Section 3 introduces the forecasting model of estimated eigenvalues by PCA used in the empirical analysis. Section 4 gives the result of the simulation study. Section 5 implements the estimator on a large portfolio using individual stocks based on the Nikkei 225. Finally, Section 6 concludes.

2. Factor Model and PCA

2.1. Factor Structure

We assume that the log-price Y follows a continuous-time factor model,

$$Y_t = \beta X_t + Z_t, \tag{1}$$

where Y_t is a d-dimensional vector process, X_t is a r-dimensional latent common factor process, Z_t is the d-dimensional idiosyncratic component, and β is a $d \times r$ constant-factor loading matrix. In addition, X_t and Z_t are independent. In this paper, the number of factors r is unknown. Here, we assume that X_t and Z_t are continuous Itô semi-martingale, as with [24,25] as follows:

$$X_t = \int_0^t h_s ds + \int_0^t \eta_s dW_s, \;\; Z_t = \int_0^t f_s ds + \int_0^t \gamma_s dB_s.$$

Then, the integrated covariance matrices of X_t, Z_t, and Y_t are defined under Assumptions 1, 2, and 3, and the sparsity assumption of [25] as follows:

$$\Sigma_{X_t} = \int_0^t \eta_s \eta_s' ds, \quad \Sigma_{Z_t} = \int_0^t \gamma_s \gamma_s' ds,$$

$$\Sigma_{Y_t} = \beta \Sigma_{X_t} \beta' + \Sigma_{Z_t}. \tag{2}$$

Although Jian et al. [29] consider the factor model following Assumption 1, 2, 3, 4, and 5 of [24], we assume more general sparsity of [25] and we do not assume that idiosyncratic component is block diagonal.

2.2. Sparsity

To estimate an HDCM, a certain condition of sparsity is necessary for dimension reduction and factor model. However, the sparsity assumption of the covariance matrix itself is inappropriate from the viewpoint of the factor model. To solve this problem, we assume that the covariance matrix of the idiosyncratic component Σ_Z is sparse, and then the form of Equation (2) becomes a low-rank plus sparse structure. A low-rank plus sparsity structure of the residual covariance matrix turns out to be a good match for asset high-frequency data [24] and guarantees a well-conditioned estimator as well as its precision matrix [25].

We use four types of thresholding functions, hard-, soft-, adaptive lasso (AL) and smoothly clipped absolute deviation (SCAD) threshold, for Σ_Z as following:

$$s_\lambda^{\text{Hard}}(z) = z\mathbf{1}(|z| > \lambda), \;\; s_\lambda^{\text{Soft}}(z) = \text{sign}(z)(|z| - \lambda)_+, \;\; s_\lambda^{AL}(z) = \text{sign}(z)(|z| - \lambda^{\eta+1}|z|^{-\eta})_+,$$

$$s_\lambda^{SCAD}(z) = \begin{cases} \text{sign}(z)(|z| - \lambda)_+, & |z| \le 2\lambda; \\ \frac{(a-1)z - \text{sign}(z)a\lambda}{a-2}, & 2\lambda < |z| \le a\lambda; \\ z, & a\lambda < |z|. \end{cases}$$

where we set $a = 3.7$ and $\eta = 1$ same as [32]. We adopt these thresholding functions and estimate the residual covariance matrix as follows:

$$\tilde{\Sigma}_{Z_t,ij}^S = \begin{cases} \hat{\Sigma}_{Z_t,ij}, & i = j; \\ s_{\lambda_{ij}}(\hat{\Sigma}_{Z_t,ij}), & i \neq j. \end{cases}$$

Dai et al. [25] denote that despite these estimations lead to the same convergence rate from their analysis, the results of finite sample performance of the covariance matrix in their simulation study and empirical analysis are quite different.

2.2.1. Thresholding Method

Following [25], the thresholding λ_{ij} in sparse functions is estimated as follows:

$$\lambda_{ij} = \tau\sqrt{\hat{\Sigma}_{Z_t,ii}\hat{\Sigma}_{Z_t,jj}},$$

where τ is a constant to be determined. Under the finite sample, we use a grid search to guarantee positive semi-definite. We divide into K pieces in $\tau \in [0, 1]$ and gradually increase τ until the final high-dimensional covariance matrix becomes positive semi-definite. As τ becomes larger, the degree of sparsity of the residual covariance increases, and, finally, the matrix becomes a diagonal matrix [25]. Thus, an estimated HDCM always becomes positive semi-definite.

2.2.2. The Number of Factors

If the log-price is observed by latent common factors, we have to estimate the number of factors. The consistent estimator of the number of latent factors is proposed by [24] under the continuous-time setting without random matrix theory. We adopt their estimator, which minimizes the penalized function using an estimator of the integrated covariance matrix $\hat{\Sigma}_t$:

$$\hat{r}^t = \arg\min_{1 \le j \le r_{\max}} \left(\frac{\lambda_j(\hat{\Sigma}_{Y_t})}{d} + j \times g(M, d) \right) - 1, \tag{3}$$

where $r_{\max}$ is 20. In theory, the choice of $r_{\max}$ is not important. This is simply used to avoid making economically meaningless choice of r in finite samples [24]. The function $g(n, d)$ is defined as follows:

$$g(M, d) = 0.02 \times \hat{\lambda}^t_{\min\left(\frac{d}{2}, \frac{M}{2}\right)}(\hat{\Sigma}_{Y_t}) \left(\frac{\log d}{M} \right)^{\frac{1}{4}}. \tag{4}$$

2.3. *PCA for High-Frequency Data*

To estimate an HDCM, we show the PCA for the realized covariance matrix estimated by high-frequency data following [29]. Here, $y_{j,t}$ is the j-th intraday log-return observed on day t. The realized covariance matrix is defined as follows:

$$\hat{\Sigma}_{Y_t} = \sum_{j=1}^{M} y_{j,t} y'_{j,t}.$$

We assume $d > M$; thus, the realized covariance matrix is estimated under this assumption.

