Article:
Yang, Jin, Lian, Heng and Zhang, Wenyang (2023) A Class of Structured High Dimensional Dynamic Covariance Matrices. Communications in Mathematics and Statistics. ISSN: 2194-671X

https://doi.org/10.1007/s40304-022-00321-7

Commun Math Stat manuscript No.
(will be inserted by the editor)

A Class of Structured High Dimensional Dynamic Covariance Matrices

Jin Yang · Heng Lian · Wenyang Zhang

Received: date / Accepted: date

Abstract High dimensional covariance matrices have attracted much attention of statisticians and econometricians during the past decades. Vast literature is devoted to the research in high dimensional covariance matrices. However, most of them are for constant covariance matrices. In many applications, constant covariance matrices are not appropriate, e.g. in portfolio allocation, dynamic covariance matrices would make much more sense. Simply assuming each entry of a covariance matrix is a function of time to introduce a dynamic structure would not work. In this paper, we are going to introduce a class of high dimensional dynamic covariance matrices in which a kind of additive structure is embedded. We will show the proposed high dimensional dynamic covariance matrices have many advantages in applications. An estimation pro-

We would like to acknowledge support for this project from National Natural Science Foundation of China (Grant Numbers 11931014, 11901315).

Jin Yang
School of Statistics and Data Sciences, Nankai University, Tianjin, China
E-mail: michael.jin.yang@nankai.edu.cn

Heng Lian
Department of Mathematics, City University of Hong Kong, Hong Kong, China
E-mail: henglian@cityu.edu.hk

Wenyang Zhang
Department of Mathematics, The University of York, York, United Kingdom
E-mail: wenyang.zhang@york.ac.uk

cedure is also proposed to estimate the proposed high dimensional dynamic covariance matrices. Asymptotic properties are built to justify the proposed estimation procedure. Intensive simulation studies show the proposed estimation procedure works very well when sample size is finite. Finally, we apply the proposed high dimensional dynamic covariance matrices, together with the proposed estimation procedure, to portfolio allocation. The results look very interesting.

Keywords Additive Structure · B-spline · Factor Models · High Dimensional Dynamic Covariance Matrices · Portfolio Allocation

Mathematics Subject Classification (2020) MSC 62G05 · MSC 62H12 · MSC 62P20

1 Introduction

Covariance matrices are a very important tool in data analysis, their applications appear in many disciplines such as engineering, psychology, finance, economics, to name but a few. Traditionally, sample covariance matrices are used to estimate covariance matrices. However, when the size of a covariance matrix is large, the sample covariance matrix would not work for the estimation of a function of the covariance matrix, such as the inverse of the covariance matrix (precision matrix), because the estimation errors would accumulate quickly to an unacceptable level due to the large size of the matrix. There is much literature about high dimensional covariance matrices, see, [28, 26, 13, 3, 4, 7, 25, 29, 14, 2, 5, 17, 19, 1, 16], and the references therein.

Most literature about high dimensional covariance matrices is for constant covariance matrices. However, in some applications, constant covariance matrices may not be appropriate, e.g. in portfolio allocation. The optimal portfolio allocation today may not be optimal tomorrow. Therefore, when forming portfolio allocation, it would make much more sense to use dynamic covariance matrices. A natural way to introduce dynamic into a covariance matrix is to assume each entry of the covariance matrix is an unknown function of time,

and estimate this function by data. However, this approach would not work very well in the applications relating to prediction, such as portfolio allocation and risk management, this is because the unknown function can go either up or down smoothly after the last time point where we have observation, which means we can not estimate the unknown function well at the future time point of interest. Another approach to introduce dynamic into a covariance matrix is to estimate the covariance matrix only based on the observations in a moving window. This approach is basically the same as assuming each entry of the covariance matrix is an unknown function of time, and estimating this unknown function by the local constant estimation, therefore, such approach is not ideal either. Treating the covariance matrices at different time points as a time series, and directly applying the concepts in time series to introduce a dynamic structure into the covariance matrices would not work, this is because such approach would result in too many unknown parameters and functions to estimate, especially for high dimensional cases which is what this paper is about. Besides, the algorithm would be too complicated and the computation involved would be too expensive.

To estimate high dimensional precision matrices more accurately, [13] proposed a factor model based structure for high dimensional covariance matrices. Although the covariance matrice there are still constant covariance matrice, their approach has built a bridge to connect the research in high dimensional covariance matrices to regression analysis. In this paper, making use of this bridge and the autoregressive idea, based on an additive structure, we introduce a dynamic structure to the coefficients in the regression models for the components of the high dimensional random vector to which we are interested in its covariance matrix, thereby, a dynamic structure is introduced to the covariance matrix. Putting it in a generic context, we have therefore introduced a class of structured high dimensional dynamic covariance matrices. In this paper, we will show the rationale of this class of structured high dimensional dynamic covariance matrices, construct an easy to implement estimation procedure for them, and apply them to portfolio allocation. We will show the

portfolio allocation based on the proposed dynamic covariance matrices yields better return than some commonly used approaches.

The rest of this paper is organised as follows. We begin in Section 2 with a description of the proposed class of structured high dimensional dynamic covariance matrices. In Section 3, we construct an estimation procedure for the proposed class of covariance matrices. Asymptotic properties of the proposed estimation procedure are built in Section 4 to justify the proposed estimation procedure theoretically. In Section 5, we describe how to apply the proposed dynamic covariance matrices to portfolio allocation. Intensive simulation studies are conducted in Section 6 to show how well the proposed estimation procedure works, and how better the portfolio allocation formed based on the proposed dynamic covariance matrices is, compared with some commonly used portfolio allocations. Finally, in Section 7, we apply the portfolio allocation, formed based on the proposed dynamic covariance matrices, to a data set which is freely available from

http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

and compare its returns with that of some commonly used approaches.

2 A class of structured high dimensional dynamic covariance matrices

In this section, based on the factor models, we are going to introduce a class of structured high dimensional dynamic covariance matrices. The factor models in this paper refer to the common factor models where the factors are observable. See [8, 9] for more details about the common factor models.

Suppose $(\mathbf{X}_t^{\mathrm{T}}, \mathbf{Y}_t^{\mathrm{T}})$, $t = 1, \cdots, n$, is a time series, where $\mathbf{Y}_t$ is a p_n dimensional vector and $\mathbf{X}_t$ is a q dimensional common factor. An underlying assumption throughout this paper is that $p_n \longrightarrow \infty$ when $n \longrightarrow \infty$, and q is fixed. In practice, q is often small, e.g. in the Fama-French three factor models, q is 3. Also, we assume that $\{\mathbf{X}_t, t = 1, \cdots, n\}$ is a stationary Markov process.

The standard common factor models are

$$\mathbf{Y}_t = \mathbf{a} + \mathbf{C}\mathbf{X}_t + \mathbf{e}_t$$

which do not account for any dynamic feature, and the covariance matrices based on this model would be constant matrices, see [13]. To account for the dynamic feature, we borrow the autoregressive idea, namely, we assume $\mathbf{a}$ and $\mathbf{C}$ depend on the observation at $t-1$. To avoid the high dimensionality of $\mathbf{Y}_{t-1}$, we assume $\mathbf{a}$ and $\mathbf{C}$ are functions of $\mathbf{X}_{t-1}$. This is reasonable, because $\mathbf{X}_{t-1}$ is the common factor, $\mathbf{Y}_{t-1}$ has been largely accounted for by $\mathbf{X}_{t-1}$. This modelling idea takes us to the following models

$$\mathbf{Y}_t = \mathbf{a}(\mathbf{X}_{t-1}) + \mathbf{C}(\mathbf{X}_{t-1})\mathbf{X}_t + \boldsymbol{\epsilon}_t, \tag{2.1}$$

where $\mathbf{X}_t = (x_{t,1},\ \cdots,\ x_{t,q})^{\mathrm{T}}$,

$$\mathbf{a}(\mathbf{X}_{t-1}) = \Big(a_1(\mathbf{X}_{t-1}),\ \cdots,\ a_{p_n}(\mathbf{X}_{t-1})\Big)^{\mathrm{T}}, \quad \mathbf{C}(\mathbf{X}_{t-1}) = \Big(c_{j,k}(\mathbf{X}_{t-1})\Big)_{p_n\times q},$$

$\boldsymbol{\epsilon}_t$s are random errors which are independent of $\mathbf{X}_t$s, and

$$E(\boldsymbol{\epsilon}_t|\{\boldsymbol{\epsilon}_l:\ l<t\}) = \mathbf{0}, \quad \mathrm{cov}(\boldsymbol{\epsilon}_t|\{\boldsymbol{\epsilon}_l:\ l<t\}) = \boldsymbol{\Sigma}_{0,t} = \mathrm{diag}\left(\sigma^2_{1,t},\ \cdots,\ \sigma^2_{p_n,t}\right).$$

Under model (2.1), by simple calculation, we have the conditional covariance matrix

$$\mathrm{cov}(\mathbf{Y}_t|\mathcal{F}_{t-1}) = \mathbf{C}(\mathbf{X}_{t-1})\boldsymbol{\Sigma}_x(\mathbf{X}_{t-1})\mathbf{C}(\mathbf{X}_{t-1})^{\mathrm{T}} + \boldsymbol{\Sigma}_{0,t} \tag{2.2}$$

where $\mathcal{F}_t$ is the $\sigma-$algebra generated by $\{(\mathbf{X}_l^{\mathrm{T}},\ \boldsymbol{\epsilon}_l^{\mathrm{T}}):\ l\le t\}$, and

$$\boldsymbol{\Sigma}_x(\mathbf{X}_{t-1}) = \mathrm{cov}(\mathbf{X}_t|\mathbf{X}_{t-1}).$$

Due to "curse of dimensionality", we have to impose some kind of structure on the unknown multivariate functions $a_j(\cdot)$s and $c_{j,k}(\cdot)$s, in order to get decent estimators of these functions. Addtive structure is one of the most commonly assumed structure in multiple nonparametric regression, we therefore assume the $a_j(\cdot)$s and $c_{j,k}(\cdot)$s in (2.1) have addtive structure, namely,

$$a_j(\mathbf{X}_{t-1}) = a_{j,0} + \sum_{l=1}^{q} a_{j,l}(x_{t-1,l}), \quad j = 1,\ \cdots,\ p_n.$$

$$c_{j,k}(\mathbf{X}_{t-1}) = c_{j,k,0} + \sum_{l=1}^{q} c_{j,k,l}(x_{t-1,l}), \quad j = 1,\ \cdots,\ p_n,\ k = 1,\ \cdots,\ q.$$

We also assume

$$E\Big(a_{j,l}(x_{t-1,l})\Big) = 0, \quad E\Big(c_{j,k,l}(x_{t-1,l})\Big) = 0, \tag{2.3}$$

$j = 1, \cdots, p_n$, $k = 1, \cdots, q$, $l = 1, \cdots, q$, to make the model identifiable.

We don't impose any condition on the matrix $\boldsymbol{\Sigma}_x(\cdot)$ because its size is q, which is often small in practice. As we are going to apply the proposed dynamic covariance matrix to portfolio allocation, it is reasonable to assume the $\sigma^2_{k,t}$s in $\boldsymbol{\Sigma}_{0,t}$ in (2.2) follow GARCH models, that is

$$\sigma^2_{k,t} = \alpha_{k,0} + \sum_{i=1}^{m} \alpha_{k,i}\epsilon^2_{k,t-i} + \sum_{j=1}^{s} \gamma_{k,j}\sigma^2_{k,t-j}, \quad t = 2, \cdots, n, \tag{2.4}$$

for each $k = 1, \cdots, p_n$ and for some integers m and s.

Clearly, (2.2) together with the conditions, imposed on the unknown functions involved or the variances of the random errors, represent a large class of structured high dimensional dynamic covariance matrices. The main focus of this paper is on (2.2) in which the $c_{j,k,0}$s and $c_{j,k,l}(\cdot)$s involved in $\mathbf{C}(\cdot)$, the $\alpha_{k,0}$s, $\alpha_{k,i}$s and $\gamma_{k,j}$s involved in $\boldsymbol{\Sigma}_{0,t}$, and $\boldsymbol{\Sigma}_x(\cdot)$ are unknown and need to be estimated. The $a_{j,0}$s and $a_{j,l}(\cdot)$s involved in $\mathbf{a}(\cdot)$ in model (2.1) are also unknown and need to be estimated.

