Quantitative Finance, 2020
Vol. 20, No. 11, 1879–1887, https://doi.org/10.1080/14697688.2018.1473632

Forecasting high-dimensional realized volatility matrices using a factor model

KEREN SHEN*, JIANFENG YAO and WAI KEUNG LI

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam, Hong Kong

(*Received 6 June 2016; accepted 1 May 2018; published online 7 June 2018*)

Modelling and forecasting covariance matrices of asset returns play a crucial role in many financial fields, such as portfolio allocation and asset pricing. The availability of high-frequency intraday data enables the modelling of the realized covariance matrix directly. However, most models in the literature suffer from the curse of dimensionality, i.e. the number of parameters needed increases at the rate of the square of the number of assets. To solve the problem, we propose a factor model with a diagonal Conditional Autoregressive Wishart model for the factor realized covariance matrices. Consequently, the positive definiteness of the estimated covariance matrix is ensured with the proposed model. Asymptotic theory is derived for the estimated parameters. In the extensive empirical analysis, we find that the number of parameters can be reduced significantly; to only about one-tenth of the benchmark model. Furthermore, the proposed model maintains a comparable performance with a benchmark vector autoregressive model for different forecast horizons.

Keywords: High-dimension; High-frequency; Realized covariance matrices; Factor model; Wishart distribution

JEL Classification: C32, C51, C53

1. Introduction

Modelling and forecasting covariances or volatility matrices of asset returns play a crucial role in many financial fields, such as portfolio allocation (Markowitz 1952) and asset pricing (Bollerslev *et al*. 1988). With the availability of intraday financial data nowadays, it becomes possible to estimate volatilities and co-volatilities of asset returns using high-frequency data directly, which leads to the so-called *realized covariance matrix* (Andersen *et al*. 2003, Barndorff-Nielsen and Shephard 2004, Barndorff-Nielsen *et al*. 2011). Two major problems arise in the estimation of realized covariance matrices. Firstly, transactions for different assets are typically asynchronous so that the high-frequency prices of different assets do not change simultaneously. Secondly, it is widely believed that the observed high-frequency prices are accompanied by *microstructure noise* so that the observed prices should be thought as a noisy version of the true underlying price process. Researchers have proposed several ways to tackle these problems, for example, overlap intervals and the previous tick method by Hayashi and Yoshida (2005) and Zhang (2011), respectively. Moreover, Bannouh *et al*. (2012) use a refresh time scheme and Christensen *et al*. (2010) propose the pre-averaging approach.

Once constructed, realized covariance matrices are analysed using multivariate models. There are two major issues to be resolved in modelling realized covariance matrices. The first issue is that the model should guarantee the positive definiteness of fitted covariance matrices. A natural choice in this aspect is the family of matrix-valued Wishart distributions which automatically generates random positive definite matrices without imposing additional constraints. Several models related to the Wishart distribution have been put forward. Gourieoux *et al*. (2009) propose the Wishart Autoregressive (WAR) model where the realized covariance matrix has a conditional distribution which is noncentral Wishart with a non-centrality parameter depending on lagged covariances and a fixed scaling matrix. Later, Golosnoy *et al*. (2012) propose the Conditional Autoregressive Wishart (CAW) model under which the conditional distribution is central Wishart with time dependent scaling matrices. Moreover, Yu *et al*. (2017) generalize the above two models to construct the Generalized Conditional Autoregressive Wishart model. There are other models involving

*Corresponding author. Email: rayshenkr@hku.hk, rayshen0111@gmail.com

the Wishart distribution, for instance, see Jin and Maheu (2009).

The second issue in modelling realized covariance matrices is the high-dimensionality. Indeed, covariance matrices have $d(d+1)/2$ entries for d assets; consequently, the number of parameters in the model for realized covariance matrices grows quickly with d. For example, for an unrestricted CAW(2,2) model with $d=10$ assets, as many as 456 parameters are needed so that it is quite challenging to fit such a model in practice. This is probably the major reason why all the empirical studies we find in the literature on model fitting for realized covariance matrices are all limited to a *small number*, say 3–5 assets. Another problem inherent in realized covariance matrices is that they deviate from their population counterpart, the so-called *integrated covariance matrix*, when the number of assets is large compared to the sample size (Johnstone and Lu 2009, Wang and Zou 2010). It is therefore important to build statistical models for realized covariance matrices with a large dimension, say several tens.

Improved estimators of realized covariance matrices are proposed in Wang and Zou (2010) with a so-called averaging realized volatility matrix estimator. Tao *et al*. (2011) propose the threshold averaging realized volatility matrix (TARVM) estimator which is two-scale and uses the previous-tick method and the threshold technique in constructing realized volatility matrices. Then, inspired by Zhang (2006) and Fan and Wang (2007), Tao *et al*. (2013) propose the threshold multi-scale realized volatility matrix (TMSRVM) estimator. The TARVM and TMSRVM estimators are shown to be consistent for the integrated covariance matrix when the dimension of the realized covariance matrix, the sample size of intraday points and the length of sampling days go to infinity. In addition, the TMSRVM estimator proves to have the optimal convergence rate under the existence of the microstructure noise.

As an effort to control the parametric dimension of the models, Tao *et al*. (2011) propose to first identify a small number of factors for the realized covariance matrices and to fit a vector autoregressive (VAR) model to the vectorized factor covariance matrices. They show that such a factor model significantly reduces the number of parameters needed to fit the realized covariance matrices. However, the VAR specification is not able to ensure the positive definiteness of the predicted factor covariance matrices, which is crucial for some real financial applications, for example the Markowitz portfolio problem. In addition (see below), such a VAR fit still needs $\mathcal{O}(r^4)$ number of parameters with r factors.

In this article, we adopt the factor model approach introduced in Tao *et al*. (2011) for realized covariance matrices. In order to overcome the aforementioned weakness of their VAR fit for the extracted factors, we propose a diagonal CAW model which has several advantages. Firstly, the proposed CAW model is able to guarantee automatically the positive definiteness of the covariance matrices generated from the model without imposing additional constraints. Secondly, as will be shown by extensive data analysis reported in this paper, our model has excellent empirical performance in terms of the reduction of the number of parameters compared to the VAR approach proposed in Tao *et al*. (2011): indeed, we obtain comparable forecasting performance with much fewer parameters.

In a related work, Asai and McAleer (2014) also use a combination of factor extraction and CAW modelling and report some empirical studies with seven assets which is still considered to be small dimension. In this paper, we focus on a larger collection of assets where empirical studies are carried out for 30 assets. As far as we know, this is the first model dealing with high-dimensional realized volatility matrix ensuring the positive definiteness. A further difference in this paper is that we also propose a thorough theoretical analysis of both the factor modelling and the CAW estimation.

The rest of the paper is organized as follows. Section 2 introduces the model set-up and our approach based on a factor model and a diagonal CAW model for the extracted factors. In Section 3, the asymptotic theory is established. In Section 4, we report the empirical data analysis on asset prices using the proposed model. Conclusions are presented in Section 5. Proofs of the asymptotic theory are provided in the Appendix.