2.3.1. POET Method

The eigenvalues of the realized covariance matrix $\hat{\Sigma}_{Y_t}$ are $\hat{\lambda}_1^t > \hat{\lambda}_2^t > \cdots > \hat{\lambda}_d^t$, and $\hat{\xi}_1^t, \hat{\xi}_2^t, \ldots, \hat{\xi}_d^t$ denote the corresponding eigenvectors. If $\hat{r}$ is the estimator of r, which is the number of factors, $\hat{\Sigma}_{Y_t}$ has a spectral decomposition as follows:

$$\hat{\Sigma}_{Y_t} = \sum_{j=1}^{\hat{r}} \hat{\lambda}_j^t \hat{\xi}_j^t \hat{\xi}_j^{t\prime} + \hat{\Sigma}_{Z_t}, \tag{5}$$

where $\hat{\Sigma}_{Z_t}$ is the covariance matrix of the residual process, which is calculated by $\hat{\Sigma}_{Z_t} = \sum_{j=\hat{r}+1}^{d} \hat{\lambda}_j^t \hat{\xi}_j^t \hat{\xi}_j^{t\prime}$. Here, even if the common factor X_t is an unobservable process, if Σ_Z is sufficiently sparse, $\beta \Sigma_{X_t} \beta'$ in Equation (2) can be estimated using the eigenvalues and eigenvectors of $\hat{\Sigma}_{Y_t}$ [24]. Therefore, we estimate the sparse residual covariance matrix, and then estimate a high-dimensional covariance matrix $\hat{\Sigma}_{Y_t}^S$ as follows:

$$\hat{\Sigma}_{Y_t}^S = \sum_{j=1}^{\hat{r}} \hat{\lambda}_j^t \hat{\xi}_j^t \hat{\xi}_j^{t\prime} + \hat{\Sigma}_{Z_t}^S, \tag{6}$$

where $\hat{\Sigma}_{Z_t}^S$ is the estimated sparse residual covariance matrix. This high-dimensional covariance estimator consists of the POET for low-frequency data of [19] and the PCA approach adopted in [24,25] for high-frequency data.

2.3.2. Shrinkage POET Method

The PCA which is used in Equation (5) is effective, when dimension d is fixed and the sample size (the number of observations in a day) is sufficiently large. However, it is well-known that in situations where $d > M$, the eigenvalues and eigenvectors of the realized covariance matrix are not consistent estimators in the sense that they are quite far from the true values [16]. To deal with this problem, we use shrinkage POET (SPOET), proposed by [31], which corrects biases of empirical eigenvalues and estimates an HDCM as follows:

$$\tilde{\Sigma}_{Y_t}^S = \sum_{j=1}^{r} \tilde{\lambda}_j^t \hat{\xi}_j^t \hat{\xi}_j^{t\prime} + \hat{\Sigma}_{Z_t}^S,$$

where $\tilde{\lambda}_j^t = \max\{\hat{\lambda}_j^t - \bar{c}d/M, 0\}$. In addition, as $\bar{c}$ is unknown, we have to estimate it. In this paper, we follow [31] to estimate as follows:

$$\hat{c} = (\mathrm{tr}(\hat{\Sigma}_Y^t) - \sum_{j=1}^{r} \hat{\lambda}_j)/(d - r - dr/M). \tag{7}$$

3. Forecasting Models

In this section, in order to forecast an HDCM, we introduce forecasting models based on the PCA. We denote the eigenvalues as:

$$\sigma_f^t = [\lambda_1^t, \ldots, \lambda_r^t]'.$$

Since these eigenvalues are the variances of factors, we can consider models similar to the time-series model of the realized variance of asset returns [29]. To model the eigenvalues, we use the exponentially weighted moving average (EWMA), (Vector) HAR, and (Vector)

AR models, the same as [29]. All models except the EWMA model can be easily estimated using OLS.

3.1. EWMA Model

In this paper, we use the EWMA model developed by [33] as a benchmark model, as follows:

$$\sigma_f^{t+1|t} = a\sigma_f^{t|t-1} + (1-a)\sigma^t,$$

where a is the decaying parameter that determines the weight of the observed value 1 period before the forecast, and we set $a = 0.94$ following the framework of a RiskMetrics approach [33]. As this model is easy to implement to forecast volatility and covariance, a lot of studies use it in practice.

3.2. VAR Model

We introduce the AR(1) and VAR models based on high-frequency factor model as:

$$\lambda_i^t = a_{0,i} + a_{1,i}\lambda_i^{t-1} + \varepsilon_i^t, \quad i = 1,\ldots,r, \tag{8}$$

$$\sigma_f^t = A_0 + A_1\sigma_f^{t-1} + \varepsilon_f^t, \tag{9}$$

where $a_{k,i}, A_k, k = 0,1$ are scalar parameters and parameter matrices, respectively. ε_i^t denotes the innovation term.

Andersen et al. [34] pointed out that the logarithmic standard deviations are closer to a normal distribution in general compared to the realized variance itself, and modeling and forecasting log volatility guarantee that the fitted and forecasted volatility are non-negative without any constrains. Therefore, we also apply the logarithmic eigenvalues to these models.

3.3. V-HAR Model

In this subsection, we introduce the HAR model and V-HAR model which are proposed by [7,35], respectively. These models are usually applied to forecasting both univariate and multivariate realized volatility. The HAR model is advantaged for approximating the long memory properties using daily, weekly and monthly volatility. Also, given the multivariate framework, the impact of the short- and long-term volatility of another asset can be included in a forecast of the volatility of one asset. To use these models, we calculate the weekly and monthly eigenvalues as follows:

$$\lambda_{i,W}^t = \frac{1}{5}\sum_{j=0}^{4}\lambda_i^{t-j}, \tag{10}$$

$$\lambda_{i,M}^t = \frac{1}{22}\sum_{j=0}^{21}\lambda_i^{t-j}. \tag{11}$$

In addition, we define that $\sigma_{f.W}^t = [\lambda_{1,W}^t,\ldots,\lambda_{r,W}^t]'$ and $\sigma_{f.M}^t = [\lambda_{1,M}^t,\ldots,\lambda_{r,M}^t]'$. We construct the HAR and V-HAR models using daily, weekly, and monthly eigenvalues as follows:

$$\lambda_i^t = a_{0,i} + a_{1,i}\lambda_i^{t-1} + a_{2,i}\lambda_{i,W}^{t-1} + a_{3,i}\lambda_{i,M}^{t-1} + \varepsilon_i^t, \quad i = 1,\ldots,r, \tag{12}$$

$$\sigma_f^t = A_0 + A_1\sigma_f^{t-1} + A_2\sigma_{f,W}^{t-1} + A_3\sigma_{f,M}^{t-1} + \varepsilon_f^t. \tag{13}$$

where $a_{k,i}, A_k, k = 0,\ldots,3$ are scalar parameters and parameter matrices, respectively. Similar to AR and VAR models, these models are transformed into logarithmic models.