We conclude this section by two remarks:

Remark 1 *Model (2.1) is interesting in its own right, since it combines additive modelling [20, 24, 14, 6] and varying coefficient modelling [21, 10, 11, 12, 26, 30, 23, 22, 27]. It is more flexible than either the additive models or the varying coefficient models, therefore, very useful in the data analysis where neither the additive models nor the varying coefficient models work.*

Remark 2 *The dynamic structure in (2.2) is fundamentally different to the dynamic structure introduced by assuming each entry of a covariance matrix is an additive function of the common factors. Clearly, the latter would need to estimate $p_n(1+p_n)/2$ additive functions, but (2.2) only needs to estimate qp_n, which is much smaller than $p_n(1+p_n)/2$ when p_n is large, additive functions.*

Most crucially, the latter does not allow any interaction of the common factors on the entries of the covariance matrix, which is unrealistic in real data analysis, this is because a covariance matrix is a quantity of second moment. Obviously, (2.2) *does allow interactions, which is more reasonable. Finally, the proposed method has the advantage that it automatically yields a positive-definite covariance matrix.*

3 Estimation procedure

In this section, we introduce an estimation procedure for $\mathrm{cov}(\mathbf{Y}_t|\mathcal{F}_{t-1})$. We will first estimate $\mathbf{C}(\cdot)$, $\boldsymbol{\Sigma}_x(\cdot)$, $\alpha_{k,i}$ and $\gamma_{k,j}$, and denote the resulting estimators by $\hat{\mathbf{C}}(\cdot)$, $\hat{\boldsymbol{\Sigma}}_x(\cdot)$, $\hat{\alpha}_{k,i}$ and $\hat{\gamma}_{k,j}$ for $i = 0, \cdots, m$ and $j = 1, \cdots, s$. Let $\hat{\boldsymbol{\Sigma}}_{0,t}$ be $\boldsymbol{\Sigma}_{0,t}$ with $\alpha_{k,i}$ and $\gamma_{k,j}$ being replaced by $\hat{\alpha}_{k,i}$ and $\hat{\gamma}_{k,j}$ respectively. We use

$$\widehat{\mathrm{cov}}(\mathbf{Y}_t|\mathcal{F}_{t-1}) = \hat{\mathbf{C}}(\mathbf{X}_{t-1})\hat{\boldsymbol{\Sigma}}_x(\mathbf{X}_{t-1})\hat{\mathbf{C}}(\mathbf{X}_{t-1})^{\mathrm{T}} + \hat{\boldsymbol{\Sigma}}_{0,t} \tag{3.1}$$

to estimate $\mathrm{cov}(\mathbf{Y}_t|\mathcal{F}_{t-1})$.

Throughout this paper, for any integers p and q, we use $\mathbf{0}_{p\times q}$ to denote a $p \times q$ matrix with each entry being 0, and $\mathbf{I}_p$ to denote an identity matrix of size p.

If the range of $x_{t,l}$, $t = 1, \cdots, n$, is different for different l, we can first standardise $x_{t,l}$, $t = 1, \cdots, n$, for each given l, such that the range of $x_{t,l}$, $t = 1, \cdots, n$, is the same for all ls. Therefore, without loss of generality, throughout this paper, we assume the range of $x_{t,l}$, $t = 1, \cdots, n$, is the same for all ls.

3.1 Estimation of $\mathbf{C}(\cdot)$ and $\mathbf{a}(\cdot)$

By (2.1), and for $j = 1, \cdots, p_n$, we have the following additive varying coefficient model

$$y_{j,t} = a_{j,0} + \sum_{l=1}^{q} a_{j,l}(x_{t-1,l}) + \sum_{k=1}^{q} x_{t,k}\left\{c_{j,k,0} + \sum_{l=1}^{q} c_{j,k,l}(x_{t-1,l})\right\} + \epsilon_{j,t},$$

which is

$$y_{j,t} = a_{j,0} + \sum_{k=1}^{q} x_{t,k} c_{j,k,0} + \sum_{l=1}^{q} \left\{ a_{j,l}(x_{t-1,l}) + \sum_{k=1}^{q} c_{j,k,l}(x_{t-1,l}) x_{t,k} \right\} + \epsilon_{j,t}, \tag{3.2}$$

for $t = 2, \cdots, n$.

Applying B-spline decomposition to $a_{j,l}(x_{t-1,l})$s and $c_{j,k,l}(x_{t-1,l})$s, and incorporating the identification condition (2.3) into the decomposition, we have

$$a_{j,l}(x_{t-1,l}) \approx (\mathbf{B}_{t-1,l} - \bar{\mathbf{B}}_l)^{\mathrm{T}} \mathbf{a}_{j,l}, \quad c_{j,k,l}(x_{t-1,l}) \approx (\mathbf{B}_{t-1,l} - \bar{\mathbf{B}}_l)^{\mathrm{T}} \mathbf{c}_{j,k,l}$$

where $\mathbf{B}_{t-1,l}$ is the vector of the B-spline basis functions at $x_{t-1,l}$, and

$$\bar{\mathbf{B}}_l = \frac{1}{n-1} \sum_{t=2}^{n} \mathbf{B}_{t-1,l}, \quad \mathbf{a}_{j,l} = (a_{j,l,1}, \cdots, a_{j,l,\mathcal{K}})^{\mathrm{T}}, \quad \mathbf{c}_{j,k,l} = (c_{j,k,l,1}, \cdots, c_{j,k,l,\mathcal{K}})^{\mathrm{T}},$$

where $\mathcal{K}$ is selected by cross-validation.

Replacing the $a_{j,l}(x_{t-1,l})$s and $c_{j,k,l}(x_{t-1,l})$s in (3.2) by their B-spline decompositions, we have the following synthetic linear model

$$y_{j,t} = a_{j,0} + \sum_{k=1}^{q} x_{t,k} c_{j,k,0} + \sum_{l=1}^{q} \left\{ (\mathbf{B}_{t-1,l} - \bar{\mathbf{B}}_l)^{\mathrm{T}} \mathbf{a}_{j,l} + \sum_{k=1}^{q} x_{t,k} (\mathbf{B}_{t-1,l} - \bar{\mathbf{B}}_l)^{\mathrm{T}} \mathbf{c}_{j,k,l} \right\} + \epsilon_{j,t}. \tag{3.3}$$

Let

$$\mathcal{X} = \begin{pmatrix} 1 & \mathbf{X}_2^{\mathrm{T}} & (1,\ \mathbf{X}_2^{\mathrm{T}}) \otimes (\mathbf{B}_{1,1} - \bar{\mathbf{B}}_1)^{\mathrm{T}} & \cdots & (1,\ \mathbf{X}_2^{\mathrm{T}}) \otimes (\mathbf{B}_{1,q} - \bar{\mathbf{B}}_q)^{\mathrm{T}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \mathbf{X}_n^{\mathrm{T}} & (1,\ \mathbf{X}_n^{\mathrm{T}}) \otimes (\mathbf{B}_{n-1,1} - \bar{\mathbf{B}}_1)^{\mathrm{T}} & \cdots & (1,\ \mathbf{X}_n^{\mathrm{T}}) \otimes (\mathbf{B}_{n-1,q} - \bar{\mathbf{B}}_q)^{\mathrm{T}} \end{pmatrix}$$

$$\boldsymbol{\beta}_j = \left(a_{j,0},\ c_{j,1,0},\ \cdots,\ c_{j,q,0},\ \mathbf{a}_{j,1}^{\mathrm{T}},\ \mathbf{c}_{j,1,1}^{\mathrm{T}},\ \cdots,\ \mathbf{c}_{j,q,1}^{\mathrm{T}},\ \cdots,\ \mathbf{a}_{j,q}^{\mathrm{T}},\ \mathbf{c}_{j,1,q}^{\mathrm{T}},\ \cdots,\ \mathbf{c}_{j,q,q}^{\mathrm{T}}\right)^{\mathrm{T}}$$

$$\mathbf{y}_j = (y_{j,2},\ \cdots,\ y_{j,n})^{\mathrm{T}}, \quad \boldsymbol{\epsilon}_j = (\epsilon_{j,2},\ \cdots,\ \epsilon_{j,n})^{\mathrm{T}}$$

(3.3) can be written to

$$\mathbf{y}_j = \mathcal{X} \boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j, \tag{3.4}$$

and the estimator of $\boldsymbol{\beta}_j$ is

$$\hat{\boldsymbol{\beta}}_j = (\mathcal{X}^{\mathrm{T}} \mathcal{X})^{-1} \mathcal{X}^{\mathrm{T}} \mathbf{y}_j$$

The estimators $\hat{a}_{j,0}$s, $\hat{c}_{j,k,0}$s, $\hat{\mathbf{a}}_{j,l}$s and $\hat{\mathbf{c}}_{j,k,l}$s of $a_{j,0}$s, $c_{j,k,0}$s, $\mathbf{a}_{j,l}$s and $\mathbf{c}_{j,k,l}$s are the corresponding components of $\hat{\boldsymbol{\beta}}_j$, and for any given u, the estimators of $a_{j,l}(u)$ and $c_{j,k,l}(u)$ are

$$\hat{a}_{j,l}(u) = (\mathbf{B}(u) - \bar{\mathbf{B}}_l)^{\mathrm{T}}\hat{\mathbf{a}}_{j,l}, \quad \hat{c}_{j,k,l}(u) = (\mathbf{B}(u) - \bar{\mathbf{B}}_l)^{\mathrm{T}}\hat{\mathbf{c}}_{j,k,l}$$

where $\mathbf{B}(u)$ is the $\mathbf{B}_{t-1,l}$ with $x_{t-1,l}$ being replaced by u. Therefore, the estimators of $a_j(\mathbf{X}_{t-1})$ and $c_{j,k}(X_{t-1})$ are

$$\hat{a}_j(\mathbf{X}_{t-1}) = \hat{a}_{j,0}+\sum_{l=1}^{q}(\mathbf{B}_{t-1,l}-\bar{\mathbf{B}}_l)^{\mathrm{T}}\hat{\mathbf{a}}_{j,l}, \quad \hat{c}_{j,k}(\mathbf{X}_{t-1}) = \hat{c}_{j,k,0}+\sum_{l=1}^{q}(\mathbf{B}_{t-1,l}-\bar{\mathbf{B}}_l)^{\mathrm{T}}\hat{\mathbf{c}}_{j,k,l}$$

for $j = 1, \cdots, p_n, \ k = 1, \cdots, q$.

3.2 Estimation of $\boldsymbol{\Sigma}_x(\cdot)$

In order to estimate $E(\mathbf{X}_t|\mathbf{X}_{t-1} = \mathbf{u})$ and $E(\mathbf{X}_t\mathbf{X}_t^{\mathrm{T}}|\mathbf{X}_{t-1} = \mathbf{u})$, for any given $\mathbf{u}$, we propose using the local constant estimators

$$\widehat{E}(\mathbf{X}_t|\mathbf{X}_{t-1} = \mathbf{u}) = \frac{\sum_{t=2}^{n}\mathbf{X}_t K_h(\|\mathbf{X}_{t-1} - \mathbf{u}\|)}{\sum_{t=2}^{n}K_h(\|\mathbf{X}_{t-1} - \mathbf{u}\|)}, \tag{3.5}$$

$$\widehat{E}(\mathbf{X}_t\mathbf{X}_t^{\mathrm{T}}|\mathbf{X}_{t-1} = \mathbf{u}) = \frac{\sum_{t=2}^{n}\mathbf{X}_t\mathbf{X}_t^{\mathrm{T}} K_h(\|\mathbf{X}_{t-1} - \mathbf{u}\|)}{\sum_{t=2}^{n}K_h(\|\mathbf{X}_{t-1} - \mathbf{u}\|)},$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function, usually taken to be the Epanechnikov kernel $K(u) = 0.75(1 - u^2)_+$, h is a bandwidth. This gives us the following estimator of $\boldsymbol{\Sigma}_x(\mathbf{u})$

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_x(\mathbf{u}) &= \widehat{E}(\mathbf{X}_t\mathbf{X}_t^{\mathrm{T}}|\mathbf{X}_{t-1} = \mathbf{u}) - \widehat{E}(\mathbf{X}_t|\mathbf{X}_{t-1} = \mathbf{u})\left\{\widehat{E}(\mathbf{X}_t|\mathbf{X}_{t-1} = \mathbf{u})\right\}^{\mathrm{T}} \\ &= \{\mathrm{tr}(\mathcal{W})\}^{-2}\mathbf{X}^{\mathrm{T}}\left\{\mathrm{tr}(\mathcal{W})\mathcal{W} - \mathcal{W}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathcal{W}\right\}\mathbf{X}\end{aligned} \tag{3.6}$$

where

$$\mathbf{X} = (\mathbf{X}_2, \cdots, \mathbf{X}_n)^{\mathrm{T}}, \quad \mathcal{W} = \mathrm{diag}(K_h(\|\mathbf{X}_1 - \mathbf{u}\|), \cdots, K_h(\|\mathbf{X}_{n-1} - \mathbf{u}\|)).$$

3.3 Estimation of $\boldsymbol{\Sigma}_{0,t}$

For each j ($j = 1, \cdots, p_n$), let

$$r_{j,t} = \hat{\epsilon}_{j,t} = y_{j,t} - \hat{a}_j(\mathbf{X}_{t-1}) - \sum_{k=1}^{q} \hat{c}_{j,k}(\mathbf{X}_{t-1}) x_{t,k}$$

By (2.4), for each k, $k = 1, \cdots, p_n$, we have the following synthetic GARCH model

$$\sigma_{k,t}^2 = \alpha_{k,0} + \sum_{i=1}^{m} \alpha_{k,i} r_{k,t-i}^2 + \sum_{j=1}^{s} \gamma_{k,j} \sigma_{k,t-j}^2, \quad t = 2, \cdots, n \tag{3.7}$$

which is equivalent to

$$r_{k,t}^2 = \alpha_{k,0} + \sum_{i=1}^{\max(m,s)} (\alpha_{k,i} + \gamma_{k,i}) r_{k,t-i}^2 + \eta_{kt} - \sum_{j=1}^{s} \gamma_{k,j} \eta_{k,t-j}, \quad t = 2, \cdots, n$$

where $\eta_{k,t} = r_{k,t}^2 - \sigma_{k,t}^2$, $\gamma_{k,i} = 0$ when $i > s$, and $\alpha_{k,i} = 0$ when $i > m$. Once $\alpha_{k,i}$ and $\gamma_{k,j}$ have been estimated, by substituting them into (3.7) and setting $\sigma_{kl}^2 = r_{k,l}^2$ for $l \leq \max(m, s)$, we can obtain an estimator $\hat{\sigma}_{k,t}^2$ of $\sigma_{k,t}^2$ and hence an estimator $\hat{\boldsymbol{\Sigma}}_{0,t}$ of $\boldsymbol{\Sigma}_{0,t}$.