2. Methodology

2.1. Model set-up

Suppose there are d assets and their log price process $\mathbf{X}(t) = \{X_1(t), \ldots, X_d(t)\}'$ follows a continuous diffusion model:

$$\mathrm{d}\mathbf{X}(t) = \boldsymbol{\mu}_t \,\mathrm{d}t + \boldsymbol{\sigma}_t \,\mathrm{d}\mathbf{W}_t, \quad t \in [0, T], \tag{1}$$

where $\boldsymbol{\mu}_t$ is a drift in $\mathbb{R}^d$, $\mathbf{W}_t$ is a standard d-dimensional Brownian motion and $\boldsymbol{\sigma}_t$ is a $d \times d$ matrix. The *integrated volatility matrix* for the t-th day is defined as

$$\boldsymbol{\Sigma}_x(t) = \int_{t-1}^{t} \boldsymbol{\sigma}_s \boldsymbol{\sigma}_s' \,\mathrm{d}s, \quad t = 1, \ldots, T. \tag{2}$$

However, it is commonly admitted that the microstructure noise is inherent in the high-frequency price process so that we are not able to observe directly $X_i(t)$, but $Y_i(t_{i\ell})$, a noisy version of $X_i(\cdot)$ at times $t_{i\ell}$, $\ell = 1, \ldots, n_i$, $i = 1, \ldots, d$. Here, n_i is the total trading times and $t_{i\ell}$ is the ℓ-th trading times of asset i during a giving trading day t. The observations $Y_i(t_{i\ell})$ is allowed to be non-synchronized, i.e. $t_{i\ell} \neq t_{j\ell}$ for any $i \neq j$. In this paper, we assume that

$$Y_i(t_{i\ell}) = X_i(t_{i\ell}) + \epsilon_i(t_{i\ell}), \tag{3}$$

where $\epsilon_i(t_{i\ell})$ are iid microstructure noise with mean zero and variance η_i, and $\epsilon_i(\cdot)$ and $X_i(\cdot)$ are independent with each other.

2.2. Realized covariance matrix estimator

Several issues arise for the estimation of $\boldsymbol{\Sigma}_x(t)$: (1) asynchronous observations of different assets; (2) microstructure noise and (3) the number of assets can be larger than the sample size. In this paper, we adopt the threshold Multi-Scale

Realized Volatility Matrix estimator (threshold MSRVM) proposed by Tao *et al.* (2013), denoted by $\hat{\mathbf{\Sigma}}_x(t), t = 1, \ldots, T$. The threshold MSRVM estimator has many attractive properties, for instance, it is consistent for the high-dimensional integrated co-volatility matrix with the optimal convergence rate. Briefly, the idea of the threshold MSRVM estimator is the following: the previous-tick method is used to construct the raw realized covariance matrices. Then, a multi-scale estimator is evaluated which is actually a kind of average of those raw estimators. In addition, the multi-scale estimator is regularized using a thresholding method, that is, matrix entries under a threshold are set to be zero.

2.3. Matrix factor model

We adopt the following factor model to reduce the large dimension of ICV $\mathbf{\Sigma}_x(t)$:

$$\mathbf{\Sigma}_x(t) = \mathbf{A}\mathbf{\Sigma}_f(t)\mathbf{A}' + \mathbf{\Sigma}_0, \tag{4}$$

for $t = 1, \ldots, T$, where $\mathbf{\Sigma}_f(t)$ are $r \times r$ positive definite factor covariance matrices, $\mathbf{\Sigma}_0$ is a $d \times d$ positive definite constant matrix and $\mathbf{A}$ is a $d \times r$ factor loading matrix normalized by the constraint $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$. As in a standard factor model, only the left-hand side of equation (4) is observed. The unknown quantities on the right-hand side are estimated using the method in Tao *et al.* (2011). Given the threshold MSRVM estimator $\hat{\mathbf{\Sigma}}_x(t)$, let

$$\bar{\hat{\mathbf{\Sigma}}}_x = \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{\Sigma}}_x(t) \tag{5}$$

and

$$\bar{\hat{\mathbf{S}}}_x = \frac{1}{T}\sum_{t=1}^{T}\{\hat{\mathbf{\Sigma}}_x(t) - \bar{\hat{\mathbf{\Sigma}}}_x\}^2. \tag{6}$$

Next, the estimator $\hat{\mathbf{A}}$ is obtained using the r orthonormal eigenvectors of $\bar{\hat{\mathbf{S}}}_x$, corresponding to its r largest eigenvalues, as its columns. Finally, the estimated factor covariance matrices are

$$\hat{\mathbf{\Sigma}}_f(t) = \hat{\mathbf{A}}'\hat{\mathbf{\Sigma}}_x(t)\hat{\mathbf{A}}, \tag{7}$$

for $t = 1, \ldots, T$; and $\mathbf{\Sigma}_0$ is estimated by

$$\hat{\mathbf{\Sigma}}_0 = \bar{\hat{\mathbf{\Sigma}}}_x - \hat{\mathbf{A}}\hat{\mathbf{A}}'\bar{\hat{\mathbf{\Sigma}}}_x\hat{\mathbf{A}}\hat{\mathbf{A}}'. \tag{8}$$

2.4. CAW modelling for factor covariance matrix

With the estimated factor covariance matrices $\{\hat{\mathbf{\Sigma}}_f(t)\}$ calculated by equations (7) and (8), we construct a dynamic structure by fitting a diagonal CAW model to $\tilde{\mathbf{\Sigma}}_f(t)$, where $\tilde{\mathbf{\Sigma}}_f(t) := \mathbf{\Sigma}_f(t) + \mathbf{A}'\mathbf{\Sigma}_0\mathbf{A}$. Here, $\tilde{\mathbf{\Sigma}}_f(t)$ is modelled rather than $\mathbf{\Sigma}_f(t)$ since $\hat{\mathbf{\Sigma}}_f(t)$ is in fact a consistent estimator of the former, and it is impossible to construct a consistent estimator for the latter, to our knowledge.

The model is defined as follows. Let $\mathcal{F}_{t-1} = \sigma(\tilde{\mathbf{\Sigma}}_f(s), s < t)$ be the past history of the process at time t. Conditional on $\mathcal{F}_{t-1}$, $\tilde{\mathbf{\Sigma}}_f(t)$ follows a central Wishart distribution

$$\tilde{\mathbf{\Sigma}}_f(t)|\mathcal{F}_{t-1} \sim \mathcal{W}_n(\nu, \mathbf{S}_f(t)/\nu), \tag{9}$$

with ν being the degrees of freedom and the scaling matrix $\mathbf{S}_f(t)$. Moreover, the scaling matrix $\mathbf{S}_f(t)$ follows a linear recursion of order (p, q)

$$\mathbf{S}_f(t) = CC' + \sum_{i=1}^{p} B_i\mathbf{S}_f(t-i)B_i' + \sum_{j=1}^{q} A_j\tilde{\mathbf{\Sigma}}_f(t-j)A_j', \tag{10}$$

where A_j, B_i and C are all $r \times r$ matrices of coefficients.

In summary, the CAW process depends on the parameters $\{\nu, C, (B_i)_{1\le i\le p}, (A_j)_{1\le j\le q}\}$ without additional constraints, so that the total number of parameters is equal to $(p+q)r^2 + r(r+1)/2 + 1 = \mathcal{O}(r^2)$ which still grows quickly with the number of factors r and the order p and q. Since the main aim of the paper is to propose a practically feasible model for a large number of assets while retaining efficiency, we will restrict ourselves to *diagonal* coefficient matrices C, $(B_i)_{1\le i\le p}$ and $(A_j)_{1\le j\le q}$. Therefore, the number of parameters becomes $(p+q+1)r + 1 = \mathcal{O}(r)$. Notice that this set-up is also supported by the fact that in the literature (McCurdy and Stengos 1992, Engle and Kroner 1995), researchers tend to use diagonal volatility models to avoid over-parameterization and argue that the variances and the covariances rely more on its own past than the history of other variances or covariances. In the empirical study developed below, we find that the diagonal models achieve a comparable performance with unrestricted ones, while being much more parsimonious and requiring far less computing time. Notice, however, the asymptotic theory developed below is also valid for unrestricted matrices A_j's, B_i's and C.