Using these models, we can obtain the forecasted HDCM, $\hat{S}_{t+1}$, as follows:

$$\hat{S}_{t+1} = \sum_{j=1}^{\hat{r}} \check{\lambda}_j^{t+1} \hat{\xi}_j^t \hat{\xi}_j^{t\prime} + \hat{\Sigma}^S_{Z_t}, \tag{14}$$

where $\check{\lambda}_j^{t+1}$ denotes the forecasted j-th eigenvalues at $t+1$, $\hat{\xi}_j$ is the eigenvectors corresponding to the forecasted eigenvalues, and $\hat{\Sigma}^S_{Z_t}$ is the sparse residual covariance matrix at t. Hence, in this model, we use the eigenvectors and sparse residual covariance matrix at t and the forecasted eigenvalues at $t+1$ to forecast an HDCM.

4. Simulation Study

SPOET outperforms POET and the sample covariance matrix for i.i.d. data using a simulation study [31]. We now investigate the small sample performance of the POET and SPOET for a large portfolio, and show that SPOET is more accurate than POET even when applied to estimating the integrated covariance matrix in continuous Itô semi-martingale.

4.1. Simulation Design

In order to investigate the small sample performance, we performed the simulation study in a simple way according to [36,37], which treats the observed realized covariance matrix as the latent integrated covariance matrix. In this paper, we consider not the realized covariance matrix, but the high-dimensional covariance matrix estimated using (S)POET as the integrated covariance matrix. Regarding the observed realized covariance matrix, we discuss this in the Data section of the empirical analysis Section 5.1, below. Their simulation method can simulate empirically realistic sample paths of daily covariance matrices using the observed data. We use a diurnal pattern because the generated returns do not allow stochastic variation in covariances within a day. The intraday volatility pattern is modeled by means based on a diurnal U-shape function, $\sigma_d(u)$. Therefore, we generate the intraday volatility pattern $\sigma_d(u)$, the spot covariance matrix $\Sigma(u)$, and intraday asset returns as follows:

$$dP(u) = \Sigma(u)^{1/2} dW(u), \tag{15}$$

$$\Sigma(u) = \sigma_d(u)\Sigma, \tag{16}$$

$$\sigma_d(u) = C + Ae^{-au} + Be^{-b(1-u)}, \tag{17}$$

where we set $A = 0.75, B = 0.25, C = 0.88929198$, and $a = b = 10$, respectively, following [36,37]. In this simulation study, similar to our empirical analysis described below, we consider 202, 100, and 50 dimensional covariance matrix of 1392 days. We generate one-second prices for each day, and the realized covariance matrix is estimated by 10- and 5-min returns.

4.2. Simulation Result

We evaluate the performance of POET and SPOET by comparing the size of the estimated eigenvalues and the norm for each estimator. Table 1 and Figure 1 show the results.

First, we confirm the size of estimated eigenvalues for each estimator. The average of the first, second, and third eigenvalues of POET and SPOET for the 10- and 5-min interval realized covariance matrix are shown in Table 1. This table shows the mean, maximum and minimum values for each eigenvalue. All eigenvalues estimated by SPOET are closer to the true values than those by POET for all dimensions and all eigenvalues.

Then, we compare the estimation performance of POET and SPOET using $||\hat{S} - S||_2$, $||\hat{S} - S||_F$, MSE (mean square error). $\hat{S}$ and S denote the estimation and the integrated covariance matrix, respectively. Our result is reported in Figure 1; the x axis shows the dimension and the interval, for example, $200(10)$ means the 202 dimensional covariance estimated using 10-min interval intraday returns. Although the colors of the line show the thresholding types, we cannot see the difference between them. The biggest thing to notice is that no matter under what loss functions, interval, or dimension, SPOET always outperforms POET. Except for MSE, the errors become smaller as the return interval becomes shorter and the dimension of the covariance matrix becomes smaller. These results show that what is stated by [31], "*It affirms the claim that shrinkage of spiked eigenvalues is necessary to maintain good performance when the spikes are not sufficiently large*" is also true for the estimation of HDCM using price process.

Table 1. Simulation results of eigenvalues.

		10-min			**5-min**		
		λ_1	λ_2	λ_3	λ_1	λ_2	λ_3
Stocks		Mean					
200	TRUE	194.543	62.640	33.081	194.543	62.640	33.081
	POET	325.517	109.111	58.144	328.368	105.507	56.128
	SPOET	313.308	96.902	45.935	322.158	99.297	49.918
100	TRUE	104.448	35.652	20.452	104.448	35.652	20.452
	POET	181.294	63.451	36.127	174.270	60.168	35.083
	SPOET	175.185	57.342	30.018	171.146	57.044	31.959
50	TRUE	49.150	18.694	11.180	49.150	18.694	11.180
	POET	83.807	31.995	19.540	83.246	31.389	18.679
	SPOET	81.243	29.431	16.977	81.916	30.059	17.349
		Max					
200	TRUE	7843.507	581.130	487.202	7843.507	581.130	487.202
	POET	10,095.698	1050.907	806.341	15,710.724	985.235	727.615
	SPOET	10,059.541	941.379	681.527	15,692.497	927.323	669.703
100	TRUE	3811.211	271.343	192.545	3811.211	271.343	192.545
	POET	6609.435	608.361	299.445	4421.615	524.372	335.179
	SPOET	6579.344	554.550	245.634	4412.434	506.948	310.397
50	TRUE	1662.291	209.630	117.863	1662.291	209.630	117.863
	POET	2477.492	448.610	230.080	3922.754	382.988	169.577
	SPOET	2469.275	420.733	202.203	3918.187	369.209	155.799
		Min					
200	TRUE	29.612	10.916	6.306	29.612	10.916	6.306
	POET	47.262	18.211	13.739	38.446	14.410	10.803
	SPOET	40.198	14.555	9.429	36.110	12.581	8.916
100	TRUE	13.115	6.480	4.319	13.115	6.480	4.319
	POET	22.374	10.855	8.103	23.954	9.939	7.046
	SPOET	19.595	8.971	6.208	22.705	8.972	6.046
50	TRUE	6.555	2.437	1.915	6.555	2.437	1.915
	POET	11.404	5.572	3.550	12.459	4.673	3.687
	SPOET	10.144	4.857	2.835	12.025	4.301	3.315

Notes: This table reports the size of eigenvalues which are first three of the integrated covariance, the POET estimator, and SPOET estimator.