For each k ($k = 1, \cdots, p_n$), let $\boldsymbol{\theta}_k = (\alpha_{k,0}, \cdots, \alpha_{k,m}, \gamma_{k,1}, \cdots, \gamma_{k,s})^{\mathrm{T}}$. We are going to use a quasi-maximum likelihood approach to estimate $\boldsymbol{\theta}_k$. We define the negative quasi log-likelihood function of $\boldsymbol{\theta}_k$ as

$$Q_{k,n}(\boldsymbol{\theta}_k) = n^{-1} \sum_{t=2}^{n} \left\{ \frac{r_{k,t}^2}{\sigma_{k,t}^2(\boldsymbol{\theta}_k)} + \log \sigma_{k,t}^2(\boldsymbol{\theta}_k) \right\} \tag{3.8}$$

where $\sigma_{k,t}^2(\boldsymbol{\theta}_k)$ are recursively defined by (3.7) with initial values being either

$$r_{k,0}^2 = \cdots = r_{k,1-m}^2 = \sigma_{k,0}^2 = \cdots = \sigma_{k,1-s}^2 = \alpha_{k,0} \tag{3.9}$$

or

$$r_{k,0}^2 = \cdots = r_{k,1-m}^2 = \sigma_{k,0}^2 = \cdots = \sigma_{k,1-s}^2 = r_{k,0}^2. \tag{3.10}$$

By minimising $Q_{k,n}(\boldsymbol{\theta}_k)$ with respect to $\boldsymbol{\theta}_k$, we use the minimiser $\hat{\boldsymbol{\theta}}_k$ to estimate $\boldsymbol{\theta}_k$, therefore, an estimator $\hat{\boldsymbol{\Sigma}}_{0,t}$ of $\boldsymbol{\Sigma}_{0,t}$ is obtained.

Finally, we note that in terms of computation, all three steps of estimation can be carried out with time complexity linear in n and p_n, since the estimation is carried out for each $j \in \{1, \ldots, p_n\}$ separately. The final computation of the covariance matrix in equation (3.1) has a time complexity $O(p_n^2)$ and computation of its inverse generally has a complexity of $O(p_n^3)$.

4 Asymptotic properties

The main interest of this paper is to estimate $\mathrm{cov}(\mathbf{Y}_t|\mathcal{F}_{t-1})$. To measure the accuracy of an estimator $\widehat{\mathbf{M}}$ of a matrix $\mathbf{M}$ of size p_n, we use the entropy loss norm, proposed by James and Stein (1961),

$$\left\|\widehat{\mathbf{M}}-\mathbf{M}\right\|_{\boldsymbol{\Sigma}}=p_n^{-1/2}\left\|\mathbf{M}^{-1/2}\left\{\widehat{\mathbf{M}}-\mathbf{M}\right\}\mathbf{M}^{-1/2}\right\|_F,$$

where $\|A\|_F$ is the Frobenius norm of matrix A. To facilitate our presentation, we focus on the convergence of $\widehat{\mathrm{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)-\mathrm{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$, after obtaining the data $\{(\mathbf{X}_1,\mathbf{Y}_1),\cdots,(\mathbf{X}_n,\mathbf{Y}_n)\}$.

We impose the following assumptions.

(C1) For $i=1,\ldots,p_n$, $(y_{it},\mathbf{X}_t,\boldsymbol{\epsilon}_{it}), t=1,\ldots,n$ is stationary and α-mixing with mixing coefficient $\alpha_i(l)\le\rho^l$ for some $\rho\in(0,1)$. ϵ_{it} has mean zero and is independent of $\{\mathbf{X}_t\}$. The support of $x_{t,j}$ is bounded.

(C2) The functions $a_{j,k,l}$ are twice continuously differentiable.

(C3) $E[\mathbf{X}_t^{\otimes 2}]$ has eigenvalues bounded and bounded away from zero, where for any matrix $\mathbf{A}$, $\mathbf{A}^{\otimes 2}=\mathbf{A}\mathbf{A}^{\mathrm{T}}$.

(C4) $\{\mathbf{X}_t, t=1,\ldots,n\}$ is a stationary Markov chain. $E[x_{t,j}|\mathbf{X}_t=\mathbf{u}]$ and $E[x_{t,j}x_{t,j'}|\mathbf{X}_t=\mathbf{u}]$ are twice continuously differentiable in $\mathbf{u}$.

(C5) For each i, $(\boldsymbol{\epsilon}_{it},\sigma_{it}^2)$ is a strictly stationary and ergodic GARCH process with $\sup_i E[\sigma_{it}^{2d}]<\infty$ for some $d>2$.

(C6) For each i, the innovations $\nu_{it}=\boldsymbol{\epsilon}_{it}/\sigma_{it}$ are i.i.d. with a nondegenerate distribution, $E\nu_{it}^2=1$ and $\sup_i E[\nu_{it}^{2d}]<\infty$ with the same d as defined in (C8).

(C7) Let $\boldsymbol{\Omega}$ be a compact subset of $(0,\infty)^{m+s+1}$. $\sup_{(\boldsymbol{\alpha}_i,\boldsymbol{\gamma}_i)\in\boldsymbol{\Omega}}\sum_{j=1}^s\gamma_{i,j}<1$, and $(\boldsymbol{\alpha}_{0i},\boldsymbol{\gamma}_{0i})$ is an interior point of $\boldsymbol{\Omega}$.

(C8) Let $\mathcal{A}(z)=\sum_{j=1}^m\alpha_{0i,j}z^j$ and $\mathcal{B}(z)=1-\sum_{j=1}^s\gamma_{0i,j}z^j$. $\mathcal{A}(z)$ and $\mathcal{B}(z)$ have no common roots on the complex plane $\mathbb{C}$, $\mathcal{A}(1)\neq 0$, $\alpha_{0i,m}+\gamma_{0i,s}\neq 0$.

(C9) The kernel function $K(z)$ is a symmetric density function that is bounded on a bounded support and satisfies the Lipschitz condition. The bandwidth h satisfies $h\asymp n^{-c}$ with $0<c<1/(q+1)$.

(C10) The number of knots for splines satisfies $\mathcal{K} \asymp n^a$ with $1/8 < a < 1/3$, where $\mathcal{K}$ is the number of basis functions in the B-splines construction. $p_n \asymp n^b$ with $b < d/2 - 1$.

Remark 3 *(C1) contains some mild regularity assumptions. Assuming $x_{t,j}$ to be bounded is common in estimation with B-splines since the basis functions are constructed on a compact interval. On the other hand,[13] used the more stringent assumption that the data are independent and identically distributed. (C2) contains smoothness condition for the component functions. (C3) is a mild assumption which should be assumed even for linear models. (C4) and (C9) are the same as assumed in [19] for the estimation of $\boldsymbol{\Sigma}_x$. (C5)-(C6) are mild regularity assumptions for the GARCH model. Compared to [13], we need to use higher-order moments for the noise since we try to model and estimate the parameters in the noise process. Assumptions (C7) and (C8) implies (2.4) admits a unique strictly stationary solution, and the parameters are identified, which are the same as those used in [18]. (C10) restricts the number of spline knots and the divergence rate of p_n. In particular, p_n can increase polynomially with n, and more stringent moment assumption with larger d allows larger p_n.*

Theorem 1 *Under assumptions (C1)-(C10),*

$$\begin{aligned}&\|\widehat{\mathrm{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n) - \mathrm{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)\|^2_{\boldsymbol{\Sigma}} \\ &= O_p\left(p_n\left(\frac{\mathcal{K}^2(\log n)^2}{n^2} + \mathcal{K}^{-8}\right) + \left(\frac{\mathcal{K}\log n}{n} + \mathcal{K}^{-4}\right) + p_n^{-1}\left(h^4 + \frac{\log n}{nh^q}\right)\right).\end{aligned}$$

Theorem 2 *Under assumptions (C1)-(C10), and that*

$$\left(p_n\left(\frac{\mathcal{K}^2(\log n)^2}{n^2} + \mathcal{K}^{-8}\right) + \left(\frac{\mathcal{K}\log n}{n} + \mathcal{K}^{-4}\right) + p_n^{-1}\left(h^4 + \frac{\log n}{nh^q}\right)\right) = o(p_n^{-1}),$$

we have

$$\begin{aligned}&\|\widehat{\mathrm{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)^{-1} - \mathrm{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)^{-1}\|^2_{\boldsymbol{\Sigma}} \\ &= O_p\left(p_n\left(\frac{\mathcal{K}^2(\log n)^2}{n^2} + \mathcal{K}^{-8}\right) + \left(\frac{\mathcal{K}\log n}{n} + \mathcal{K}^{-4}\right) + p_n^{-1}\left(h^4 + \frac{\log n}{nh^q}\right)\right).\end{aligned}$$

Remark 4 *Up to the logarithmic term the optimal choice of for the number of basis $\mathcal{K}$ is the standard choice $\mathcal{K} \asymp n^{1/5}$. For the bandwidth, the optimal choice of $h \asymp n^{-\frac{1}{4+q}}$. The first two terms in the rates of Theorem 1 are associated with the error in estimating the conditional mean of y_{it} while the last term comes from the errors in estimating $\boldsymbol{\Sigma}_x(\mathbf{X}_{t-1})$. The error in estimating $\boldsymbol{\Sigma}_{0,t}$ is also contained in the second term, since the dominating error comes from using the residuals in place of $\boldsymbol{\epsilon}_{it}$ in estimating the parameters in the GARCH model. Although the error in estimating the covariance matrix is most often used in theoretical results, the inverse of the covariance matrix appears in portfolio allocation and thus Theorem 2 is more relevant for our application.*

5 Application to portfolio allocation

In this section, we briefly describe the construction of an estimated optimal portfolio allocation based on the proposed additive structure and the associated estimation procedure. Since the formula for optimal portfolio allocation contains $E(\mathbf{Y}_t|\mathcal{F}_{t-1})$, we shall introduce its estimator $\widehat{E}(\mathbf{Y}_t|\mathcal{F}_{t-1})$ first. By taking conditional expectation of (2.1), we have

$$E(\mathbf{Y}_t|\mathcal{F}_{t-1}) = \mathbf{a}(\mathbf{X}_{t-1}) + \mathbf{C}(\mathbf{X}_{t-1})E(\mathbf{X}_t|\mathbf{X}_{t-1}).$$

Therefore, we use

$$\widehat{E}(\mathbf{Y}_t|\mathcal{F}_{t-1}) = \hat{\mathbf{a}}(\mathbf{X}_{t-1}) + \hat{\mathbf{C}}(\mathbf{X}_{t-1})\widehat{E}(\mathbf{X}_t|\mathbf{X}_{t-1}) \tag{5.1}$$

to estiamte $E(\mathbf{Y}_t|\mathcal{F}_{t-1})$, where $\widehat{E}(\mathbf{X}_t|\mathbf{X}_{t-1})$ is defined in (3.5).

Similar with [19], the estimated optimal portfolio allocation vector $\hat{\mathbf{w}}$ of p_n risky assets, to be held between times $t-1$ and t, is defined as the solution to

$$\begin{aligned} &\min_{\mathbf{w}} \mathbf{w}^{\mathrm{T}} \widehat{\mathrm{cov}}(\mathbf{Y}_t|\mathcal{F}_{t-1})\mathbf{w} \\ &\text{subject to } \mathbf{w}^{\mathrm{T}}\mathbf{1}_{p_n} = 1 \text{ and } \mathbf{w}^{\mathrm{T}}\widehat{E}(\mathbf{Y}_t|\mathcal{F}_{t-1}) = \delta, \end{aligned} \tag{5.2}$$

where δ is the target rerurn imposed on the portfolio. The solution $\hat{\mathbf{w}}$ is given by

$$\hat{\mathbf{w}} = \frac{c_3 - c_2\delta}{c_1c_3 - c_2^2}\widehat{\mathrm{cov}}(\mathbf{Y}_t|\mathcal{F}_{t-1})^{-1}\mathbf{1}_{p_n} + \frac{c_1\delta - c_2}{c_1c_3 - c_2^2}\widehat{\mathrm{cov}}(\mathbf{Y}_t|\mathcal{F}_{t-1})^{-1}\widehat{E}(\mathbf{Y}_t|\mathcal{F}_{t-1}), \tag{5.3}$$

where

$$c_1 = \mathbf{1}_{p_n}^{\mathrm{T}} \widehat{\mathrm{cov}}(\mathbf{Y}_t|\mathcal{F}_{t-1})^{-1}\mathbf{1}_{p_n}, \quad c_2 = \mathbf{1}_{p_n}^{\mathrm{T}} \widehat{\mathrm{cov}}(\mathbf{Y}_t|\mathcal{F}_{t-1})^{-1},$$
$$c_3 = \widehat{E}(\mathbf{Y}_t|\mathcal{F}_{t-1})^{\mathrm{T}} \widehat{\mathrm{cov}}(\mathbf{Y}_t|\mathcal{F}_{t-1})^{-1} \widehat{E}(\mathbf{Y}_t|\mathcal{F}_{t-1}).$$

We remark that one can easily impose further constraints to restrict short selling or limit the gross exposure of the portfolio [see, e.g., 15]. The crucial point is that one still needs to estimate the covariance matrix of asset returns to calculate the portfolio allocation. Throughout the numerical examples of this article, we use (5.2) since it is a natural starting point and frequently appears in the literature.