The estimation of the parameters $\theta = (\nu, \mathrm{diag}(C)', \mathrm{diag}(B_i)'_{1\le i\le p}, \mathrm{diag}(A_j)'_{1\le j\le q})'$ of the diagonal CAW(p, q) model is carried out by maximizing the log-likelihood function using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization procedure. Positivity of the diagonal elements C_{kk}, $A_{11,j}$ and $B_{11,i}$ are enforced, where $1 \le k \le r$. The log-likelihood function is

$$\begin{aligned}\mathcal{L}(\theta) = \frac{1}{T}\sum_{t=1}^{T}\Bigg\{&-\frac{\nu r}{2}\ln(2) - \frac{r(r-1)}{4}\ln(\pi)\\ &-\sum_{i=1}^{r}\ln\Gamma\left(\frac{\nu+1-i}{2}\right) - \frac{\nu}{2}\ln\left|\frac{\mathbf{S}_f(t)}{\nu}\right|\\ &+\left(\frac{\nu-r-1}{2}\right)\ln|\tilde{\mathbf{\Sigma}}_f(t)| - \frac{1}{2}\mathrm{tr}(\nu\mathbf{S}_f(t)^{-1}\tilde{\mathbf{\Sigma}}_f(t))\Bigg\}.\end{aligned} \tag{11}$$

In practice, initial values for $\mathbf{S}_f(t)$ are needed to run the maximization of the log-likelihood function. For example, if the order $(p, q) = (2, 2)$ is used, then the initial values $\mathbf{S}_f(1)$ and $\mathbf{S}_f(2)$ are needed. In the empirical analysis of this paper, we take $\mathbf{S}_f(1) = \hat{\mathbf{\Sigma}}_f(1)$ and $\mathbf{S}_f(2) = \hat{\mathbf{\Sigma}}_f(2)$ as $\mathbf{S}_f(t)$ is the conditional expectation of $\tilde{\mathbf{\Sigma}}_f(t)$, for any t.

3. Asymptotic theory

Given a d-dimensional vector $\mathbf{x} = (x_1, \ldots, x_d)'$ and a $d \times d$ matrix $\mathbf{U} = (U_{ij})$, define vector and matrix norms as

$$||\mathbf{x}||_2 = \left(\sum_{i=1}^{d} |x_i|^2\right)^{1/2}, \quad ||\mathbf{U}||_2 = \sup\{||\mathbf{U}\mathbf{x}||_2, ||\mathbf{x}||_2 = 1\}. \tag{12}$$

In fact, $||\mathbf{U}||_2$ is the spectral norm, equal to the square root of the largest eigenvalue of $\mathbf{U}'\mathbf{U}$.

In addition, define the Frobenius norm of the $d \times d$ matrix $\mathbf{U} = (U_{ij})$ as

$$||U||_F = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d} |U_{ij}|^2}. \tag{13}$$

The asymptotic theory below uses the following assumptions.

(A1) All row vectors of $\mathbf{A}'$ and $\mathbf{\Sigma}_0$ in the factor model (4) satisfy the sparsity condition (14) below. We say that a d-dimensional vector $\mathbf{x} = (x_1, \ldots, x_d)'$ is sparse if

$$\sum_{j=1}^{d} |x_i|^{\delta} \le C\pi(d), \tag{14}$$

where $0 \le \delta < 1$, $\pi(d)$ is a deterministic function of d that grows slowly in d, such as $\pi(d) = 1$ or $\ln(d)$, and C is a positive constant.

(A2) The factor model (4) has r fixed factors, with $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$, and matrices $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_f$ satisfy

$$||\mathbf{\Sigma}_0||_2 < \infty$$
$$\max_{1\le t\le T} |\mathbf{\Sigma}_{f,jj}(t)| = \mathcal{O}_d(B(T)), \quad j = 1, \ldots, r, \tag{15}$$

where $1 \le B(T) = o(T)$.

(A3) $\max_{1\le t\le T} \| \hat{\mathbf{\Sigma}}_x(t) - \mathbf{\Sigma}_x(t) \|_2 = \mathcal{O}_d(A(d,T,n))$, for some rate function $A(d,T,n)$ such that $A(d,T,n)B^5(T) = o(1)$ with $B(T)$ defined in equation (A2).

(A4) A_j's and B_i's are such that the series $\mathbf{S}_f(t)$ under the CAW model (9) and (10) are stationary and ergodic.

(A5) The parameter set Θ for all parameters A_j's, B_i's, C and ν is compact for the CAW model (9) and (10).

(A6) The Hessian matrix $\partial^2\mathcal{L}(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j(\partial^2\hat{\mathcal{L}}(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j)$ converges to some deterministic matrix function $D(\boldsymbol{\theta})(\hat{D}(\boldsymbol{\theta}))$ as T goes to infinity which is of full rank for all $\boldsymbol{\theta} \in \Theta$.

The first two conditions are used to prove the consistency of $\hat{\mathbf{\Sigma}}_f(t)$. The assumption (A3) for the threshold MSRVM estimator follows Tao *et al.* (2013) where we take $A(d,T,n) = \pi(d)[e_n(d^2T)^{1/\beta}]^{1-\delta}\ln T$ and $B(T) = \ln T$, with $e_n \sim n^{-1/4}$. Here, β relates to the moment conditions on the microstructure noise which can be taken large enough so that $A(d,T,n)B^5(T)$ will go to 0 as n, d and T go to infinity.

THEOREM 1 *Suppose the models* (1), (3) *and* (4) *satisfy Conditions (A1) –(A3). Denote the ordered eigenvalues of* $\bar{\tilde{\mathbf{S}}}_x$ *by* $\lambda_1 \ge \cdots \ge \lambda_d$. *Let* $\mathbf{a}_1, \ldots, \mathbf{a}_r$ *be the eigenvectors of* $\bar{\tilde{\mathbf{S}}}_x$ *corresponding to the r largest eigenvalues* $\lambda_1, \ldots, \lambda_r$. *Also set* $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_r$ *be the r largest eigenvalues of* $\bar{\tilde{\mathbf{S}}}_x$ *and* $\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_r$ *the corresponding eigenvectors. Let* $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_r)$ *and* $\hat{\mathbf{A}} = (\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_r)$. *As n, d, T go to infinity, we have*

$$||\mathbf{A}'\hat{\mathbf{A}} - \mathbf{I}_r||_2 = \mathcal{O}_d(A(d,T,n)B(T)),$$
$$||\hat{\mathbf{\Sigma}}_f(t) - \mathbf{\Sigma}_f(t) - \mathbf{A}'\mathbf{\Sigma}_0\mathbf{A}||_2 = \mathcal{O}_d(A^{1/2}(d,T,n)B^{3/2}(T)). \tag{16}$$

THEOREM 2 *Suppose that* $\hat{\boldsymbol{\theta}}$ *is the maximized log-likelihood estimator of* $\boldsymbol{\theta}$ *based on the data* $\hat{\mathbf{\Sigma}}_f(t)$ *from the CAW model and* $\tilde{\boldsymbol{\theta}}$ *is the maximized log-likelihood estimator based on the true data* $\tilde{\mathbf{\Sigma}}_f(t)$ *from the same CAW model. Then under Conditions (A1) –(A6),*

$$\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} = \mathcal{O}_d(A^{1/4}(d,T,n)B^{5/4}(T)). \tag{17}$$

REMARK 1 Suppose $d \sim T$, $n = T^{\gamma}$ as they go to infinity, and $\delta = 1/2$, it can be shown that $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} \sim \mathcal{O}_d(T^{-(\gamma/4-3/\beta)/8})$. For large enough β, we have $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} \sim \mathcal{O}_d(T^{-\gamma/32})$, which goes to zero when T goes to infinity.