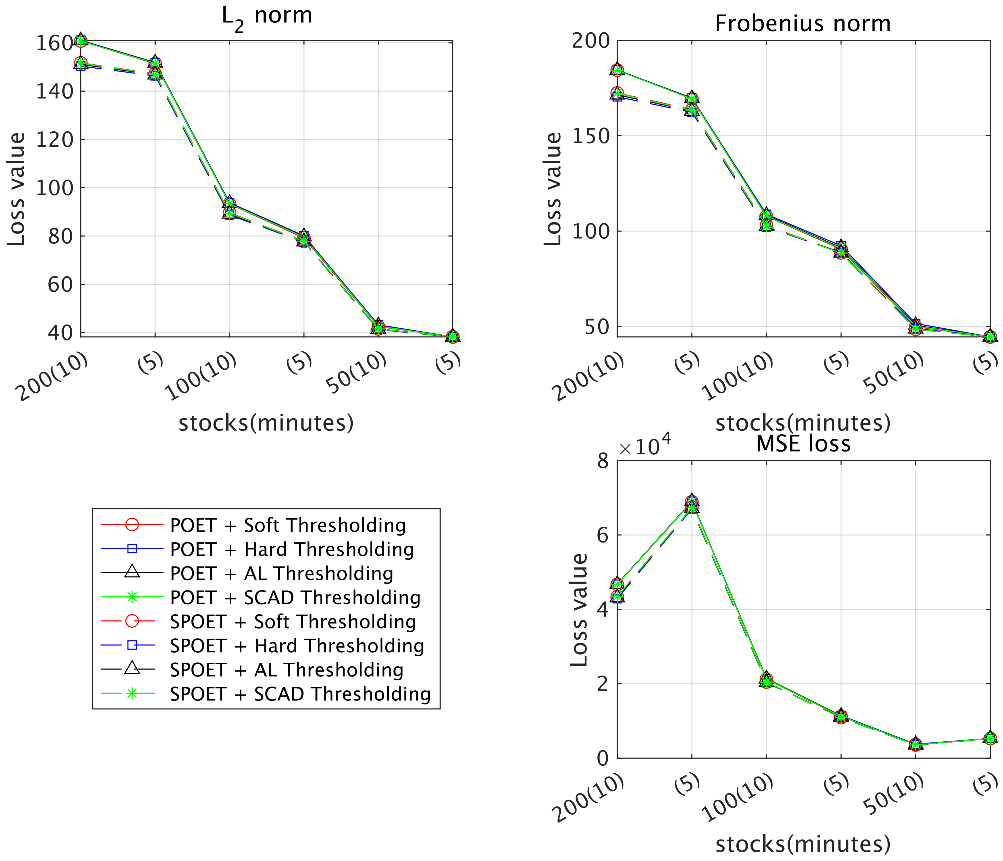

Figure 1. Simulation results. Notes: the x axis shows the number of stocks and the time interval of realized covariance matrix.

5. Empirical Analysis

First, we explain the data we used and its descriptive statistics. Then, the forecasting models are evaluated by loss functions and a variance of portfolio, which is estimated by forecasted covariance matrix.

5.1. Data

We use the high-frequency data of individual stocks included in the Nikkei 225, which we bought from Nikkei NEEDS-TICK data. The sample period covers 1392 days, from 1 January 2015 to 31 December 2020. We adopt a maximum of 202 individual stocks that have traded continuously during the sample period. In addition, we consider not only 202 stocks but also 100 and 50 stocks, and, for each dimension, we estimate the realized covariance matrix using 10- and 5-min interval intraday returns. In order to estimate the realized covariance matrix, we use MFE Toolbox (https://www.kevinsheppard.com/MFE_Toolbox (accessed on 1 November 2022)), which was published by Prof. Kevin Sheppard. However, only for the realized covariance matrix of 50 stocks, we did not consider the matrix with 5-min intervals. This is because high-frequency intraday returns with 5-min intervals have 60 observations in a day, which is not appropriate to the objective of this study, i.e., the situation where the sample size is smaller than the dimension of the matrix.

Figure 2 shows the time series of first, second, and third eigenvalues estimated by POET and their autocorrelation. Similar to [29], each eigenvalue series shows a variation similar to volatility. In addition, since the autocorrelation is significant and positive, the autoregressive models, like the AR and HAR models, are effective to model the eigenvalues. The results of the estimated eigenvalues by SPOET can be observed as the same as POET; therefore, we omit them here.

Table 2 presents the size of the first, second, and third eigenvalues of the 200 dimensional realized covariance matrix estimated by POET and SPOET. We can find that when the number of factors is three, SPOET can estimate the shrinkage eigenvalues for mean, max, and min.