6 Simulation studies

In this section, we are going to use a simulated example to show how well the proposed estimation procedure and portfolio allocation work.

We generate data from model (2.1) together with (2.4). We set n to be either 1000 or 2000, p_n either 50 or 100. We also set

$$q = 3, \quad m = 1, \quad s = 1, \quad \alpha_{0,k} = 0.001, \quad \alpha_{1,k} = 0.1, \quad \gamma_{1,k} = 0.1.$$

For $j = 1, \cdots, p_n$, we set

$$a_j(\mathbf{X}_{t-1}) = 1.15 + \sum_{l=1}^{q} 0.25 \sin(2\pi x_{t-1,l}), \quad c_{j,1}(\mathbf{X}_{t-1}) = -0.2 + \sum_{l=1}^{q} 0.2 \sin(\pi x_{t-1,l}),$$

$$c_{j,2}(\mathbf{X}_{t-1}) = -0.22 + \sum_{l=1}^{q} 0.2 \sin(3\pi x_{t-1,l}), \quad c_{j,3}(\mathbf{X}_{t-1}) = 0.96 + \sum_{l=1}^{q} 0.2 \sin(4\pi x_{t-1,l}).$$

For $t = 0, \cdots, n+1$, we generate $\mathbf{X}_t$ independently from a uniform distribution on $[-1, 1]^q$, $\mathbf{Z}_t$ from p_n-variate standard normal distribution, and $\boldsymbol{\epsilon}_t$ through $\boldsymbol{\epsilon}_t = \boldsymbol{\Sigma}_{0,t}^{1/2}\mathbf{Z}_t$ recursively. Once both $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ have been generated, $\mathbf{Y}_t$ can be generated through (2.1) for $t = 1, \cdots, n+1$.

We will initially pretend that $(\mathbf{X}_{n+1}^{\mathrm{T}}, \mathbf{Y}_{n+1}^{\mathrm{T}})$ is unknown to us, and this will not be used in the estimation of $\mathrm{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$. The purpose of generating an additional data point $(\mathbf{X}_{n+1}^{\mathrm{T}}, \mathbf{Y}_{n+1}^{\mathrm{T}})$ is to enable us to calculate the one-period simple return

$$R(\hat{\mathbf{w}}) = \hat{\mathbf{w}}^{\mathrm{T}}\mathbf{Y}_{n+1} \tag{6.1}$$

of a portfolio allocation $\hat{\mathbf{w}}$ formed at time n based on data $(\mathbf{X}_t^{\mathrm{T}},\ \mathbf{Y}_t^{\mathrm{T}})$, $t = 1,\ \cdots,\ n$.

We evaluate the performance of the portfolio allocation $\hat{\mathbf{w}}$ by the Sharpe ratio

$$\mathrm{SR}(\hat{\mathbf{w}}) = \frac{E\left\{R(\hat{\mathbf{w}})\right\}}{\mathrm{SD}\left\{R(\hat{\mathbf{w}})\right\}},$$

where $\mathrm{SD}\left\{R(\hat{\mathbf{w}})\right\}$ is the standard deviation of $R(\hat{\mathbf{w}})$. A zero risk-free rate is assumed for simplicity.

To evaluate the performance of an estimator $\hat{\mathbf{M}}$ of matrix $\mathbf{M}$ (if $\mathbf{M}$ is a vector of dimension p, we treat it as a $p \times 1$ matrix), we use the following metric:

$$\Delta(\hat{\mathbf{M}}, \mathbf{M}) = \frac{\|\hat{\mathbf{M}} - \mathbf{M}\|_F}{\|\mathbf{M}\|_F}.$$

The kernel function in the estimation procedure is taken to be the Epanechnikov kernel $K(u) = 0.75(1-u^2)_+$, and the proposed bandwidth is based on a k-nearest neighbors bandwidth with k being selected by cross-validation. We define the cross-validation statistic by

$$CV(k) = \sum_{t=n-M}^{n} \left\| \mathbf{Y}_t - \hat{\mathbf{a}}^{(t-1)}(\mathbf{X}_{t-1}) - \hat{\mathbf{C}}^{(t-1)}(\mathbf{X}_{t-1})\mathbf{X}_t \right\|,$$

where $\hat{\mathbf{a}}^{(t-1)}(\cdot)$ and $\hat{\mathbf{C}}^{(t-1)}(\cdot)$ are the respective estimators of $\mathbf{a}(\cdot)$ and $\mathbf{C}(\cdot)$ using a k-nearest neighbors bandwidth based on $(\mathbf{X}_l^{\mathrm{T}},\ \mathbf{Y}_l^{\mathrm{T}})$, $l = 1,\ \cdots,\ t-1$, and M is a look-back integer parameter such that $M < n-1$. For each of the four cases: $(n = 1000,\ p_n = 50)$, $(n = 1000,\ p_n = 100)$, $(n = 2000,\ p_n = 50)$, $(n = 2000,\ p_n = 100)$, we do 1000 simulations. The results about the accuracy of the estimators $\hat{\mathbf{a}}$, $\hat{\mathbf{C}}$, $\hat{E}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$, $\hat{\boldsymbol{\Sigma}}_{0,n+1}$, and $\hat{\boldsymbol{\Sigma}}_x(\mathbf{X}_n)$ of $\mathbf{a} = \mathbf{a}(\mathbf{X}_n)$, $\mathbf{C} = \mathbf{C}(\mathbf{X}_n)$, $\boldsymbol{\Sigma}_{0,n+1}$, and $\boldsymbol{\Sigma}_x(\mathbf{X}_n)$, are presented in Table 1 which shows these estimators work very well.

We use $\widehat{\mathrm{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)^{-1}$ to estimate $\mathrm{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)^{-1}$, the results about the accuracy of the proposed estimators $\widehat{\mathrm{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ and $\widehat{\mathrm{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)^{-1}$ are presented in Table 2, which again shows the proposed estimators work very well.

To have a more visible idea about how well the proposed estimator of $\mathrm{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ or of $\mathrm{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)^{-1}$ fares, compared with the alternatives, we

Table 1 Performance of Parameter Estimators

	$n = 1000$ $p_n = 50$	$n = 1000$ $p_n = 100$	$n = 2000$ $p_n = 50$	$n = 2000$ $p_n = 100$
$E\{\Delta(\hat{\mathbf{a}}, \mathbf{a})\}$	0.0109	0.0109	0.0075	0.0075
$SD\{\Delta(\hat{\mathbf{a}}, \mathbf{a})\}$	0.0002	0.0002	0.0001	0.0001
$E\{\Delta(\hat{\mathbf{C}}, \mathbf{C})\}$	0.0379	0.0379	0.0245	0.0246
$SD\{\Delta(\hat{\mathbf{C}}, \mathbf{C})\}$	0.0211	0.0210	0.0130	0.0130
$E\{\Delta(\hat{E}(\mathbf{Y}_{n+1}\|\mathcal{F}_n), E(\mathbf{Y}_{n+1}\|\mathcal{F}_n))\}$	0.0255	0.0255	0.0177	0.0178
$SD\{\Delta(\hat{E}(\mathbf{Y}_{n+1}\|\mathcal{F}_n), E(\mathbf{Y}_{n+1}\|\mathcal{F}_n))\}$	0.0223	0.0222	0.0180	0.0179
$E\{\Delta(\hat{\Sigma}_{0,n+1}, \Sigma_{0,n+1})\}$	0.1800	0.1804	0.1107	0.1106
$SD\{\Delta(\hat{\Sigma}_{0,n+1}, \Sigma_{0,n+1})\}$	0.0503	0.0504	0.0355	0.0338
$E\{\Delta(\hat{\Sigma}_x(\mathbf{X}_n), \Sigma_x(X_n))\}$	0.0747	0.0747	0.0571	0.0571
$SD\{\Delta(\hat{\Sigma}_x(\mathbf{X}_n), \Sigma_x(X_n))\}$	0.0481	0.0481	0.0417	0.0417

$E\{\cdot\}$ and $SD\{\cdot\}$ are sample mean and sample standard deviation. For example, $E\{\Delta(\hat{\mathbf{a}}, \mathbf{a})\}$ and $SD\{\Delta(\hat{\mathbf{a}}, \mathbf{a})\}$ are the sample mean and sample standard deviation of $\Delta(\hat{\mathbf{a}}, \mathbf{a})$ over the 1000 replications, respectively.

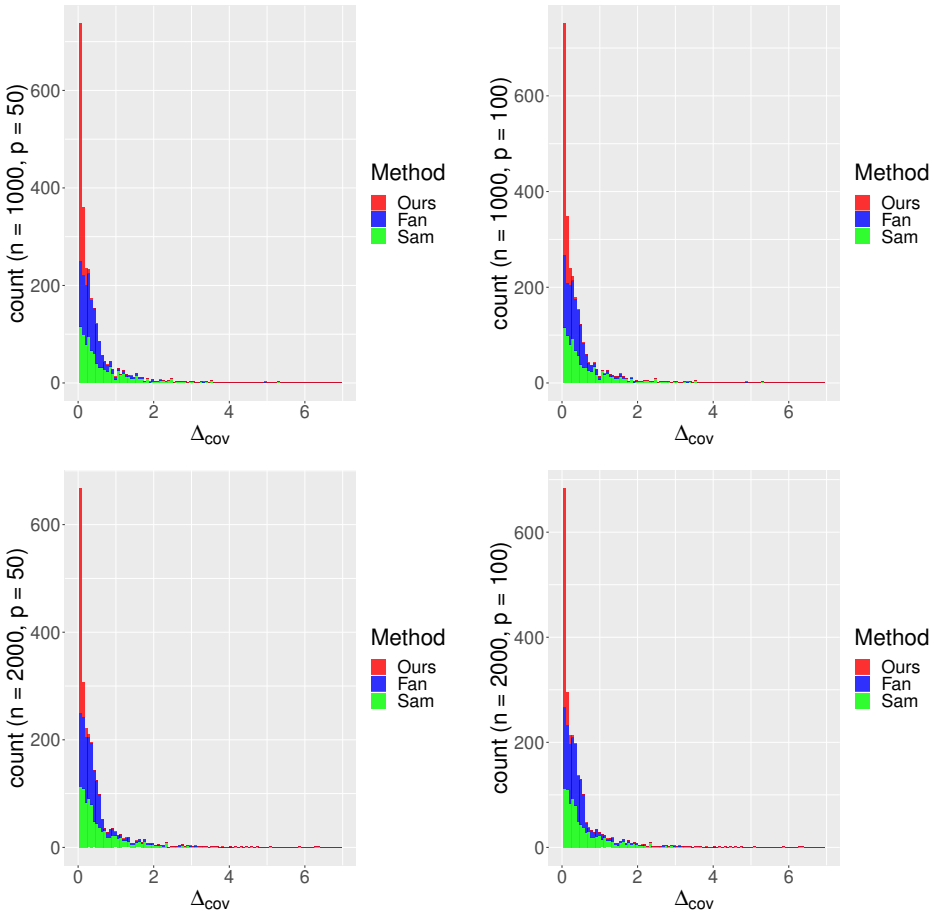
Table 2 Performances of $\widehat{\text{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ and $\widehat{\text{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)^{-1}$

	$n = 1000$ $p_n = 50$	$n = 1000$ $p_n = 100$	$n = 2000$ $p_n = 50$	$n = 2000$ $p_n = 100$
$E\{\Delta(\widehat{\mathbf{Cov}}_{n+1}, \mathbf{Cov}_{n+1})\}$	0.069	0.069	0.049	0.050
$SD\{(\Delta(\widehat{\mathbf{Cov}}_{n+1}, \mathbf{Cov}_{n+1})\}$	0.055	0.055	0.042	0.042
$E\{\Delta(\widehat{\mathbf{Cov}}_{n+1}^{-1}, \mathbf{Cov}_{n+1}^{-1})\}$	0.205	0.190	0.146	0.140
$SD\{\Delta(\widehat{\mathbf{Cov}}_{n+1}^{-1}, \mathbf{Cov}_{n+1}^{-1})\}$	0.022	0.022	0.018	0.017

In this table, $\mathbf{Cov}_{n+1} = cov(\mathbf{Y}_{n+1}|\mathcal{F}_n)$, and $\widehat{\mathbf{Cov}}_{n+1} = \widehat{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$.

produce the histograms, in Figures 1 and 2, of $\Delta\left(\widehat{\text{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n),\ \text{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)\right)$ (denoted by Δ_{COV}) or of $\Delta\left(\widehat{\text{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)^{-1},\ \text{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)^{-1}\right)$ (denoted by $\Delta_{\text{COV}^{-1}}$) over the 1000 simualtions conducted for each scenario, $\widehat{\text{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ is constructed by either the proposed method or alternatives. Figures 1 and 2 show the proposed method outperforms the alternatives regardless for estimating $\text{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ or $\text{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)^{-1}$.