4. Empirical data analysis

We apply the proposed methodology to modelilng 30 stocks traded at the New York Stock Exchange (NYSE) in this section.

4.1. Data description

We use 30 stocks traded at NYSE, which consist of 27 components of Dow Jones Index: 3M (MMM), American Express (AXP), AT&T (T), Boeing (BA), Caterpillar (CAT), Chevron (CVX), Coca-Cola (KO), Dupont (DD), ExxonMobil (XOM), General Electric (GE), Goldman Sachs (GS), The Home Depot (HD), IBM (IBM), Johnson & Johnson (JNJ), JPMorgan Chase (JPM), McDonald's (MCD), Merck (MRK), Nike (NKE), Pfizer (PFE), Procter & Gamble (PG), Travelers (TRV), UnitedHealth Group (UNH), United Technologies (UTX), Verizon (VZ), Visa (V), Wal-Mart (WMT) and Walt Disney (DIS), and 3 former components of Dow Jones Index: Honeywell (HON), Citigroup (C) and American International Group (AIG). The raw tick-by-tick trading data are downloaded from TAQ database of Wharton Research Data Service. The data period starts at 3 January 2012 and ends on 31 December 2012, with totally 250 trading days.

We firstly conduct data cleaning with the procedures introduced in Brownlees and Gallo (2006) and Barndorff-Nielsen *et al.* (2009). The steps are the following:

1. Delete entries with a time stamp outside 9:30 am–4:00 pm when the exchange is open.
2. Delete entries with a time stamp inside 9: 30–10:00 am or 3: 30–4:00 pm to eliminate the open and end effect of price fluctuation.
3. Delete entries with the transaction price equal to zero.
4. If multiple transactions have the same time stamp, use the median price.

5. Delete entries with prices which are outliers. Let $\{p_i\}_{i=1}^{N}$ be an ordered tick-by-tick price series. We treat the i-th price as an outlier if $|p_i - \bar{p}_i(m)| > 3s_i(m)$, where $\bar{p}_i(m)$ and $s_i(m)$ denote the sample mean and sample standard deviation of a neighbourhood of m observations around i, respectively. For the beginning prices which may not have enough left hand side neighbours, we get $m-i$ neighbours from $i+1$ to $m+1$. Similar procedures are taken for the ending prices. We take $m=5$ here.

Then, we construct the threshold MSRVM estimator based on the cleaned tick-by-tick data following the steps in Tao *et al.* (2013). We set the threshold to be 5% of the largest of the absolute value of entries in the matrix. This generates a series of 250 matrices of $\hat{\mathbf{\Sigma}}_x(t)$, which are 30 by 30. Descriptive statistics of selected realized variances and covariances are provided in table 1.

In addition, two plots for realized variances and two plots for realized covariances are shown in figure 1. We find the following properties:

1. All 30 realized variances and 435 covariances are skewed to the right, with mean skewness 1.39.
2. All realized variances and covariances have bigger kurtosis than that of the normal distribution, with mean kurtosis 5.92 showing fat tails.

Table 1. Descriptive statistics for some selected realized variances and covariances.

Stock	Mean $*10^{-5}$	Maximum $*10^{-4}$	Minimum $*10^{-5}$	SD $*10^{-5}$	Skewness	Kurtosis
			Realized variance			
AIG	17.4	9.16	3.05	11.2	2.47	13.1
AXP	6.12	6.38	1.45	4.71	7.67	91.5
BA	5.46	2.19	1.11	3.25	1.70	7.33
C	16.7	9.62	2.80	10.9	2.56	15.3
			Realized covariance			
AIG-AXP	4.30	1.65	−1.99	3.13	1.00	3.76
AIG-BA	3.32	1.32	−2.74	2.82	1.03	3.93
AIG-C	7.72	4.54	−1.47	5.99	1.85	9.37
AXP-BA	2.41	1.01	−0.59	1.95	1.26	4.37
AXP-C	5.03	2.24	−0.75	3.53	1.61	6.84
BA-C	3.69	1.75	−1.40	3.13	1.66	6.30

Note: We report the descriptive statistics of the realized variances and covariances of the dataset, namely, mean, maximum, minimum, standard deviation, skewness and kurtosis. We only show some entries due to limited space.

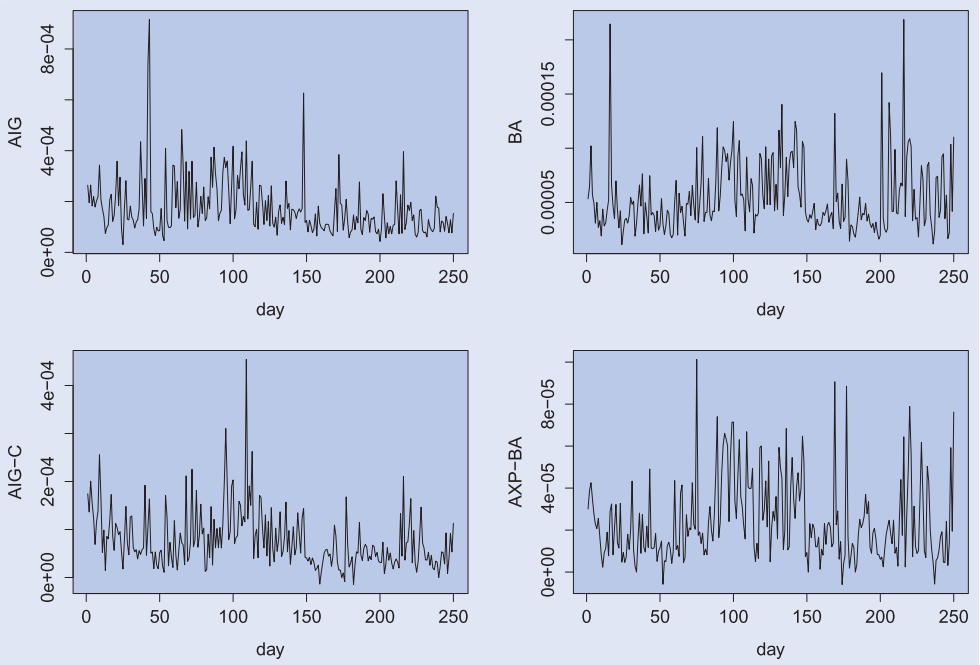

Figure 1. The plot of selected realized variances and covariances. Two plots for realized variances and two plots for realized covariances are shown in figure 1. AIG and BA are chosen for the realized variances while the realized covariances between AIG and C, and between AXP and BA are shown in the following figures.

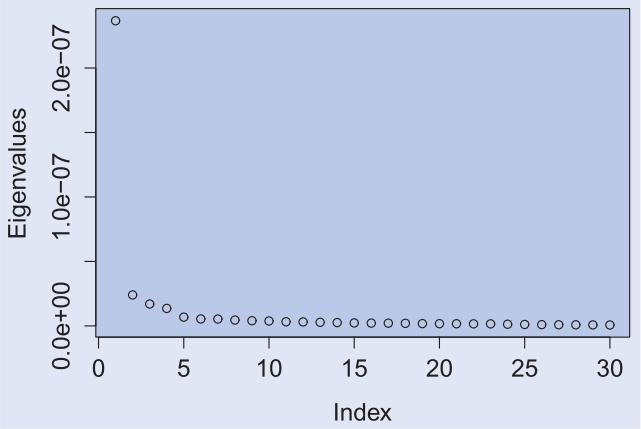

Figure 2. Plots of the eigenvalues of $\bar{\hat{\mathbf{S}}}_x$ for the dataset when $k = 220$.