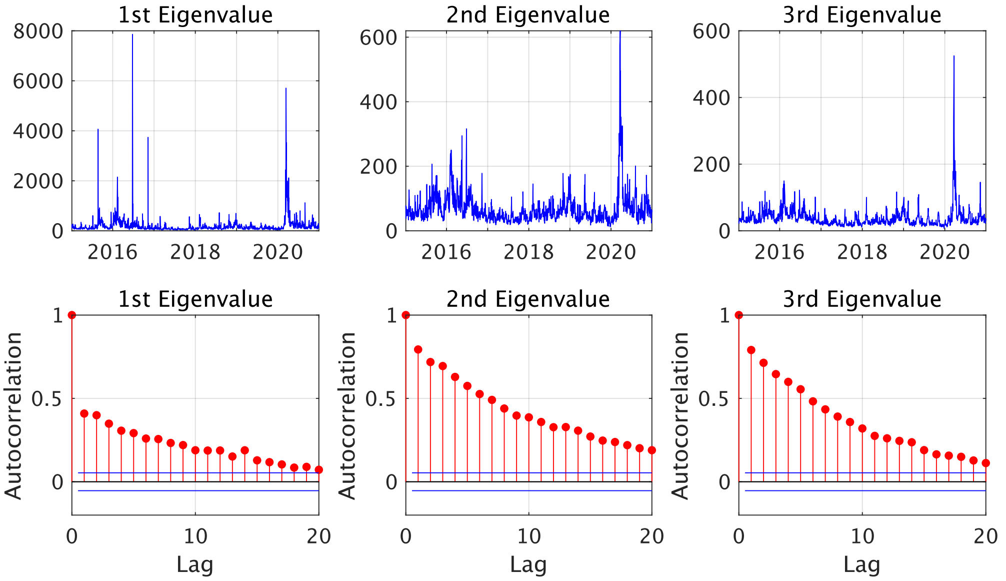

Figure 2. Eigenvalues estimated by POET and sample autocorrelation functions.

Table 2. The eigenvalues estimated by POET and SPOET of 200 dimensional matrix.

	10-min			**5-min**		
	λ_1	λ_2	λ_3	λ_1	λ_2	λ_3
			Mean			
POET	199.61	67.38	37.20	175.16	56.31	35.14
SPOET	191.78	59.55	29.37	170.06	51.21	30.04
			Max			
POET	7859.83	619.82	524.94	4930.99	1379.47	331.24
SPOET	7835.62	546.92	452.04	4909.13	1361.80	288.60
			Min			
POET	31.52	12.46	7.69	22.70	14.25	7.54
SPOET	28.49	10.13	5.36	20.65	12.24	5.90

5.2. *Out-of-Sample*

We evaluate all models using the rolling-window method during the out-of-sample period. These models are reestimated everyday and set 500 days as the rolling window. The process of evaluating the forecasting performance requires using some loss functions, the Diebold–Mariano (DM) test proposed by [38] and the model confidence set (MCS) developed by [39]. Then, based on the forecasted covariance matrix, we construct a portfolio and calculate the variances of returns that are generated by each portfolio.

5.2.1. Loss Functions and MCS

In this paper, to evaluate the forecasting performance at time t, we use the Frobenius distance and MSE which are known to be robust in the presence of noisy covariance matrix proxies [40].

$$\text{Frobenius:}\ \ \text{tr}\big[(\hat{S}_t - \Sigma_t)'(\hat{S}_t - \Sigma_t)\big], \tag{18}$$

$$\text{MSE:}\ \ \text{vech}(\Sigma_t - \hat{S}_t)'\text{vech}(\Sigma_t - \hat{S}_t), \tag{19}$$

where $\hat{S}_t$ is the forecasted HDCM and Σ_t denotes an integrated covariance matrix at t. As a proxy of an integrated covariance matrix, we use an HDCM based on an ex-post observed realized covariance matrix. Then, the 80% MCS is calculated by the result of loss functions. We also calculated the 90% and 70% MCSs, and their results selected the same models with 80% MCS, therefore, we describe the result of 80% MCS. The MCS is calculated by the block bootstrap method with the length of the block beginning at two and the number of bootstrap samples being 10,000.

Tables 3 and 4 show the number of factors for each dimension and the results of loss functions and MCS using the realized covariance matrix estimated by 10- and 5-min intraday returns. The number of factors is estimated by Equations (3) and (4). The results use the soft thresholding method for the sparse estimation of the residual covariance matrix. When other sparse estimations are used, the value of the loss changes, but the results remain the same. Additionally, the MCS is used for the results of a total of 18 models using POET and SPOET for each number of stocks.

Table 3. Average forecasting losses for 10-min interval intraday returns.

10 min	**200 Stocks**		**100 Stocks**		**50 Stocks**	
Observations	30					
Factors	3		4		6	
Frobenius	POET	SPOET	POET	SPOET	POET	SPOET
AR	219.06	210.41 ***	117.65	114.12 ***	54.99	54.47 ***
VAR	202.31	195.70 ***	108.42	105.75 ***	50.79	50.65 **
HAR	204.99	196.76 ***	111.17	107.84 ***	52.45	52.01 ***
V-HAR	199.47	192.11 ***	108.09	105.41 ***	50.67	50.24 ***
AR(log)	192.98	184.47 ***	105.79	102.33 ***	49.61	49.14 ***
VAR(log)	189.82	**181.66** ***	103.50	**100.30** ***	48.62	**48.26**
HAR(log)	188.63	**180.09** ***	103.21	**99.75** ***	48.52	**48.06** ***
V-HAR(log)	188.57	**179.98** ***	102.81	**99.41** ***	48.39	**47.96**
EWMA	212.02	203.34 ***	115.63	112.10 ***	54.73	54.18 ***
MSE						
AR	5.1307	4.8785	1.4675	1.4156	0.3484	0.3446 *
VAR	4.8692	4.6519	1.3938	1.3534	0.3273	0.3262
HAR	4.5806	**4.3417** *	1.3232	**1.2742** *	0.3220	0.3185 *
V-HAR	4.3779	**4.1552** **	1.2834	**1.2416** **	0.3061	**0.3028**
AR(log)	5.3709	5.1806	1.5393	1.4987	0.3652	0.3629
VAR(log)	5.2869	5.1063	1.5049	1.4711	0.3629	0.3624
HAR(log)	4.8197	4.6170	1.3723	1.3304	0.3308	0.3284
V-HAR(log)	4.8930	4.6684 *	1.3854	1.3413	0.3391	0.3368
EWMA	6.0374	5.7767	1.6822	1.6272	0.4088	0.4034 *

Notes: The values of MSE are $\times 10^{-4}$. The selected models by 80% MCS is shown in bold. *, ** , and *** denote significance at 10%, 5%, and 1% levels for DM test.