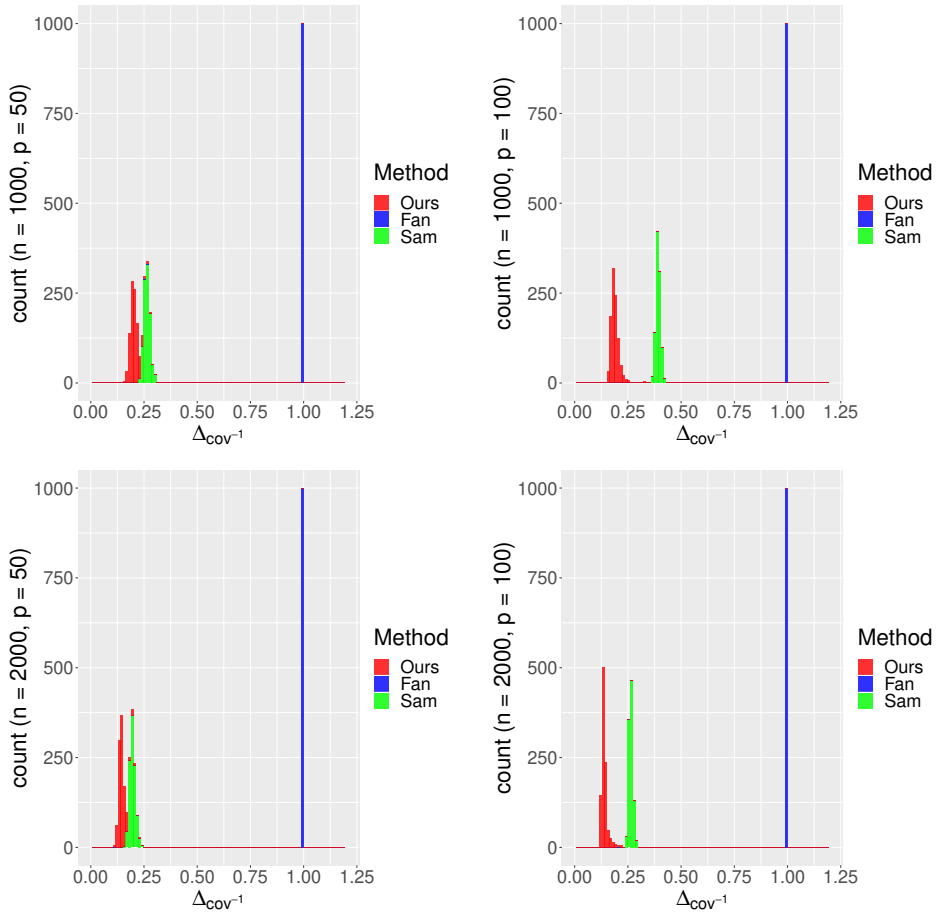
Fig. 1 The accuracy of the estimators of $\mathbf{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$



The histograms for the proposed method are in (red), the sample covariance matrix in (green), the method proposed by Fan, Fan, and Lv (2008) in (blue).

Using a target return $\delta = 1\%$, we now examine the performance of the proposed portfolio allocation by computing the return as described in (6.1). To see how much gain can be made by using our proposed structure, we make a comparison with portfolio allocations based on Markowitzs formula but where the covariance matrix is estimated using the sample covariance matrix and the factor model given in [13]. The mean, standard deviation, and Sharpe ratio

Fig. 2 The accuracy of the estimators of $\mathbf{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)^{-1}$



The histograms for the proposed method are in (red), the sample covariance matrix in (green), the method proposed by Fan, Fan, and Lv (2008) in (blue).

of the returns based on the 1000 simulations are presented in Table 3. Table 3 shows, our portfolio allocation performs much better than others, in terms of Sharpe ratio. This suggests there is significant gain from making use of the proposed dynamic structure of the covariance matrix.

Table 3 The Sharpe Ratios for Various Portfolio Allocations

	$(n = 1000,\ p_n = 50)$	$(n = 1000,\ p_n = 100)$	$(n = 2000,\ p_n = 50)$	$(n = 2000,\ p_n = 100)$
$E\{R(\hat{w})\}$	1.12%	1.13%	1.19%	1.19%
$E\{R(\hat{w}_1)\}$	1.16%	1.14%	1.24%	1.21%
$E\{R(\hat{w}_2)\}$	1.17%	1.14%	1.25%	1.22%
$SD\{R(\hat{w})\}$	0.78%	0.73%	0.85%	0.80%
$SD\{R(\hat{w}_1)\}$	1.00%	0.86%	1.23%	1.00%
$SD\{R(\hat{w}_2)\}$	0.97%	0.82%	1.21%	0.97%
$SR(\hat{w})$	1.43	1.54	1.40	1.49
$SR(\hat{w}_1)$	1.16	1.32	1.00	1.21
$SR(\hat{w}_2)$	1.20	1.39	1.03	1.25

In this table, $\hat{w}$ is the proposed portfolio allocation, $\hat{w}_1$ and $\hat{w}_2$ are the portfolio allocations formed by Markowitzs formula respectively using the sample covariance matrix and the covariance matrix given in [13].

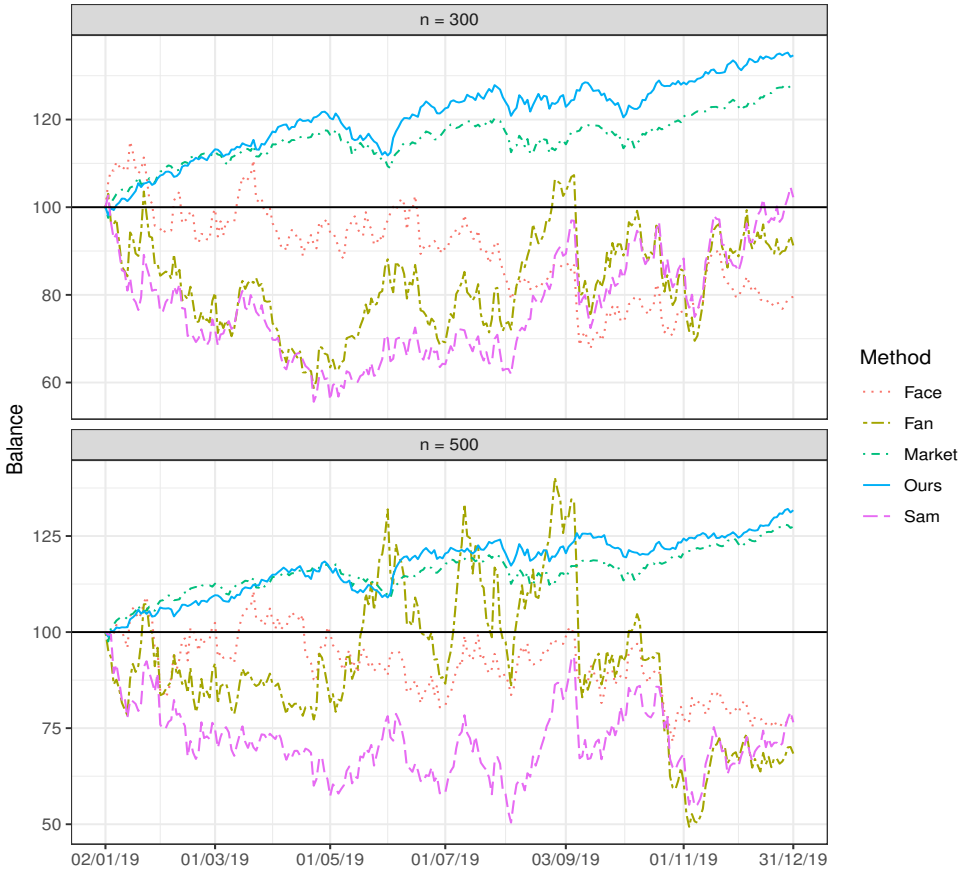
7 Real data analysis

In this section, we are going to use a real data example to demonstrate the gain of using the proposed dynamic structure for the covariance matrix in Markowitz's formula when forming a portfolio allocation. Except the market portfolio (denoted by $Market$, it serves as an important benchmark indicating whether we are in a bull or bear market), all portfolio allocations in this section are formed by Markowitz's formula but with different approaches to deal with the covariance matrix involved in the formula. The portfolio allocation based on the proposed dynamic structure and estimation for the covariance matrix is denoted by $Ours$, based on the structure and estimation in [19] is denoted by $Face$, based on the sample covariance matrix is denoted by Sam, and based on the structure and estimation in [13] is denoted by Fan.

We are going to compare the returns of $Ours$, $Face$, Sam, Fan, and $Market$. In all cases, we use the same target return $\delta = 1\%$. The kernel function used in the construction of $Ours$ is still taken to be the Epanechnikov kernel, and the bandwidths are selected by the method described in Section 6.

The dataset for us to study is from the Kenneth French's website, http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html, it is about the returns of $p_n = 49$ industry portfolios during 2019. The dataset is used directly and not altered or manipulated in any way. The response variable $\mathbf{Y}_t$ is chosen to be the vector of the daily simple returns of $p_n = 49$ industry portfolios (value weighted) minus the risk-free rate (one-month Treasury bill rate). The observable factors $x_{1,t}$, $x_{2,t}$ and $x_{3,t}$ are taken to be the market, size and value factors, respectively, from the Fama-French three-factor model.

Fig. 3 Trading Strategies



We compare the four portfolio allocations (*Ours*, *Face*, *Sam*, and *Fan*) along with *market* during 2019 using a simple trading strategy. We trade on each trading day, which is $T = 252$ trading days in this year and we assume we have initial balance of £100. Besides, we assume no transaction costs, allow for short selling, and assume that all possible portfolio allocations are attainable. Our trading strategy consists of forming a portfolio allocation $\hat{\mathbf{w}}$ at the end of each trading day and holding it until the end of the next trading day. Between day $t-1$ and day t, we obtain the portfolio return

$$R_t(\hat{\mathbf{w}}) = \hat{\mathbf{w}}^{\mathrm{T}}\mathbf{Y}_t + R_{f,t},$$

where $\hat{\mathbf{w}}$ is formed based on $(\mathbf{X}_{t-j}^{\mathrm{T}},\ \mathbf{Y}_{t-j}^{\mathrm{T}})$, $j = 1,\ \cdots,\ n$ for some look-back integer n and $R_{f,t}$ is the risk-free rate on day t. With the realized returns $R_t(\hat{\mathbf{w}})$, $t = 1,\ \cdots,\ T$, we can calculate the annualized Sharpe ratio

$$\mathrm{SR}(\hat{\mathbf{w}}) = \frac{\bar{R}(\hat{\mathbf{w}})}{SD(R)}\sqrt{T},$$

where

$$\bar{R}(\hat{\mathbf{w}}) = \frac{1}{T}\sum_{t=1}^{T}\left\{R_t(\hat{\mathbf{w}}) - R_{f,t}\right\}, \quad SD(R) = \left[\frac{1}{T}\sum_{t=1}^{T}\left\{R_t(\hat{\mathbf{w}}) - R_{f,t} - \bar{R}(\hat{\mathbf{w}})\right\}^2\right]^{1/2}.$$

The annualized Sharpe ratio for each of the five portfolio allocations under comparison is presented in Table 4 when $n = 300$ or 500. Meanwhile, we plot the balance for each portfolio at the end of each trading day in Figure 3. Table 4 and Figure 3 show clearly that *Ours* performs significantly better than other four.

We remark that *Ours*, *Face*, *Sam*, and *Fan* are all constructed based on Markowitz's formula, the difference between them lies in the way to estimate the covariance matrix of returns, which appears in Markowitz's formula. The crucial point is that *Sam* and *Fan* do not consider the dynamic feature of the covariance matrix in their estimation. Although *Face* takes into account the dynamic feature, it still does not do very well in terms of the balance left on the final trading day in 2019. *Ours* employs an additive dynamic structure for the covariance matrix, such structure seems working very well for this dataset and yields the best return.

Table 4 Annualized Sharpe ratios

	$n = 300$	$n = 500$
Ours	2.443	2.098
Face	-0.271	-0.220
Sam	0.339	0.059
Fan	0.203	0.015
Market	1.924	1.924

Appendix: Proofs of the Theorems

We write $a_{j,l}$ as $a_{j,0,l}$, $c_{j,k,l}$ as $a_{j,k,l}$, and $c_{j,k}$ as $a_{j,k}$. We write the model more succinctly as

$$y_{it} = \overline{\mathbf{X}}_t^{\mathrm{T}} \overline{\mathbf{a}}_i(\mathbf{X}_{t-1}) + \epsilon_{it},$$

where $\overline{\mathbf{X}}_t = (1, \mathbf{X}_t^{\mathrm{T}})^{\mathrm{T}}$, $\overline{\mathbf{a}}_i(\mathbf{X}_{t-1}) = (a_{i,0}(\mathbf{X}_{t-1}), a_{i,1}(\mathbf{X}_{t-1}), \ldots, a_{i,q}(\mathbf{X}_{t-1}))^{\mathrm{T}}$, $a_{i,k}(\mathbf{X}_{t-1}) = a_{i,k,0} + a_{i,k,1}(x_{t-1,1}) + \cdots + a_{i,k,q}(x_{t-1,q})$. Using B-splines, we model the functions as $a_{i,k,l}(x_{t-1,l}) \approx \mathbf{B}^{\mathrm{T}}(x_{t-1,l})\boldsymbol{\theta}_{i,k,l}$ where $\mathbf{B}$ now denotes the *centered* basis functions. Let $\boldsymbol{\theta}_{i,k} = (a_{i,k,0}, \boldsymbol{\theta}_{i,k,1}^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_{i,k,q}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\Theta}_i = (\boldsymbol{\theta}_{i,0}, \ldots, \boldsymbol{\theta}_{i,q})$, $\boldsymbol{\theta}_i = \mathrm{vec}(\boldsymbol{\Theta}_i) = (\boldsymbol{\theta}_{i,0}^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_{i,q}^{\mathrm{T}})^{\mathrm{T}}$, and $\mathbf{B}(\mathbf{X}_{t-1}) = (1, \mathbf{B}^{\mathrm{T}}(x_{t-1,1}), \ldots, \mathbf{B}^{\mathrm{T}}(x_{t-1,q}))^{\mathrm{T}}$. Then the least squares problem is

$$\min \sum_{i,t} (y_{it} - \overline{\mathbf{X}}_t \boldsymbol{\Theta}_i^{\mathrm{T}} \mathbf{B}(\mathbf{X}_{t_1}))^2 = \min \sum_{i,t} (y_{it} - (\overline{\mathbf{X}}_t \otimes \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t_1}))\boldsymbol{\theta}_i)^2$$

where $\otimes$ denotes the Kronecker product.