3. The realized variances and covariances have significant fluctuations during the year, indicated by the graphs.

For the next two subsections, we show the estimated results for both the diagonal CAW and VAR models when the first $k = 220$ days are treated as data points.

4.2. *Model fitting*

The eigenvalues of the sample variance matrix $\bar{\hat{\mathbf{S}}}_x$ are evaluated and are shown in figure 2. We choose four factors for the model, as there is a discernible drop between the fourth and the fifth eigenvalue, though the second to fourth eigenvalues are much less than the biggest one.

Let $\hat{\mathbf{A}}$ be the eigenvectors of $\bar{\hat{\mathbf{S}}}_x$ corresponding to the four largest eigenvalues. We calculate the factor volatility matrices $\hat{\mathbf{\Sigma}}_f(t)$, which are 4 by 4. Then, we fit the diagonal CAW model to the matrix series. Different orders are used to make comparison, namely $(p, q) = (0, 1)$, $(p, q) = (1, 1)$, $(p, q) = (1, 2)$, $(p, q) = (2, 1)$ and $(p, q) = (2, 2)$. For every estimation, we randomly choose 60 initial values for each optimization of the log-likelihood and choose the one with the largest log-likelihood value to give the estimated parameters.

4.3. *VAR model*

For comparison purposes, we fit the VAR model to the vectorized factor covariance matrix $\hat{\mathbf{\Sigma}}_f(t)$, which is a vector with 10 entries. Using the package "vars" in R, we select the VAR(1) model as all the model selection criteria, namely Akaike information criterion (AIC), Hannan Quinn (HQ), Schwarz criterion (SC) and final prediction error (FPE), choose the order 1, as shown in table 2.

The model is

$$\text{vech}\{\tilde{\mathbf{\Sigma}}_f(t)\} = \mathbf{A}_0 + \mathbf{A}_1 \text{vech}\{\tilde{\mathbf{\Sigma}}_f(t-1)\} + \mathbf{e}(t),$$

where $\mathbf{A}_0$ is a 10-dimensional vector, $\mathbf{A}_1$ is a 10×10 square matrix and $\mathbf{e}(t)$ is a 10-dimensional vector white noise process with zero mean and finite fourth moments. Both $\mathbf{A}_0$ and $\mathbf{A}_1$ are estimated by the least squares method.

We denote the estimator of $\mathbf{A}_1$ by $\hat{\mathbf{A}}_1$. We find that $|\hat{\mathbf{A}}_1| = -1.03 \times 10^{-7}$ and the biggest absolute eigenvalue is $0.4906 < 1$ which ensures the stationarity of the VAR model.

4.4. *Performance comparison in out-of-sample forecasting*

We compare the out-of-sample one-day-ahead forecast performance of the diagonal CAW and other models including unrestricted VAR(1), diagonal VAR(1) and Exponentially Weighted Moving Average (EWMA) models. For the EWMA model, we set $\lambda = 0.94$. The prediction of the one-day-ahead realized covariance is calculated by: (1) predict the one-day-ahead factor covariance matrix by conditional expectation and (2) plug the forecast factor covariance matrix into the factor model (4) to get the predicted realized covariance matrix.

The predictive accuracy is measured with both the Frobenius norm and the spectral norm. We take the first k days as data and forecast the next day, where $k = 220, \ldots, 249$. Every model is re-estimated and the new forecasts are generated based upon the new parameter estimates. Then, we take the average of errors during 30 periods to do the comparison. Table 3 contains the results of the prediction error of four models using different norms. Here, FN stands for the Frobenius norm and SN for the spectral norm.

The main findings are as follows.

1. The diagonal CAW models with order $(p, q) = (1, 1)$ performs the best among the CAW models as it has the smallest error under both Frobenius and spectral norms. In general, all CAW models have similar performance except the ones with order $(p, q) = (0, 1)$, which indicates the possibility of over-parameterization with orders larger than $(1, 1)$.
2. The diagonal CAW(1) model have slightly better performance compared with that of VAR(1) model.
3. The diagonal CAW model needs far less parameters than the VAR model does. Here, the best CAW model

Table 2. The selection of the order of VAR model.

Oder	1	2	3	4	5
AIC(n) $*10^2$	− 1.949	− 1.943	− 1.937	− 1.933	− 1.932
HQ(n) $*10^2$	− 1.939	− 1.925	− 1.910	− 1.898	− 1.888
SC(n) $*10^2$	− 1.924	− 1.897	− 1.871	− 1.846	− 1.824
FPE(n) $*10^{-84}$	0.234	0.416	0.796	1.292	1.751

Note: We fit the VAR model to the vectorized factor covariance matrix $\hat{\mathbf{\Sigma}}_f(t)$, which is a vector with 10 entries. Using the package 'vars' in R, we select the VAR(1) model as all the model selection criteria, namely Akaike information criterion (AIC), Hannan Quinn (HQ), Schwarz criterion (SC) and final prediction error (FPE), choose the order 1.

Table 3. Forecast errors for CAW and other models using different norms.

Order	CAW Number of parameters	CAW FN $*10^{-4}$	CAW SN $*10^{-4}$
$p=0, q=1$	9	6.10	5.58
$p=1, q=1$	13	5.27	4.80
$p=1, q=2$	17	5.28	4.81
$p=2, q=1$	17	5.28	4.83
$p=2, q=2$	21	5.28	4.82
	Others		
	Number of parameters	FN $*10^{-4}$	SN $*10^{-4}$
VAR(1)	110	5.27	4.85
Diagonal VAR(1)	20	5.41	4.95
EWMA	0	5.65	5.27

Note: We report the results of the prediction accuracy of the two models using different norms. Here, FN is for the Frobenius norm and SN for the spectral norm.

with order $(p, q) = (1, 1)$ only needs 13 parameters, nearly a tenth of the number of parameters that VAR(1) model needs.

4. We do predictions for 2–5 days ahead in addition to one-day forecast. We find that the performance of predictions of longer horizons is similar with that for the one day, in general.
5. The diagonal VAR(1) model has even less parameters than the diagonal CAW model with $(p, q) = (1, 1)$, but with worse prediction accuracy.
6. The EWMA model does not need any parameters, but its performance is nearly the worst among all models by only beating CAW with $(p, q) = (0, 1)$.

As a result, we conclude that the diagonal CAW model with order $(p, q) = (1, 1)$ is recommended compared to the VAR model.

5. Conclusions

In the literature, most models dealing with the realized covariance matrix focus on a small number of assets, which become unfeasible when the dimension is large. In order to solve the problem, we propose a factor model with diagonal CAW model fitted to the factor covariance matrix. Our model performs comparably with the VAR model while requiring far less parameters. For example, in the data analysis, the CAW model with order $(p, q) = (1, 1)$ performs similarly with the VAR model, measured both in the Frobenius norm and in the spectral norm, but only needs nearly one-tenth of the number of parameters of the latter. In addition, the model ensures the positive definiteness for the predicted covariance matrices, which solves the problem inherent in the VAR model proposed by Tao *et al*. (2011). Notice that the additive microstructure model (3) might be not that realistic in practice, for example, the model ignores the discreteness of asset price (Li *et al.* 2018). The reason for its adoption is mainly for simplicity of theoretical analysis. Thus, it would be valuable to extend the methodology of the paper by accommodating other microstructure noise. This would indeed be possible as far as the microstructure noise fulfils the requirement that the difference between the RCov and ICV is small (see Assumption (A3)). Finally, diagnostic tests for the proposed model is worth considering in future.

Acknowledgments

The authors are grateful to a referee for careful reading and critical comments.

Funding

This research was supported by Hong Kong Research Grants Council General Research Fund [grant numbers 17303315 and 17332416].