Table 3 shows the results of the Frobenius distance of all models. First, we compare forecast values of POET and SPOET with the same time-series model using the DM test which is the statistical hypothesis testing based on the difference of the loss to compare the forecasting accuracy between the two models. Therefore, *, **, and *** in the table denote that the value is a more accurate forecast than the competitor at 10%, 5%, and 1% significant levels; for example, in the 200 dimension, since the forecast value 210.41 of the AR model with SPOET has ***, it is more accurate than the forecast of the AR model with POET. In Table 3, for 200 and 100 dimensions, all forecast values of SPOET have better performance than POET for Frobenius loss. For 50 dimensions, almost all the models with SPOET are significant at 1% and 5%. On the other hand, for MSE, although a few models are improved, almost none of the models seem to improve. Therefore, overall, estimating an HDCM using SPOET improces the accuracy of forecasting compared to using POET. Then, we select the best models in terms of forecasting performance. The VAR (log), HAR (log), V-HAR (log) models driven by logarithmic eigenvalues are selected by MCS for all dimensions. In addition, from the perspective of MSE, the models selected by MCS are HAR, and V-HAR models of SPOET with 200 and 100 stocks, and the V-HAR model of SPOET with 50 stocks. Table 4 shows almost the same result as Table 3.

These results say that it is possible for the eigenvalues estimated by POET, in other words, the eigenvalues estimated by classical PCA, to be modeled and forecasted with biases under the high dimension. On the other hand, since the biases of the eigenvalues are corrected by SPOET, the forecasted values obtained using SPOET are more accurate than those obtained using POET. In addition, we compare the AR model with the HAR model, and the VAR model with the V-HAR model. Both the HAR model and the V-HAR model show smaller losses; hence, approximating the long memory property for eigenvalue-driven models is effective.

Table 4. Average forecasting losses for 5-min interval intraday returns.

5 min	**200 Stocks**		**100 Stocks**	
Observations	60			
Factors	3		4	
Frobenius	POET	SPOET	POET	SPOET
AR	170.86	165.89 ***	94.80	93.33 ***
VAR	157.61	153.61 ***	87.50	86.38
HAR	158.93	154.26 ***	88.93	87.58 ***
V-HAR	156.93	152.52 ***	88.05	86.72 ***
AR(log)	153.33	148.61 ***	86.49	85.11
VAR(log)	150.30	**145.84** ***	84.16 **	**82.94**
HAR(log)	149.11	**144.33** ***	84.01	**82.62**
V-HAR(log)	148.74	**143.98** ***	83.56	**82.12**
EWMA	171.51	166.62 ***	95.52	94.05 ***
MSE				
AR	3.2853	3.1792	1.0234	1.0062
VAR	3.0918	3.0000	0.9562	0.9441
HAR	2.9739	**2.8748**	0.9212	0.9056
V-HAR	2.8736	**2.7764** *	0.8970	**0.8818** **
AR(log)	3.7911	3.7246	1.1351	1.1243
VAR(log)	3.7055	3.6469	1.0873	1.0823
HAR(log)	3.3425	3.2647	1.0035	0.9913
V-HAR(log)	3.3448	3.2621	0.9946	0.9816
EWMA	4.5155	4.3992	1.2842	1.2645

Notes: The values of MSE are $\times 10^{-4}$. The selected models by 80% MCS is shown in bold. *, ** , and *** denote significance at 10%, 5%, and 1% levels for DM test.

5.2.2. Portfolio Performance

In order to determine the best model among our proposed and benchmark models, we compare their forecasting performance in an economic context. In this paper, we consider that the model which generates a smaller variance portfolio than other models is better. The portfolio is estimated by the minimum variance portfolio without short selling. The weight of each stock including a portfolio can be calculated based on the results of the following optimization problem:

$$\begin{aligned} \min_{\omega_t} \quad & \omega_t' \hat{S}_t \omega_t, \\ s.t. \quad & \sum_{i=1}^{N} \omega_{t,i} = 1, 0 \leq \omega_{t,i} \leq 1, \end{aligned} \tag{20}$$

where ω_t denotes the vector of portfolio weight at t, and $\hat{S}_t$ denotes the high-dimensional covariance matrix forecasted by each model.

Figures 3 and 4 show the average portfolio variance estimated by POET, SPOET, and each forecasting model. The x axis shows the types of models and the y axis shows the portfolio variance. The color of each line denotes the thresholding method of the residual covariance matrix, the red, blue, black, and green show the soft, hard, AL, and SCAD thresholding, respectively. In addition, the results of POET are indicated by the solid lines and those for SPOET are indicated by the dotted lines.

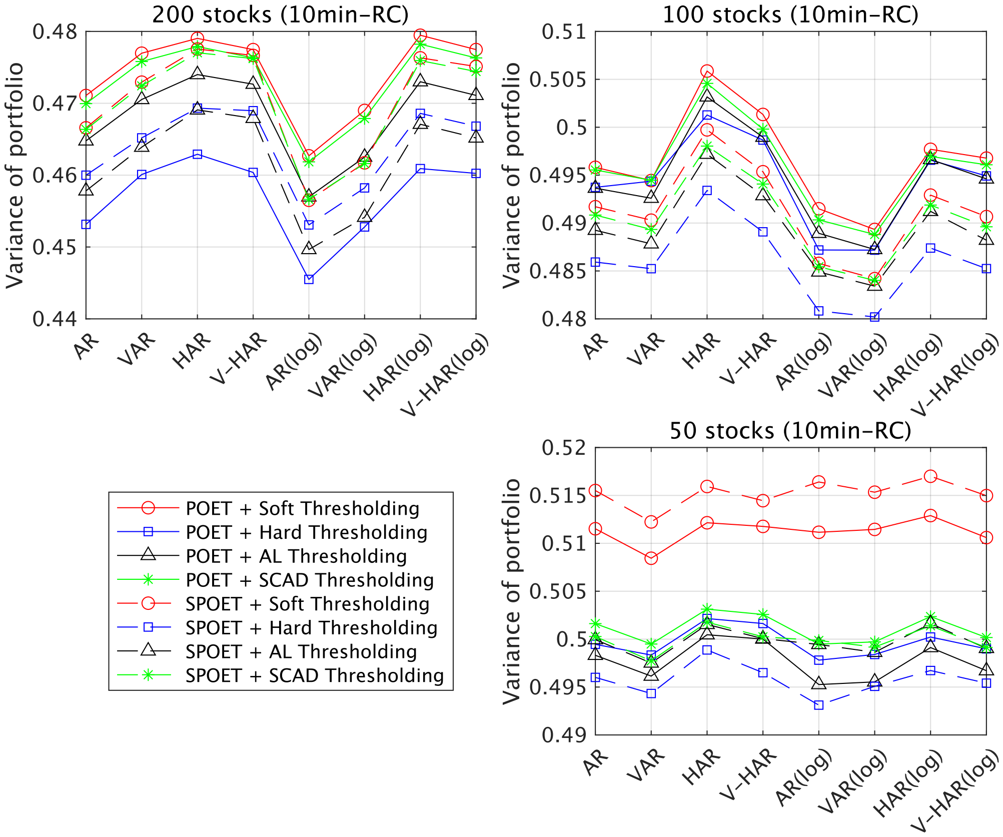

Figure 3. The variance of portfolios estimated by each model using 10-min interval intraday returns.