Below we use C to denote a generic positive constant whose exact value can change even on the same line. Whenever we use the constant $C_1 > 0$ in $1/n^{C_1}$, C_1 will denote a constant that can be chosen to be arbitrarily large. We use $\|.\|_{op}$ to denote the operator norm of a matrix (the operator norm is the same as the largest singular value of the matrix) and use $\|.\|$ to denote the Frobenius norm of a matrix or the Euclidean norm of a vector. We use $\|.\|_{L^2}$ to denote the L^2 norm of functions and $\|.\|_\infty$ is the sup-norm for vectors (maximum absolute value of the components). Since we will very frequently use tail probability, for simplicity of notation we write $P(X > Ca) = O(b)$ as $X = O_t(a; b)$, where a is possibly random, while $X = o_t(a; b)$ means $P(X > \delta a) = O(b)$ for any $\delta > 0$. $O_v(a), O_{p,v}(a), O_{t,v}(a; b)$ denotes a (possibly random) vector such that its Euclidean norm is of order $O(a), O_p(a), O_t(a; b)$, respectively.

Let $\boldsymbol{\theta}_{0,j,k,l}$ be the spline approximation coefficient that satisfies $\sup_x |a_{j,k,l}(x) - \mathbf{B}^{\mathrm{T}}(x)\boldsymbol{\theta}_{0,j,k,l}| \leq CK^{-2}$ which is possible under smoothness assumption (C2). Set $\boldsymbol{\theta}_{0,i,k} = (a_{i,k,0}, \boldsymbol{\theta}_{0,i,k,1}^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_{0,i,k,q}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\Theta}_{0i} = (\boldsymbol{\theta}_{0,i,0}, \ldots, \boldsymbol{\theta}_{0,i,q})$, and $\boldsymbol{\theta}_{0i} = \mathrm{vec}(\boldsymbol{\Theta}_{0i}) = (\boldsymbol{\theta}_{0,i,0}^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_{0,i,q}^{\mathrm{T}})^{\mathrm{T}}$.

First we consider the asymptotic property of the initial mean estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_{p_n}^{\mathrm{T}})^{\mathrm{T}}$. Although the results are relatively standard, we aim to obtain bounds that hold uniformly

over different responses $i = 1, \ldots, p_n$ in order to obtain the rates on the conditional variance of $\mathbf{Y}_t$. To this effect, tail bounds for the estimators are derived.

Proposition 1 *Let* $r_n = \sqrt{\mathcal{K}/n} + \mathcal{K}^{-2}$ *and* $r'_n = \sqrt{\mathcal{K}\log n/n} + \mathcal{K}^{-2}$. *We have* $\max_i \|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{0i}\| = O_t(r'_n; p_n\mathcal{K}^d(\log n)^{3d}/n^{d-1})$.

Proof of Proposition 1. We show that

$$\inf_{\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0i}\| = Cr'_n} \sum_{t=1}^{n}(y_{it} - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_i\overline{\mathbf{X}}_t)^2 - \sum_{t=1}^{n}(y_{it} - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_{0i}\overline{\mathbf{X}}_t)^2 > 0$$

with high probability.

We have

$$\begin{aligned}
&\sum_{t=1}^{n}(y_{it} - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_i\overline{\mathbf{X}}_t)^2 - \sum_{t=1}^{n}(y_{it} - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_{0i}\overline{\mathbf{X}}_t)^2 \\
&= \sum_t (\mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_i\overline{\mathbf{X}}_t) - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_{0i}\overline{\mathbf{X}}_t)^2 \\
&\qquad -2(\epsilon_{it} - r_{it})(\mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_i\overline{\mathbf{X}}_t - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_{0i}\overline{\mathbf{X}}_t),
\end{aligned}$$

where $r_{it} = \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_{0i}\overline{\mathbf{X}}_t - \overline{\mathbf{X}}_t^{\mathrm{T}}\bar{\mathbf{a}}_i(\mathbf{X}_{t-1})$ with $|r_{it}| = O(\mathcal{K}^{-2})$.

Furthermore,

$$\sum_t (\mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_i\overline{\mathbf{X}}_t) - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_{0i}\overline{\mathbf{X}}_t)^2 = n(\boldsymbol{\theta}_i^{\mathrm{T}} - \boldsymbol{\theta}_{0i}^{\mathrm{T}})\widehat{\mathbf{A}}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0i}),$$

where

$$\widehat{\mathbf{A}} = \frac{1}{n}\sum_{t=1}^{n}\left(\overline{\mathbf{X}}_t \otimes \mathbf{B}(\mathbf{X}_{t-1})\right)\left(\overline{\mathbf{X}}_t^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\right). \tag{A.1}$$

By Lemma 1, eigenvalues of $\widehat{\mathbf{A}}$ are bounded and bounded away from zero, with probability at least $1 - O(1/n^{C_1})$. Thus

$$\sum_t (\mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_i\overline{\mathbf{X}}_t) - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_{0i}\overline{\mathbf{X}}_t)^2 \geq Cn\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0i}\|^2, \tag{A.2}$$

with probability at least $1 - O(1/n^{C_1})$.

Next, using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\sum_t r_{it}(\mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_i\overline{\mathbf{X}}_t) - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_{0i}\overline{\mathbf{X}}_t) \\
&= C\sqrt{n}\mathcal{K}^{-2} \cdot O_t(\sqrt{n\|\boldsymbol{\theta} - \boldsymbol{\theta}\|^2}; 1/n^{C_1}).
\end{aligned} \tag{A.3}$$

Finally consider the term $\epsilon_{it}(\mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_i\overline{\mathbf{X}}_t) - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_{0i}\overline{\mathbf{X}}_t)$. We have

$$\begin{aligned}
&\sum_t \epsilon_{it}(\mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_i\overline{\mathbf{X}}_t) - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_{0i}\overline{\mathbf{X}}_t) \\
&= \sum_t (\boldsymbol{\theta}_i^{\mathrm{T}} - \boldsymbol{\theta}_{0i}^{\mathrm{T}})\left(\overline{\mathbf{X}}_t \otimes \mathbf{B}(\mathbf{X}_{t-1})\right)\epsilon_{it}
\end{aligned}$$

$$\leq \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0i}\| \left\| \sum_t \overline{\mathbf{X}}_t \otimes \mathbf{B}(\mathbf{X}_{t-1})\epsilon_{it} \right\|.$$

Lemma 2 bound the second term above in terms of tail probability. By this lemma, we have

$$\begin{aligned} &\left| \sum_t \epsilon_{it}(\mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_i\overline{\mathbf{X}}_t - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_{0i}\overline{\mathbf{X}}_t) \right| \\ &= O_t(\sqrt{n\mathcal{K}\log n}; \mathcal{K}^d(\log n)^{3d}/n^{d-1})). \end{aligned} \tag{A.4}$$

Combining (A.2), (A.3) and (A.4),

$$\sum_{t=1}^n (y_{it} - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_i\overline{\mathbf{X}}_t)^2 - \sum_{t=1}^n (y_{it} - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\boldsymbol{\Theta}_{0i}\overline{\mathbf{X}}_t)^2 > 0$$

with probability at least $1 - O(\mathcal{K}^d(\log n)^{3d}/n^{d-1})$ uniformly over $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0i}\|^2 = C(r_n')^2$.

Thus $\|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{0i}\| = O_t(r_n'; \mathcal{K}^d(\log n)^{3d}/n^{d-1})$ and, as an immediate corollary by Boole's inequality, $\max_i \|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{0i}\|) = O_t(r_n'; p_n\mathcal{K}^d(\log n)^{3d}/n^{d-1})$. □

Now we consider the convergence rate for the parameters in the GARCH model. The main difference from the standard setting is that the residuals here are estimated. We write the estimated residual as $\widehat{\epsilon}_{it}$. We define

$$\begin{aligned} &\Delta \\ :=& \max_{i,t} |\widehat{\epsilon}_{it} - \epsilon_{it}| \\ =& \max_{i,t} |\mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\widehat{\boldsymbol{\Theta}}_i\overline{\mathbf{X}}_t - \overline{\mathbf{a}}_i^{\mathrm{T}}(\mathbf{X}_{t-1})\overline{\mathbf{X}}_t| \\ \leq& \max_{i,t} C\|(\widehat{\boldsymbol{\Theta}}_i - \boldsymbol{\Theta}_{0i})^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{t-1})\| + C\mathcal{K}^{-2} \\ =& O_p\left(\sqrt{\frac{\mathcal{K}\log n}{n}} + \mathcal{K}^{-2}\right). \end{aligned}$$

Let $\boldsymbol{\vartheta}_i = (\alpha_{i,0}, \alpha_{i,1}, \ldots, \alpha_{i,m}, \gamma_{i,1}, \ldots, \gamma_{i,s})^{\mathrm{T}}$ be all the parameters in the GARCH model, with true parameter value $\boldsymbol{\vartheta}_{0i} = (\alpha_{0i,0}, \alpha_{0i,1}, \ldots, \alpha_{0i,m}, \gamma_{0i,1}, \ldots, \gamma_{0i,s})^{\mathrm{T}}$.

Proposition 2 *Denote our quasi-likelihood estimator of $\boldsymbol{\vartheta}_{0i}$ by $\widehat{\boldsymbol{\vartheta}}_i$, we have* $\max_i \|\widehat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_{0i}\| = O_t(\Delta; \frac{p_n(\log n)^{3d/2}}{n^{d/2-1}})$, *where $\Delta := \max_{i,t} |\widehat{\epsilon}_{it} - \epsilon_{it}|$ as is defined above.*

Proof of Proposition 2. Let $\widetilde{\sigma}_{it}^2(\boldsymbol{\vartheta}_i)$ be defined iteratively by

$$\widetilde{\sigma}_{it}^2(\boldsymbol{\vartheta}_i) = \alpha_{i,0} + \sum_{j=1}^m \alpha_{i,j}\epsilon_{it-j}^2 + \sum_{j=1}^s \gamma_{i,j}\widetilde{\sigma}_{it-j}^2(\boldsymbol{\vartheta}_i),$$

and $\widehat{\sigma}_{it}^2(\boldsymbol{\vartheta}_i)$ defined iteratively by

$$\widehat{\sigma}_{it}^2(\boldsymbol{\vartheta}_i) = \alpha_{i,0} + \sum_{j=1}^m \alpha_{i,j}\widehat{\epsilon}_{it-j}^2 + \sum_{j=1}^s \gamma_{i,j}\widehat{\sigma}_{it-j}^2(\boldsymbol{\vartheta}_i),$$

both with the initial values given by (3.9) or (3.10) . The negative quasi-log-likelihood is given by $\widehat{\mathcal{L}}_i(\boldsymbol{\vartheta}_i) = \frac{1}{n}\sum_{t=1}^n \widehat{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_i)$ with $\widehat{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_i) = \widehat{\epsilon}_{it}^2/\widehat{\sigma}_{it}^2(\boldsymbol{\vartheta}_i) + \log\widehat{\sigma}_{it}^2(\boldsymbol{\vartheta}_i)$. Similarly let $\widetilde{\mathcal{L}}_i(\boldsymbol{\vartheta}_i) = \frac{1}{n}\sum_{t=1}^n \widetilde{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_i)$ with $\widetilde{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_i) = \epsilon_{it}^2/\widetilde{\sigma}_{it}^2(\boldsymbol{\vartheta}_i) + \log\widetilde{\sigma}_{it}^2(\boldsymbol{\vartheta}_i)$.