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P., Modeling and forecasting realized volatility. *Econometrica*, 2003, **71**, 579–625.

Asai, M. and McAleer, M.J., Forecasting co-volatilities via factor models with asymmetry and long memory in realized covariance (No. TI 2014-037/III). Tinbergen Institute Discussion Paper Series, 2014.

Bannouh, K., Martens, M.P.E., Oomen, R.C. and Van Dijk, D.J., Realized mixed-frequency factor models for vast dimensional covariance estimation (No. ERS-2012-017-F&A). ERIM Report Series Research in Management, 2012.

Barndorff-Nielsen, O.E. and Shephard, N., Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica*, 2004, **72**, 885–925.

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A. and Shephard, N., Realized kernels in practice: Trades and quotes. *J. Econom.*, 2009, **12**(3), C1–C32.

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A. and 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**(2), 149–169.

Bickel, P.J. and Levina, E., Regularized estimation of large covariance matrices. *Ann. Stat.*, 2008, **36**(1), 199–227.

Bollerslev, T., Engle, R.F. and Wooldridge, J., A capital asset pricing model with time varying covariances. *J. Political Econ.*, 1988, **95**, 116–131.

Brownlees, C.T. and Gallo, G.M., Financial econometric analysis at ultra-high frequency: Data handling concerns. *Comput. Stat. Data. Anal.*, 2006, **51**(4), 2232–2245.

Christensen, K., Kinnebrock, S. and Podolskij, M., Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *J. Econom.*, 2010, **159**(1), 116–133.

Engle, R.F. and Kroner, K.F., Multivariate simultaneous generalized arch. *Econ. Theory*, 1995, **11**, 122–150.
Fan, J. and Wang, Y., Multi-scale jump and volatility analysis for high-frequency financial data. *J. Am. Stat. Assoc.*, 2007, **102**(480), 1349–1362.
Golosnoy, V., Gribisch, B. and Liesenfeld, R., The conditional autoregressive Wishart model for multivariate stock market volatility. *J. Econom.*, 2012, **167**(1), 211–223.
Gourieoux, C., Jasiak, J. and Sufana, R., The Wishart autoregressive process of multivariate stochastic volatility. *J. Econom.*, 2009, **150**(2), 167–181.
Hayashi, T. and Yoshida, N., On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 2005, **11**(2), 359–379.
Jin, X. and Maheu, J.M., Modelling realized covariances. Working Paper 382, Department of Economics, University of Toronto, 2009.
Johnstone, I.M. and Lu, A.Y., On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Stat. Assoc.*, 2009, **104**(486), 682–693.
Li, Y., Zhang, Z. and Li, Y., A unified approach to volatility estimation in the presence of both rounding and random market microstructure noise. *J. Econom.*, 2018, **203**(2), 187–222.
Markowitz, H.M., Portfolio selection. *J. Finance*, 1952, **7**(1), 77–91.
McCurdy, T.H. and Stengos, T., A comparison of risk-premium forecasts implied by parametric versus nonparametric conditional mean estimators. *J. Econom.*, 1992, **52**(1), 225–244.
Tao, M., Wang, Y., Yao, Q. and Zou, J., Large volatility matrix inference via combining low-frequency and high-frequency approaches. *J. Am. Stat. Assoc.*, 2011, **106**(495), 1025–1040.
Tao, M., Wang, Y. and Chen, X., Fast convergence rates in estimating large volatility matrices using high-frequency financial data. *Econ. Theory*, 2013, **29**(04), 838–856.
Wang, Y. and Zou, J., Vast volatility matrix estimation for high-frequency financial data. *Ann. Stat.*, 2010, **38**(2), 943–978.
Yu, P.L.H., Li, W.K. and Ng, F.C., The generalized conditional autoregressive Wishart model for multivariate realized volatility. *J. Bus. Econ. Stat.*, 2017, **35**(04), 513–527.
Zhang, L., Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli*, 2006, **12**(6), 1019–1043.
Zhang, L., Estimating covariation: Epps effect, microstructure noise. *J. Econom.*, 2011, **160**(1), 33–47.

Appendix. Proof of theorems

For convenience, we denote $A(d,T,n)$ and $B(T)$ by A and B in the proof part.

Proof of Theorem 1 Following Theorem 1 in Tao *et al.* (2011), we can easily show that $\| \bar{\hat{\mathbf{S}}}_x - \bar{\mathbf{S}}_x \|_2 = \mathcal{O}_p(AB)$.

Then, we claim that

$$\max_{1\le j\le r} | \hat{\lambda}_j - \lambda_j | = \mathcal{O}_p(AB), \tag{A1}$$

$$\max_{1\le j\le r} \| \hat{\mathbf{a}}_j - \mathbf{a}_j \|_2 = \mathcal{O}_p(AB). \tag{A2}$$

Let $P = \bar{\hat{\mathbf{S}}}_x - \bar{\mathbf{S}}_x$ with ordered eigenvalues $\rho_1 \ge \cdots \ge \rho_d$. Then, we have $\rho_d \le \hat{\lambda}_j - \lambda_j \le \rho_1$. As a result,

$$| \hat{\lambda}_j - \lambda_j | \le \max(| \rho_1 |, | \rho_d |) = \| \bar{\hat{\mathbf{S}}}_x - \bar{\mathbf{S}}_x \|_2 , \tag{A3}$$

which proves equation (A1). Equation (A2) follows from Theorem 1 in Bickel and Levina (2008) and the same argument in the proof of Theorem 5 in the same paper.

For the j-th diagonal entry of $\mathbf{A}'\hat{\mathbf{A}} - \mathbf{I}_r$,

$$\mathbf{a}_j'\hat{\mathbf{a}}_j - 1 = -\| \hat{\mathbf{a}}_j - \mathbf{a}_j \|_2^2 /2 = \mathcal{O}_p(A^2B^2) \le \mathcal{O}_p(AB) \tag{A4}$$

as we assume AB goes to zero. For off-diagonal entry (k,j) $(k \ne j)$,

$$\begin{aligned} | \mathbf{a}_k'\hat{\mathbf{a}}_j | &= | \mathbf{a}_k'(\hat{\mathbf{a}}_j - \mathbf{a}_j) | \\ &\le \| \mathbf{a}_k' \|_2 \| \hat{\mathbf{a}}_j - \mathbf{a}_j \|_2 = \| \hat{\mathbf{a}}_j - \mathbf{a}_j \|_2 \\ &= \mathcal{O}_p(AB), \end{aligned} \tag{A5}$$

which proves the first result.

To prove the second result, we separate the left-hand side of the equation into three parts:

$$\begin{aligned} &\hat{\mathbf{\Sigma}}_f(t) - \mathbf{\Sigma}_f(t) - \mathbf{A}'\mathbf{\Sigma}_0\mathbf{A} \\ &= \hat{\mathbf{A}}'[\hat{\mathbf{\Sigma}}_x(t) - \mathbf{\Sigma}_x(t)]\hat{\mathbf{A}} + \hat{\mathbf{A}}'\mathbf{\Sigma}_x(t)\hat{\mathbf{A}} - \mathbf{\Sigma}_f(t) - \mathbf{A}'\mathbf{\Sigma}_0\mathbf{A} \\ &= \hat{\mathbf{A}}'[\hat{\mathbf{\Sigma}}_x(t) - \mathbf{\Sigma}_x(t)]\hat{\mathbf{A}} + [(\mathbf{A}'\hat{\mathbf{A}})'\mathbf{\Sigma}_f(t)\mathbf{A}'\hat{\mathbf{A}} - \mathbf{\Sigma}_f(t)] \\ &\quad + [\hat{\mathbf{A}}'\mathbf{\Sigma}_0\hat{\mathbf{A}} - \mathbf{A}'\mathbf{\Sigma}_0\mathbf{A}]. \end{aligned} \tag{A6}$$