In Figure 3, the result of 10-min realized covariance matrix is shown. For the 200 dimensions, the pair of POET and hard thresholding have the best performance among competing models. However, comparing POET and SPOET with the same sparse estimation shows that the results with SPOET generate portfolios with smaller variance except for hard thresholding. For the 100 dimensions, the pair of SPOET and hard are the best. Additionally, for all models, the forecasting models with SPOET have better performance than those with POET. Finally, for the 50 dimensions, we cannot find the differences between POET and SPOET, and the soft thresholding performs worse than other thresholding methods.

Figure 4 shows the result of a 5-min realized covariance matrix. The differences in performance between POET and SPOET become smaller than in the case of the 10-min realized covariance matrix. This is perhaps because increasing the number of intraday returns makes the classical PCA performance improve. However, for both dimensions, the performances of SPOET are still better than POET.

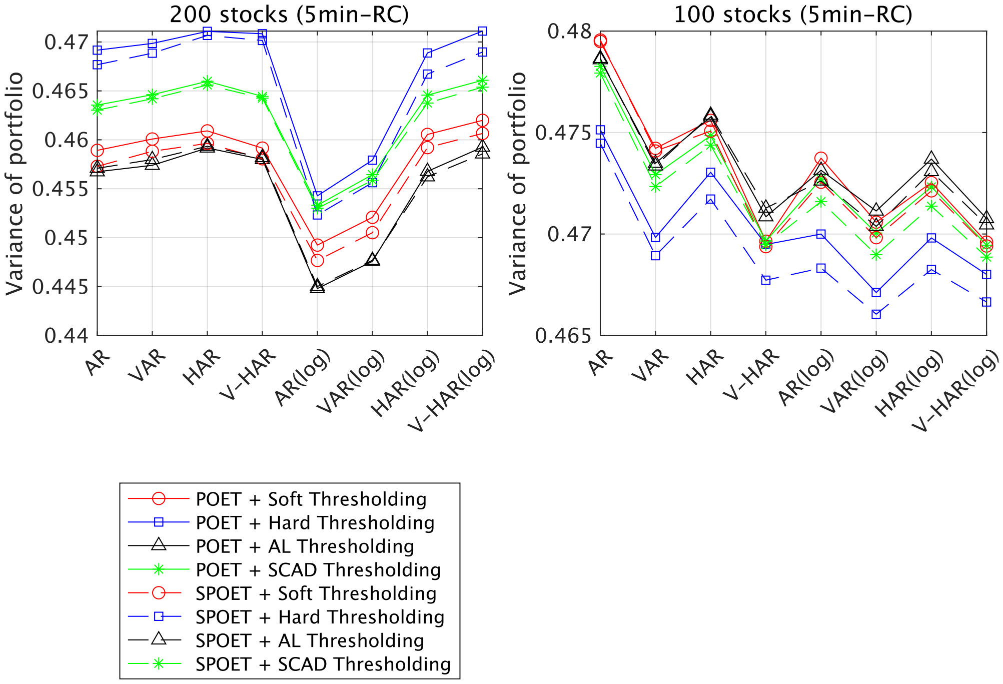

Figure 4. The variance of portfolios estimated by each model using 5-min interval intraday returns.

6. Conclusions

In this study, we constructed the HDCM forecasting models using high-dimensional PCA. In particular, the previous studies show that to estimate the latent factors, POET is used. However, it is known that when the dimension is greater than the sample size, the eigenvalues estimated by classical PCA have biases. Therefore, in order to estimate the eigenvalues more accurately, we adopted SPOET which corrects biases of empirical eigenvalues. In addition, we combined eigenvalues and time-series models to forecast eigenvalues and covariance matrix.

In the simulation study, we generated the asset returns based on the estimated HDCM as the integrated covariance matrix and it shows that SPOET is also effective for the price process. Especially, the empirical eigenvalues of SPOET were closer to the true values than those of POET.

In the empirical analysis, we constructed some forecasting models of HDCM using a number of individual stocks traded on Nikkei 225. Almost all our proposed models which use SPOET show better performance than the other models which use POET. In addition, in terms of economic performance, our models can generate a smaller variance than benchmarks in most cases. This study applied SPOET discussed under the i.i.d. setting to the continuous Itô semi-martingale setting for simulation study and empirical analysis. Thus, theoretical results are needed in the future.

Author Contributions: Conceptualization, H.S. and T.M.; software, H.S.; data curation, H.S. and T.M.; formal analysis, H.S.; writing—original draft preparation, H.S.; writing—review and editing, H.S. and T.M.; supervision, T.M. All authors have read and agreed to the published version of the manuscript.

Funding: This study is partly supported by the Institute of Statistical Mathematics (ISM) cooperative research program (2022-ISMCRP-2024), JSPS KAKENHI Grant Number 21K01433, and Grant-in-Aid for JSPS Fellows Grant Number 22J10285.