In the first part of the proof, we establish consistency of $\widehat{\boldsymbol{\vartheta}}_i$ uniformly over i. By the proof of Theorem 7.1 in [18], we only need to show that

$$\sup_{1\le i\le p_n, \boldsymbol{\vartheta}_i\in\boldsymbol{\Omega}} |\widehat{\mathcal{L}}_i(\boldsymbol{\vartheta}_i) - \widetilde{\mathcal{L}}_i(\boldsymbol{\vartheta}_i)| = o_p(1).$$

Let

$$\widehat{\underline{\sigma}}_{it}(\boldsymbol{\vartheta}_i) = \begin{pmatrix} \widehat{\sigma}^2_{it}(\boldsymbol{\vartheta}_i) \\ \widehat{\sigma}^2_{it-1}(\boldsymbol{\vartheta}_i) \\ \vdots \\ \widehat{\sigma}^2_{it-s+1}(\boldsymbol{\vartheta}_i) \end{pmatrix}, \ \widehat{\underline{c}}_{it}(\boldsymbol{\vartheta}_i) = \begin{pmatrix} \alpha_{i,0} + \sum_{j=1}^m \alpha_{i,j}\widehat{\epsilon}^2_{it-j} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \ B(\boldsymbol{\vartheta}_i) = \begin{pmatrix} \gamma_{i,1} & \gamma_{i,2} & \cdots & \gamma_{i,s} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix},$$

and similarly define $\widetilde{\underline{\sigma}}_{it}(\boldsymbol{\vartheta}_i)$ and $\widetilde{\underline{c}}_{it}(\boldsymbol{\vartheta}_i)$. In the following the dependence of these quantities on $\boldsymbol{\vartheta}_i$ is often suppressed for simplicity of notation. By assumption (C10), we have $\sup_{\boldsymbol{\vartheta}_i\in\boldsymbol{\Omega}} \|B\|_{op} =: \rho < 1$. Using $\widehat{\epsilon}^2_{it} - \epsilon^2_{it} = \Delta^2_{it} + 2\epsilon_{it}\Delta_{it}$ and $|\Delta_{it}| \le \Delta$, where $\Delta_{it} := \widehat{\epsilon}_{it} - \epsilon_{it}$, we have

$$\|\widehat{\underline{c}}_{it} - \widetilde{\underline{c}}_{it}\| = |\sum_{j=1}^m \alpha_{i,j}(\Delta^2_{it-j} + 2\epsilon_{it-j}\Delta_{it-j})| \le C(\Delta^2 + \Delta\sum_{j=1}^m \alpha_{i,j}|\epsilon_{it-j}|),$$

and using $\widehat{\underline{\sigma}}^2_{it} = \widehat{\underline{c}}_{it} + B\widehat{\underline{\sigma}}^2_{it-1}$,

$$\begin{aligned} &\|\widehat{\underline{\sigma}}^2_{it} - \widetilde{\underline{\sigma}}^2_{it}\| \\ &= \|(\widehat{\underline{c}}_{it} - \widetilde{\underline{c}}_{it}) + B(\widehat{\underline{c}}_{it-1} - \widetilde{\underline{c}}_{it-1}) + \cdots + B^{t-1}(\widehat{\underline{c}}_{i1} - \widetilde{\underline{c}}_{i1})\| \\ &\le C(\Delta^2 + \Delta\sum_{k=0}^{t-1}\rho^k\sum_{j=1}^m \alpha_{i,j}|\epsilon_{it-k-j}|). \end{aligned}$$

Furthermore,

$$\begin{aligned} &(1/n)\sum_{t=1}^n\sum_{k=0}^{t-1}\rho^k|\epsilon_{it-k-j}| \\ &= (1/n)\sum_{k=0}^{n-1}\rho^k\sum_{t=k+1}^n|\epsilon_{it-k-j}| \\ &\le \frac{1}{n(1-\rho)}(|\epsilon_{i1}| + \cdots + |\epsilon_{in}|), \end{aligned}$$

and using Theorem 2.18 (ii) of [12], similar to the arguments used in the proof of Lemma 2

$$\begin{aligned} &P\left(\frac{1}{n}(|\epsilon_{i1}| + \cdots + |\epsilon_{in}|) > C + a\right) \\ &\le C\exp\left\{-C\frac{a^2 n/\log n}{1/\log n + b_n a}\right\} + O(1/n^{C_1}) + C\frac{n}{b_n^{2d}}, \end{aligned}$$

if $a > C/b_n^{2d-1}$. Choosing a to be a constant and $b_n \sim n/(\log n)^2$, the right hand size above is $O(\frac{(\log n)^{4d}}{n^{2d-1}})$. Thus

$$\frac{1}{n}\sum_{t=1}^n \|\widehat{\underline{\sigma}}^2_{it} - \widetilde{\underline{\sigma}}^2_{it}\| = O_t(\Delta; \frac{(\log n)^{4d}}{n^{2d-1}}).$$

Using similar arguments, we can get

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\epsilon_{it}^2|\widehat{\sigma}_{it}^2-\widetilde{\sigma}_{it}^2| \\
&\le \Delta^2\left(\frac{1}{n}\sum_{t=1}^{n}\epsilon_{it}^2\right)+\Delta\left(\frac{1}{n}\sum_{t=1}^{n}\epsilon_{it}^2\sum_{k=0}^{t-2}\rho^k\sum_{j=1}^{m}\alpha_{i,j}|\epsilon_{it-k-j}|\right) \\
&= O_t(\Delta;\frac{(\log n)^d}{n^{\frac{1}{2}d-1}}).
\end{aligned}$$

These bounds lead to, since $\widehat{\sigma}_{it}^2$ and $\widetilde{\sigma}_{it}^2$ are bounded away from zero (they are larger than $\alpha_{i,0}$),

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{t=1}^{n}(\frac{\widehat{\epsilon}_{it}^2}{\widehat{\sigma}_{it}^2}-\frac{\epsilon_{it}^2}{\widetilde{\sigma}_{it}^2})\right| \\
&\le \frac{1}{n}\sum_{t=1}^{n}\frac{|\widehat{\epsilon}_{it}^2-\epsilon_{it}^2|}{\widehat{\sigma}_{it}^2}+\frac{1}{n}\sum_{t=1}^{n}\epsilon_{it}^2\left|\frac{\widehat{\sigma}_{it}^2-\widetilde{\sigma}_{it}^2}{\widehat{\sigma}_{it}^2\widetilde{\sigma}_{it}^2}\right| \\
&\le C\left(\frac{1}{n}\sum_{t=1}^{n}(\Delta^2+\Delta|\epsilon_{it}|)+\frac{1}{n}\sum_{t=1}^{n}\epsilon_{it}^2|\widehat{\sigma}_{it}^2-\widetilde{\sigma}_{it}^2|\right) \\
&= O_t(\Delta;\frac{(\log n)^d}{n^{d/2-1}}),
\end{aligned}$$

and using $|\log(x/y)|\le|x-y|/\min\{x,y\}$,

$$\begin{aligned}
&|\frac{1}{n}\sum_{t=1}^{n}\log\widehat{\sigma}_{it}^2-\frac{1}{n}\sum_{t=1}^{n}\log\widetilde{\sigma}_{it}^2| \\
&=\left|\frac{1}{n}\sum_{t=1}^{n}\log\frac{\widehat{\sigma}_{it}^2}{\widetilde{\sigma}_{it}^2}\right| \\
&\le C\frac{1}{n}\sum_{t=1}^{n}|\widehat{\sigma}_{it}^2-\widetilde{\sigma}_{it}^2| \\
&= O_t(\Delta;\frac{(\log n)^{4d}}{n^{2d-1}}).
\end{aligned}$$

Thus

$$\sup_{1\le i\le p_n,\boldsymbol{\vartheta}_i\in\Omega}|\widehat{\mathcal{L}}_i(\boldsymbol{\vartheta}_i)-\widetilde{\mathcal{L}}_i(\boldsymbol{\vartheta}_i)|=o_p(1).$$

Now we proceed to establish the convergence rate. We further define $\sigma_{it}^2(\boldsymbol{\vartheta}_i)$ to be the unique strictly stationary solution of the GARCH model (2.4), and define $\mathcal{L}_i(\boldsymbol{\vartheta}_i)=\frac{1}{n}\sum_t\mathcal{L}_{it}(\boldsymbol{\vartheta}_i)$ with $\mathcal{L}_{it}(\boldsymbol{\vartheta}_i)=\epsilon_{it}^2/\sigma_{it}^2(\boldsymbol{\vartheta}_i)+\log\sigma_{it}^2(\boldsymbol{\vartheta}_i)$. We also define $\underline{\sigma}_{it}^2=(\sigma_{it}^2,\ldots,\sigma_{it-s+1}^2)^{\mathrm{T}}$.

Similar to previous calculations in the proof of consistency of $\widetilde{\boldsymbol{\vartheta}}_i$, it is easy to see that

$$\|\frac{\partial\widehat{\underline{c}}_{it}}{\partial\alpha_{i,j}}-\frac{\partial\widetilde{\underline{c}}_{it}}{\partial\alpha_{i,j}}\|\le C(\Delta^2+\Delta\sum_{j=1}^{m}|\epsilon_{it-j}|).$$

Taking derivative of the equation

$$\widehat{\underline{\sigma}}_{it}^2-\widetilde{\underline{\sigma}}_{it}^2=(\widehat{\underline{c}}_{it}-\widetilde{\underline{c}}_{it})+B(\widehat{\underline{c}}_{it-1}-\widetilde{\underline{c}}_{it-1})+\cdots+B^{t-1}(\widehat{\underline{c}}_{i1}-\widetilde{\underline{c}}_{i1}),$$

we get

$$\left\| \frac{\partial \widehat{\underline{\sigma}}^2_{it}}{\partial \vartheta_{i,j}} - \frac{\partial \widetilde{\underline{\sigma}}^2_{it}}{\partial \vartheta_{i,j}} \right\| \leq C(\Delta^2 + \Delta \sum_{k=0}^{t-1} \rho^k \sum_{j=1}^{m} |\epsilon_{it-k-j}|).$$

Combined with (7.30) and (7.58) of [18], we get

$$\|\widehat{\underline{\sigma}}^2_{it} - \underline{\sigma}^2_{it}\| \leq C(\Delta^2 + \Delta \sum_{k=0}^{t-1} \rho^k \sum_{j=1}^{m} |\epsilon_{it-k-j}| + \rho^t),$$

and

$$\left\| \frac{\partial \widehat{\underline{\sigma}}^2_{it}}{\partial \vartheta_{i,j}} - \frac{\partial \underline{\sigma}^2_{it}}{\partial \vartheta_{i,j}} \right\| \leq C(\Delta^2 + \Delta \sum_{k=0}^{t-1} \rho^k \sum_{j=1}^{m} |\epsilon_{it-k-j}| + \rho^t). \tag{A.5}$$

We have

$$\frac{\partial \widehat{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_{0i})}{\partial \boldsymbol{\vartheta}_i} = \left(1 - \frac{\widehat{\epsilon}^2_{it}}{\widehat{\sigma}^2_{it}}\right) \frac{1}{\widehat{\sigma}^2_{it}} \frac{\partial \widehat{\sigma}^2_{it}}{\partial \boldsymbol{\vartheta}_i},$$

and

$$\frac{\partial \mathcal{L}_{it}(\boldsymbol{\vartheta}_{0i})}{\partial \boldsymbol{\vartheta}_i} = \left(1 - \frac{\epsilon^2_{it}}{\sigma^2_{it}}\right) \frac{1}{\sigma^2_{it}} \frac{\partial \sigma^2_{it}}{\partial \boldsymbol{\vartheta}_i},$$

and thus

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_t \frac{\partial \widehat{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_{0i})}{\partial \boldsymbol{\vartheta}_i} - \frac{1}{n} \sum_t \frac{\partial \mathcal{L}_{it}(\boldsymbol{\vartheta}_{0i})}{\partial \boldsymbol{\vartheta}_i} \right\| \\
\leq &\left\| \frac{1}{n} \sum_t \left(\frac{\widehat{\epsilon}^2_{it}}{\widehat{\sigma}^2_{it}} - \frac{\epsilon^2_{it}}{\sigma^2_{it}} \right) \frac{1}{\sigma^2_{it}} \frac{\partial \sigma^2_{it}}{\partial \boldsymbol{\vartheta}_i} \right\| \\
&+ \left\| \frac{1}{n} \sum_t \frac{\epsilon^2_{it}}{\sigma^2_{it}} \left(\frac{1}{\widehat{\sigma}^2_{it}} \frac{\partial \widehat{\sigma}^2_{it}}{\partial \boldsymbol{\vartheta}} - \frac{1}{\sigma^2_{it}} \frac{\partial \sigma^2_{it}}{\partial \boldsymbol{\vartheta}_i} \right) \right\| \\
&+ \left\| \frac{1}{n} \sum_t \left(\frac{\widehat{\epsilon}^2_{it}}{\widehat{\sigma}^2_{it}} - \frac{\epsilon^2_{it}}{\sigma^2_{it}} \right) \left(\frac{1}{\widehat{\sigma}^2_{it}} \frac{\partial \widehat{\sigma}^2_{it}}{\partial \boldsymbol{\vartheta}_i} - \frac{1}{\sigma^2_{it}} \frac{\partial \sigma^2_{it}}{\partial \boldsymbol{\vartheta}_i} \right) \right\|.
\end{aligned} \tag{A.6}$$

For the first term above, we have

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_t \left(\frac{\widehat{\epsilon}^2_{it}}{\widehat{\sigma}^2_{it}} - \frac{\epsilon^2_{it}}{\sigma^2_{it}} \right) \frac{1}{\sigma^2_{it}} \frac{\partial \sigma^2_{it}}{\partial \boldsymbol{\vartheta}_i} \right\| \\
\leq &\left\| \frac{1}{n} \sum_t \left(\frac{\widehat{\epsilon}^2_{it} - \epsilon^2_{it}}{\widehat{\sigma}^2_{it}} \right) \frac{1}{\sigma^2_{it}} \frac{\partial \sigma^2_{it}}{\partial \boldsymbol{\vartheta}_i} \right\| + \left\| \frac{1}{n} \sum_t \epsilon^2_{it} \left(\frac{\widehat{\sigma}^2_{it} - \sigma^2_{it}}{\widehat{\sigma}^2_{it} \sigma^2_{it}} \right) \frac{1}{\sigma^2_{it}} \frac{\partial \sigma^2_{it}}{\partial \boldsymbol{\vartheta}_i} \right\| \\
= &\, O_t(\Delta; \frac{(\log n)^d}{n^{\frac{1}{2}d-1}}),
\end{aligned}$$

since by equation (7.54) of [18], all moments of $\frac{1}{\sigma^2_{it}} \frac{\partial \sigma^2_{it}}{\partial \boldsymbol{\vartheta}}$ exist. Using (A.5), the second term of (A.6) is again $O_t(\Delta; \frac{(\log n)^d}{n^{\frac{1}{2}d-1}})$. Finally the third term of (A.6) is