For the first term on the right-hand side of equation (A6), since

$$\| \hat{\mathbf{A}}' \|_2^2, \| \hat{\mathbf{A}} \|_2^2 = 1, \tag{A7}$$

we have

$$\begin{aligned} \| \hat{\mathbf{A}}'[\hat{\mathbf{\Sigma}}_x(t) - \mathbf{\Sigma}_x(t)]\hat{\mathbf{A}} \|_2 &\le \| \hat{\mathbf{A}}' \|_2 \| \hat{\mathbf{\Sigma}}_x(t) - \mathbf{\Sigma}_x(t) \|_2 \| \hat{\mathbf{A}} \|_2 \\ &= \mathcal{O}_p(A). \end{aligned} \tag{A8}$$

For the second term, from Condition (A2), we know that $\| \mathbf{\Sigma}_f(t) \|_2 = \mathcal{O}_p(B)$, and we have

$$\begin{aligned} \| \hat{\mathbf{A}} - \mathbf{A} \|_2^2 &= \| (\hat{\mathbf{A}}' - \mathbf{A}')(\hat{\mathbf{A}} - \mathbf{A}) \|_2 \\ &\le 2 \| \mathbf{A}'\hat{\mathbf{A}} - \mathbf{I}_r \|_2 = \mathcal{O}_p(AB). \end{aligned} \tag{A9}$$

As a result,

$$\begin{aligned} &\| (\mathbf{A}'\hat{\mathbf{A}})'\mathbf{\Sigma}_f(t)\mathbf{A}'\hat{\mathbf{A}} - \mathbf{\Sigma}_f(t) \|_2 \\ &= \| \hat{\mathbf{A}}'\mathbf{A}\mathbf{\Sigma}_f(t)\mathbf{A}'\hat{\mathbf{A}} - \hat{\mathbf{A}}'\hat{\mathbf{A}}\mathbf{\Sigma}_f(t)\hat{\mathbf{A}}'\hat{\mathbf{A}} \|_2 \\ &\le \| \hat{\mathbf{A}}' \|_2 \| \mathbf{A}\mathbf{\Sigma}_f(t)\mathbf{A}' - \hat{\mathbf{A}}\mathbf{\Sigma}_f(t)\hat{\mathbf{A}}' \|_2 \| \hat{\mathbf{A}} \|_2 \\ &\le \| (\hat{\mathbf{A}} - \mathbf{A})\mathbf{\Sigma}_f(t)\hat{\mathbf{A}}' + \mathbf{A}\mathbf{\Sigma}_f(t)(\hat{\mathbf{A}} - \mathbf{A})' \|_2 \\ &\le \| \hat{\mathbf{A}} - \mathbf{A} \|_2 \| \mathbf{\Sigma}_f(t) \|_2 (\| \hat{\mathbf{A}} \|_2 + \| \mathbf{A} \|_2) \\ &= \mathcal{O}_p(A^{1/2}B^{3/2}). \end{aligned} \tag{A10}$$

For the third term, again from Condition (A2), we know that $\|\mathbf{\Sigma}_0\|_2$ is bounded, therefore,

$$\begin{aligned} &\| \hat{\mathbf{A}}'\mathbf{\Sigma}_0\hat{\mathbf{A}} - \mathbf{A}'\mathbf{\Sigma}_0\mathbf{A} \|_2 \\ &= \| (\hat{\mathbf{A}} - \mathbf{A})'\mathbf{\Sigma}_0\hat{\mathbf{A}} + \mathbf{A}'\mathbf{\Sigma}_0(\hat{\mathbf{A}} - \mathbf{A}) \|_2 \\ &\le \| (\hat{\mathbf{A}} - \mathbf{A})' \|_2 \| \mathbf{\Sigma}_0 \|_2 \| \hat{\mathbf{A}} \|_2 + \| \mathbf{A}' \|_2 \| \mathbf{\Sigma}_0 \|_2 \| (\hat{\mathbf{A}} - \mathbf{A}) \|_2 \\ &= \| \hat{\mathbf{A}} - \mathbf{A} \|_2 \| \mathbf{\Sigma}_0 \|_2 (\| \hat{\mathbf{A}} \|_2 + \| \mathbf{A} \|_2) \\ &= \mathcal{O}_p(A^{1/2}B^{1/2}). \end{aligned} \tag{A11}$$

From equations (A8), (A10) and (A11), we conclude that

$$\hat{\mathbf{\Sigma}}_f(t) - \mathbf{\Sigma}_f(t) - \mathbf{A}'\mathbf{\Sigma}_0\mathbf{A} = \mathcal{O}_p(A^{1/2}B^{3/2}). \tag{A12}$$

■

The following lemma is needed to prove Theorem 2.

LEMMA 1 *The log-likelihood function for the CAW model given observed data* $\hat{\mathbf{\Sigma}}_f(t)$ *and true data* $\tilde{\mathbf{\Sigma}}_f(t)$ *are*

$$\hat{\mathcal{L}}(\theta) = \frac{1}{T}\sum_{t=1}^{T}\left\{-\frac{\nu r}{2}\ln(2) - \frac{r(r-1)}{4}\ln(\pi) - \sum_{i=1}^{r}\ln\Gamma\left(\frac{\nu+1-i}{2}\right) - \frac{\nu}{2}\ln\left|\frac{\hat{\mathbf{S}}_f(t)}{\nu}\right| + \left(\frac{\nu-r-1}{2}\right)\ln\left|\hat{\mathbf{\Sigma}}_f(t)\right| - \frac{1}{2}\mathrm{tr}(\nu\hat{\mathbf{S}}_f(t)^{-1}\hat{\mathbf{\Sigma}}_f(t))\right\} \tag{A13}$$

and

$$\mathcal{L}(\theta) = \frac{1}{T}\sum_{t=1}^{T}\left\{-\frac{\nu r}{2}\ln(2) - \frac{r(r-1)}{4}\ln(\pi) - \sum_{i=1}^{r}\ln\Gamma\left(\frac{\nu+1-i}{2}\right) - \frac{\nu}{2}\ln\left|\frac{\mathbf{S}_f(t)}{\nu}\right| + \left(\frac{\nu-r-1}{2}\right)\ln|\tilde{\mathbf{\Sigma}}_f(t)| - \frac{1}{2}\mathrm{tr}(\nu\mathbf{S}_f(t)^{-1}\tilde{\mathbf{\Sigma}}_f(t))\right\}, \tag{A14}$$

respectively. Then,

$$\hat{\mathcal{L}}(\theta) - \mathcal{L}(\theta) = \mathcal{O}_p(A^{1/2}B^{5/2}). \tag{A15}$$

Proof By simple algebraic manipulations,

$$\hat{\mathcal{L}}(\theta) = \mathcal{L}(\theta) + \frac{1}{T}\sum_{t=1}^{T}\left\{-\frac{\nu}{2}\left(\ln\left|\frac{\hat{\mathbf{S}}_f(t)}{\mathbf{S}_f(t)}\right|\right) + \left(\frac{\nu-r-1}{2}\right)\ln\left|\frac{\hat{\mathbf{\Sigma}}_f(t)}{\tilde{\mathbf{\Sigma}}_f(t)}\right| - \frac{1}{2}\mathrm{tr}(\nu(\hat{\mathbf{S}}_f(t)^{-1}\hat{\mathbf{\Sigma}}_f(t) - \mathbf{S}_f(t)^{-1}\tilde{\mathbf{\Sigma}}_f(t)))\right\}. \tag{A16}$$