Data Availability Statement: The real data we used can be bought from Nikkei NEEDS-TICK data.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Bauwens, L.; Laurent, S.; Rombouts, J. Multivariate GARCH models: A survey. *J. Appl. Econom.* **2006**, *21*, 79–109. [CrossRef]
2. Engle, R.; Kroner, K. Multivariate simultaneous generalized arch. *Econom. Theory* **1995**, *11*, 122–150. [CrossRef]
3. Tse, Y.; Tsui, A. A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. *J. Bus. Econ. Stat.* **2002**, *20*, 351–362. [CrossRef]
4. Engle, R. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *J. Bus. Econ. Stat.* **2002**, *20*, 339–350. [CrossRef]
5. Barndorff-Nielsen, O.; Shephard, N. Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica* **2004**, *72*, 885–925. [CrossRef]
6. Barndorff-Nielsen, O.; Hansen, P.; Lunde, A.; Shephard, N. Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *J. Econom.* **2011**, *162*, 149–169. [CrossRef]
7. Bubák, V.; Kočenda, E.; Žikeš, F. Volatility transmission in emerging European foreign exchange markets. *J. Bank. Financ.* **2011**, **35**, 2829–2841. [CrossRef]
8. Golosnoy, V.; Gribisch, B.; Liesenfeld, R. The conditional autoregressive Wishart model for multivariate stock market volatility. *J. Econom.* **2012**, *167*, 211–223. [CrossRef]
9. Bauwens, L.; Storti, G.; Violante, F. Dynamic conditional correlation models for realized covariance matrices. *CORE DP* **2012**, *60*, 104–108.
10. Engle, R.; Ledoit, O.; Wolf, M. Large dynamic covariance matrices. *J. Bus. Econ. Stat.* **2019**, *37*, 363–375. [CrossRef]
11. Nakagawa, K.; Imamura, M.; Yoshida, K. Risk-based portfolios with large dynamic covariance matrices. *Int. J. Financ. Stud.* **2018**, *6*, 52. [CrossRef]
12. Moura, G.; Santos, A.; Ruiz, E. Comparing high-dimensional conditional covariance matrices: Implications for portfolio selection. *J. Bank. Financ.* **2020**, *118*, 105882. [CrossRef]
13. De Nard, G.; Engle, R.; Ledoit, O.; Wolf, M. Large dynamic covariance matrices: Enhancements based on intraday data. *J. Bank. Financ..* **2022**, *138*, 106426. [CrossRef]
14. Trucíos, C.; Mazzeu, J.; Hallin, M.; Hotta, L.; Valls Pereira, P.L.; Zevallos, M. Forecasting conditional covariance matrices in high-dimensional time series: A general dynamic factor approach. *J. Bus. Econ. Stat.* **2021**, 1–13. [CrossRef]
15. Wang, Y.; Zou, J. Vast volatility matrix estimation for high-frequency financial data. *Ann. Stat.* **2010**, *38*, 943–978. [CrossRef]
16. Tao, M.; Wang, Y.; Yao, Q.; Zou, J. Large volatility matrix inference via combining low-frequency and high-frequency approaches. *J. Am. Stat. Assoc.* **2011**, *106*, 1025–1040. [CrossRef]
17. Kim, D.; Wang, Y.; Zou, J. Asymptotic theory for large volatility matrix estimation based on high-frequency financial data. *Stoch. Process. Their Appl.* **2016**, *126*, 3527–3577. [CrossRef]
18. Shen, K.; Yao, J.; Li, W. Forecasting high-dimensional realized volatility matrices using a factor model. *Quant. Financ.* **2020**, *20*, 1879–1887. [CrossRef]
19. Fan, J.; Liao, Y.; Mincheva, M. Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2013**, *75*, 603–680. [CrossRef]
20. Fan, J.; Furger, A.; Xiu, D. Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data. *J. Bus. Econ. Stat.* **2016**, *34*, 489–503. [CrossRef]
21. Brownlees, C.; Nualart, E.; Sun, Y. Realized networks. *J. Appl. Econom.* **2018**, *33*, 986–1006. [CrossRef]
22. Koike, Y. De-biased graphical lasso for high-frequency data. *Entropy* **2020**, *22*, 456. [CrossRef]
23. Fan, J.; Fan, Y.; Lv, J. High dimensional covariance matrix estimation using a factor model. *J. Econom.* **2008**, *147*, 186–197. [CrossRef]
24. Aït-Sahalia, Y.; Xiu, D. Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *J. Econom.* **2017**, *201*, 384–399. [CrossRef]
25. Dai, C.; Lu, K.; Xiu, D. Knowing factors or factor loadings, or neither? Evaluating estimators of large covariance matrices with noisy and asynchronous data. *J. Econom.* **2019**, *208*, 43–79. [CrossRef]
26. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
27. Fan, J.; Li, R. Variable selection via nonconcave penalized. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
28. Cai, T.; Liu, W. Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Stat. Assoc.* **2011**, *106*, 672–684. [CrossRef]
29. Jian, Z.; Deng, P.; Zhu, Z. High-dimensional covariance forecasting based on principal component analysis of high-frequency data. *Econ. Model.* **2018**, *75*, 422–431. [CrossRef]
30. Yata, K.; Aoshima, M. Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *J. Multivar. Anal.* **2012**, *105*, 193–215. [CrossRef]
31. Wang, W.; Fan, J. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann. Stat.* **2017**, *45*, 1342–1374. [CrossRef] [PubMed]
32. Rothman, A.; Levina, E.; Zhu, J. Generalized thresholding of large covariance matrices. *J. Am. Stat. Assoc.* **2009**, *104*, 177–186. [CrossRef]
33. J.P.Morgan/Reuters. *Risk Metrics. Thechnical Document*, 4th ed.; J.P.Morgan/Reuters: New York, NY, USA, 1996.
34. Andersen, T.; Bollerslev, T.; Diebold, F. Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *Rev. Econ. Stat.* **2007**, *89*, 701–720. [CrossRef]

35. Corsi, F. A simple approximate long-memory model of realized volatility. *J. Financ. Econom.* **2009**, *7*, 174–196. [CrossRef]
36. Andersen, T.; Dobrev, D.; Schaumburg, E. Jump-robust volatility estimation using nearest neighbor truncation. *J. Econom.* **2012**, *169*, 75–93. [CrossRef]
37. Bollerslev, T.; Patton, A.; Quaedvlieg, R. Modeling and forecasting (un)reliable realized covariances for more reliable financial decisions. *J. Econom.* **2018**, *207*, 71–91. [CrossRef]
38. Diebold, F.; Mariano, R. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–265.
39. Hansen, P.R.; Lunde, A.; Nason, J.M. The Model Confidence Set. *Econometrica* **2011**, *79*, 453–497. [CrossRef]
40. Laurent, S.; Rombouts, J.; Violante, F. On loss functions and ranking forecasting performances of multivariate volatility models. *J. Econom.* **2013**, *173*, 1–10. [CrossRef]