Since $\mathbf{S}_f(t)$ is the conditional mean of $\tilde{\mathbf{\Sigma}}_f(t)$, it can be proved that $\max_t \| \mathbf{S}_f(t) \|_2 = \mathcal{O}_p(B)$ and $\hat{\mathbf{S}}_f(t) - \mathbf{S}_f(t) = \mathcal{O}_p(A^{1/2}B^{3/2})$ as $\tilde{\mathbf{\Sigma}}_f(t)$ and $\hat{\mathbf{\Sigma}}_f(t) - \tilde{\mathbf{\Sigma}}_f(t)$. The basic idea of the proof is the following. Equation (10) can be treated as a linear map from $(\tilde{\mathbf{\Sigma}}_f(t))$ to $(\mathbf{S}_f(t))$: $(\mathbf{S}_f(t)) = \mathcal{L}\cdot(\tilde{\mathbf{\Sigma}}_f(t))$, where $\mathcal{L}$ is a linear operator with bounded norm. It follows that $\forall t, \| \mathbf{S}_f(t) \|_2 \leq \alpha \sup_s \| \tilde{\mathbf{\Sigma}}_f(t) \|_2$ for some constant α, which gives the required bound. A similar argument can be applied to $\hat{\mathbf{S}}_f(t) - \mathbf{S}_f(t)$.

For the first term of the additional part, since $\mathbf{S}_f(t) = \mathcal{O}_p(B)$,

$$\left\|\frac{\hat{\mathbf{S}}_f(t) - \mathbf{S}_f(t)}{\mathbf{S}_f(t)}\right\|_2 = \mathcal{O}_p(A^{1/2}B^{1/2}). \tag{A17}$$

Thus,

$$\ln\left|\frac{\hat{\mathbf{S}}_f(t)}{\mathbf{S}_f(t)}\right| = \left(\left|\frac{\hat{\mathbf{S}}_f(t)}{\mathbf{S}_f(t)}\right| - 1\right) - o\left(\left|\frac{\hat{\mathbf{S}}_f(t)}{\mathbf{S}_f(t)}\right| - 1\right) = \mathcal{O}_p(A^{1/2}B^{1/2}). \tag{A18}$$

For the second term, it is easy to prove that $\tilde{\mathbf{\Sigma}}_f(t) = \mathcal{O}_p(B)$, so that

$$\left\|\frac{\hat{\mathbf{\Sigma}}_f(t) - \tilde{\mathbf{\Sigma}}_f(t)}{\tilde{\mathbf{\Sigma}}_f(t)}\right\|_2 = \mathcal{O}_p(A^{1/2}B^{1/2}), \tag{A19}$$

which leads to

$$\ln\left|\frac{\hat{\mathbf{\Sigma}}_f(t)}{\tilde{\mathbf{\Sigma}}_f(t)}\right| = \left(\left|\frac{\hat{\mathbf{\Sigma}}_f(t)}{\tilde{\mathbf{\Sigma}}_f(t)}\right| - 1\right) - o\left(\left|\frac{\hat{\mathbf{\Sigma}}_f(t)}{\tilde{\mathbf{\Sigma}}_f(t)}\right| - 1\right) = \mathcal{O}_p(A^{1/2}B^{1/2}). \tag{A20}$$

For the third term, first note that as $\mathbf{S}_f(t) \geq \mathbf{C}\mathbf{C}^{\mathrm{T}}$, there exists a constant $w > 0$ such that the minimum eigenvalue of $\mathbf{S}_f(t)$ is larger than or equal to w. Consequently, $\|\mathbf{S}_f(t)^{-1}\|_2 \leq 1/w$ which is bounded. As a result,

$$\begin{aligned}
&\|\hat{\mathbf{S}}_f(t)^{-1}\hat{\mathbf{\Sigma}}_f(t) - \mathbf{S}_f(t)^{-1}\tilde{\mathbf{\Sigma}}_f(t)\|_2 \\
&\quad\leq \|\hat{\mathbf{S}}_f(t)^{-1}(\hat{\mathbf{\Sigma}}_f(t) - \tilde{\mathbf{\Sigma}}_f(t))\|_2 + \|(\hat{\mathbf{S}}_f(t)^{-1} - \mathbf{S}_f(t)^{-1})\tilde{\mathbf{\Sigma}}_f(t)\|_2 \\
&\quad\leq \|\hat{\mathbf{S}}_f(t)^{-1}\|_2\|(\hat{\mathbf{\Sigma}}_f(t) - \tilde{\mathbf{\Sigma}}_f(t))\|_2 + \|\hat{\mathbf{S}}_f(t)^{-1}\|_2\| \\
&\quad\quad \hat{\mathbf{S}}_f(t) - \mathbf{S}_f(t)\|_2\|\mathbf{S}_f(t)^{-1}\|_2\|\tilde{\mathbf{\Sigma}}_f(t)\|_2 \\
&\quad= \frac{1}{w}\mathcal{O}_p(A^{1/2}B^{3/2}) + \frac{1}{w^2}\mathcal{O}_p(A^{1/2}B^{3/2})\mathcal{O}_p(B) \\
&\quad= \mathcal{O}_p(A^{1/2}B^{5/2}).
\end{aligned} \tag{A21}$$

■

Proof From Lemma 1,

$$\hat{\mathcal{L}}(\hat{\theta}) - \mathcal{L}(\hat{\theta}) = \mathcal{O}_p(A^{1/2}B^{5/2}) \tag{A22}$$

and

$$\hat{\mathcal{L}}(\tilde{\theta}) - \mathcal{L}(\tilde{\theta}) = \mathcal{O}_p(A^{1/2}B^{5/2}). \tag{A23}$$

Then,

$$(\hat{\mathcal{L}}(\tilde{\theta}) - \hat{\mathcal{L}}(\hat{\theta})) + (\mathcal{L}(\hat{\theta}) - \mathcal{L}(\tilde{\theta})) \leq \mathcal{O}_p(A^{1/2}B^{5/2}). \tag{A24}$$

$$\begin{aligned}
\text{LHS of (A24)} = &\sum_i \frac{\partial\hat{\mathcal{L}}(\hat{\boldsymbol{\theta}})}{\partial\theta_i}(\tilde{\theta} - \hat{\theta}) + \sum_i\sum_j \frac{\partial^2\hat{\mathcal{L}}}{2\partial\theta_i\partial\theta_j}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^2 \\
&+ \sum_i \frac{\partial\mathcal{L}(\tilde{\boldsymbol{\theta}})}{\partial\theta_i}(\hat{\theta} - \tilde{\theta}) \\
&+ \sum_i\sum_j \frac{\partial^2\mathcal{L}}{2\partial\theta_i\partial\theta_j}(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^2 + o(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^2.
\end{aligned} \tag{A25}$$

It is known that $\partial\hat{\mathcal{L}}(\hat{\boldsymbol{\theta}})/\partial\theta_i = 0$ and $\partial\mathcal{L}(\tilde{\boldsymbol{\theta}})/\partial\theta_i = 0$. In addition, from Condition A6, there exist some constants $\tilde{\ell}$ and $\hat{\ell}$ such that uniformly,

$$\left|\frac{\partial^2\mathcal{L}}{\partial\theta_i\partial\theta_j}(\tilde{\boldsymbol{\theta}})\right| > \tilde{\ell} > 0, \tag{A26}$$

$$\left|\frac{\partial^2\hat{\mathcal{L}}}{\partial\theta_i\partial\theta_j}(\hat{\boldsymbol{\theta}})\right| > \hat{\ell} > 0. \tag{A27}$$

As a result,

$$\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} = \mathcal{O}_p(A^{1/4}B^{5/4}). \tag{A28}$$

■