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{t=1}^{n} \left(\frac{\widehat{\epsilon}^2_{it}}{\widehat{\sigma}^2_{it}} - \frac{\epsilon^2_{it}}{\sigma^2_{it}} \right) \left(\frac{1}{\widehat{\sigma}^2_{it}} \frac{\partial \widehat{\sigma}^2_{it}}{\partial \boldsymbol{\vartheta}_i} - \frac{1}{\sigma^2_{it}} \frac{\partial \sigma^2_{it}}{\partial \boldsymbol{\vartheta}_i} \right) \right\| \\
\leq C &\left(\frac{1}{n} \sum_{t=1}^{n} (\Delta^2 + \Delta |\epsilon_{it}| + \epsilon^2_{it} |\widehat{\sigma}^2_{it} - \sigma^2_{it}|)(\frac{1}{\sigma^2_{it}} \frac{\partial \sigma^2_{it}}{\partial \boldsymbol{\vartheta}_i} |\widehat{\sigma}^2_{it} - \sigma^2_{it}| + \| \frac{\partial \widehat{\sigma}^2_{it}}{\partial \boldsymbol{\vartheta}_i} - \frac{\partial \sigma^2_{it}}{\partial \boldsymbol{\vartheta}_i} \|) \right)
\end{aligned}$$

$$= O_t(\Delta^2; \frac{(\log n)^d}{n^{\frac{1}{2}d-1}}).$$

Thus we get

$$\left\| \frac{1}{n}\sum_t \frac{\partial \widehat{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_{0i})}{\partial \boldsymbol{\vartheta}_i} - \frac{1}{n}\sum_t \frac{\partial \mathcal{L}_{it}(\boldsymbol{\vartheta}_{0i})}{\partial \boldsymbol{\vartheta}_i} \right\| = O_t(\Delta; \frac{(\log n)^d}{n^{\frac{1}{2}d-1}}). \tag{A.7}$$

Similarly, using that

$$\begin{aligned}
&\frac{\partial^2 \widehat{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_i)}{\partial \boldsymbol{\vartheta}_i \partial \boldsymbol{\vartheta}_i^{\mathrm{T}}} \\
&= \left(1 - \frac{\widehat{\epsilon}_{it}^2}{\widehat{\sigma}_{it}^2}\right)\left(\frac{1}{\widehat{\sigma}_{it}^2}\frac{\partial^2 \widehat{\sigma}_{it}^2}{\partial \boldsymbol{\vartheta}_i \partial \boldsymbol{\vartheta}_i^{\mathrm{T}}}\right) + \left(2\frac{\widehat{\epsilon}_{it}^2}{\widehat{\sigma}_{it}^2} - 1\right)\left(\frac{1}{\widehat{\sigma}_{it}^2}\frac{\partial \widehat{\sigma}_{it}^2}{\partial \boldsymbol{\vartheta}_i}\right)\left(\frac{1}{\widehat{\sigma}_{it}^2}\frac{\partial \widehat{\sigma}_{it}^2}{\partial \boldsymbol{\vartheta}_i^{\mathrm{T}}}\right),
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial^2 \mathcal{L}_{it}(\boldsymbol{\vartheta}_i)}{\partial \boldsymbol{\vartheta}_i \partial \boldsymbol{\vartheta}_i^{\mathrm{T}}} \\
&= \left(1 - \frac{\epsilon_{it}^2}{\sigma_{it}^2}\right)\left(\frac{1}{\sigma_{it}^2}\frac{\partial^2 \sigma_{it}^2}{\partial \boldsymbol{\vartheta}_i \partial \boldsymbol{\vartheta}_i^{\mathrm{T}}}\right) + \left(2\frac{\epsilon_{it}^2}{\sigma_{it}^2} - 1\right)\left(\frac{1}{\sigma_{it}^2}\frac{\partial \sigma_{it}^2}{\partial \boldsymbol{\vartheta}_i}\right)\left(\frac{1}{\sigma_{it}^2}\frac{\partial \sigma_{it}^2}{\partial \boldsymbol{\vartheta}_i^{\mathrm{T}}}\right),
\end{aligned}$$

and

$$\begin{aligned}
&\left\| \frac{\partial^2 \widehat{\sigma}_{it}^2}{\partial \vartheta_{i,j} \partial \vartheta_{i,j'}} - \frac{\partial^2 \sigma_{it}^2}{\partial \vartheta_{i,j} \partial \vartheta_{i,j'}} \right\| \\
&\leq C(\Delta^2 + \Delta \sum_{k=0}^{t-1} \rho^k \sum_{j=1}^{q} |\epsilon_{it-k-j}| + \rho^t),
\end{aligned}$$

the last of which can be shown similar to (A.5), we can show that there exists a neighborhood $\mathcal{V}(\boldsymbol{\vartheta}_{0i})$ of $\boldsymbol{\vartheta}_{0i}$ such that

$$\sup_{\boldsymbol{\vartheta}_i \in \mathcal{V}(\boldsymbol{\vartheta}_{0i})} \left\| \frac{1}{n}\sum_t \frac{\partial^2 \widehat{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_i)}{\partial \boldsymbol{\vartheta}_i \partial \boldsymbol{\vartheta}_i^{\mathrm{T}}} - \frac{1}{n}\sum_t \frac{\partial^2 \mathcal{L}_{it}(\boldsymbol{\vartheta}_i)}{\partial \boldsymbol{\vartheta}_i \partial \boldsymbol{\vartheta}_i^{\mathrm{T}}} \right\| = O_t(\Delta; \frac{(\log n)^d}{n^{\frac{1}{2}d-1}}). \tag{A.8}$$

Since $\frac{\partial \mathcal{L}_{it}(\boldsymbol{\vartheta}_{0i})}{\partial \boldsymbol{\vartheta}_i}$ has mean zero, we have easily

$$\left\| \frac{1}{n}\sum_t \frac{\partial \mathcal{L}_{it}(\boldsymbol{\vartheta}_{0i})}{\partial \boldsymbol{\vartheta}_i} \right\| = O_p(n^{-1/2}). \tag{A.9}$$

Furthermore, using Theorem 2.18 (ii) of [12],

$$\left\| \frac{1}{n}\sum_t \frac{\partial \mathcal{L}_{it}(\boldsymbol{\vartheta}_{0i})}{\partial \boldsymbol{\vartheta}_i} \right\| = O_t\left(\sqrt{\log n/n}; \frac{(\log n)^{3d/2}}{n^{d/2-1}}\right). \tag{A.10}$$

Using (7.54) in [18] which stated that $\sup_{\boldsymbol{\vartheta}_i \in \mathcal{V}(\boldsymbol{\vartheta}_{0i})} \left(\frac{1}{\sigma_{it}^2}\frac{\partial \sigma_{it}^2}{\partial \boldsymbol{\vartheta}_i}\right)$ and $\sup_{\boldsymbol{\vartheta}_i \in \mathcal{V}(\boldsymbol{\vartheta}_{0i})} \left(\frac{1}{\sigma_{it}^2}\frac{\partial^2 \sigma_{it}^2}{\partial \boldsymbol{\vartheta}_i \partial \boldsymbol{\vartheta}_i^{\mathrm{T}}}\right)$ has moments of all orders, with again Theorem 2.18 (ii) of [12] we get

$$\left\| \sup_{\boldsymbol{\vartheta}_i \in \mathcal{V}(\boldsymbol{\vartheta}_{0i})} \frac{1}{n}\sum_t \frac{\partial^2 \mathcal{L}_{it}(\boldsymbol{\vartheta}_i)}{\partial \boldsymbol{\vartheta}_i \partial \boldsymbol{\vartheta}_i^{\mathrm{T}}} \right\| = O_t\left(1; \frac{(\log n)^{2d}}{n^{d-1}}\right). \tag{A.11}$$

Finally, since

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{t=1}^{n}\frac{\partial}{\partial\boldsymbol{\vartheta}_i}\widehat{\mathcal{L}}_t(\widehat{\boldsymbol{\vartheta}}_i) \\ &= \frac{1}{n}\sum_{t=1}^{n}\frac{\partial}{\partial\boldsymbol{\vartheta}_i}\widehat{\mathcal{L}}_t(\boldsymbol{\vartheta}_{0i}) + \frac{1}{n}\sum_{t=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\vartheta}_i\partial\boldsymbol{\vartheta}_i^{\mathrm{T}}}\widehat{\mathcal{L}}_t(\boldsymbol{\vartheta}_i^*)(\widehat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_{0i}), \end{aligned}$$

where $\boldsymbol{\vartheta}_i^*$ lies between $\widehat{\boldsymbol{\vartheta}}_i$ and $\boldsymbol{\vartheta}_{0i}$, we have, by (A.7)–(A.11),

$$\|\widehat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_{0i}\| = O_p(\Delta + 1/\sqrt{n}),$$

and

$$\max_i \|\widehat{\boldsymbol{\vartheta}}_i - \boldsymbol{\vartheta}_{0i}\| = O_t\left(\Delta + \sqrt{\log n/n}; \frac{p_n(\log n)^{3d/2}}{n^{d/2-1}}\right).$$

.1 Proof of Theorem 1 and Theorem 2.

Given the rates obtained for estimators of $\mathbf{a}_i$, $\boldsymbol{\alpha}_i$, $\boldsymbol{\gamma}_i$, and $\boldsymbol{\Sigma}_x$ (proved in Lemma D.1 of [19]) the proof of convergence rate for $\widehat{\mathrm{Cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ in Theorem 1 is exactly as the proof of Theorem 2 in [19] and thus omitted.

For Theorem 2, for simplicity of notation, we denote $\mathbf{M} := \mathrm{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ and $\widehat{\mathbf{M}} := \widehat{\mathrm{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$. Then

$$\begin{aligned} &\|\widehat{\mathbf{M}}^{-1} - \mathbf{M}^{-1}\|_{\boldsymbol{\Sigma}}^2 \\ &= p_n^{-1}\|\mathbf{M}^{1/2}(\widehat{\mathbf{M}}^{-1} - \mathbf{M}^{-1})\mathbf{M}^{1/2}\|^2 \\ &= p_n^{-1}\|\mathbf{M}^{1/2}\widehat{\mathbf{M}}^{-1}(\widehat{\mathbf{M}} - \mathbf{M})\mathbf{M}^{-1/2}\|^2 \\ &\le p_n^{-1}\|\mathbf{M}^{1/2}\widehat{\mathbf{M}}^{-1}\mathbf{M}^{1/2}\|_{op}^2\|\mathbf{M}^{-1/2}(\widehat{\mathbf{M}} - \mathbf{M})\mathbf{M}^{-1/2}\|^2 \\ &= \|\mathbf{M}^{1/2}\widehat{\mathbf{M}}^{-1}\mathbf{M}^{1/2}\|_{op}^2\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\boldsymbol{\Sigma}}^2 \\ &\le (2\|\mathbf{M}^{1/2}(\widehat{\mathbf{M}}^{-1} - \mathbf{M}^{-1})\mathbf{M}^{1/2}\|_{op}^2 + 2)\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\boldsymbol{\Sigma}}^2 \\ &= (2p_n\|\widehat{\mathbf{M}}^{-1} - \mathbf{M}^{-1}\|_{\boldsymbol{\Sigma}}^2 + 2)\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\boldsymbol{\Sigma}}^2. \end{aligned}$$

Since we assumed $p_n\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\boldsymbol{\Sigma}}^2 = o_p(1)$, the term $2p_n\|\widehat{\mathbf{M}}^{-1} - \mathbf{M}^{-1}\|_{\boldsymbol{\Sigma}}^2\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\boldsymbol{\Sigma}}^2$ can be moved to the left hand side to get $\|\widehat{\mathbf{M}}^{-1} - \mathbf{M}^{-1}\|_{\boldsymbol{\Sigma}}^2 = O_p(\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\boldsymbol{\Sigma}}^2)$, which proves Theorem 2. □

Lemma 1 *The eigenvalues of* $\widehat{\mathbf{A}}$ *are bounded and bounded away from zero, with probability at least* $1 - O(1/n^{C_1})$.

Proof of Lemma 1. Define

$$\mathbf{A} = E\left[\left(\overline{\mathbf{X}}_t \otimes \mathbf{B}(\mathbf{X}_{t-1})\right)\left(\overline{\mathbf{X}}_t^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1})\right)\right].$$

By assumptions (C1) and (C3), the eigenvalues of $\mathbf{A}$ are bounded and bounded away from zero.

For any $1 \le k, k' \le \mathcal{K}$ and $0 \le j_1, j_2, j_3, j_4 \le q$, we have

$$|B_k(x_{t-1,j_1})B_{k'}(x_{t-1,j_2})x_{t,j_3}x_{t,j_4}| \le \mathcal{K},$$

and

$$E[(B_k(x_{t-1,j_1})B_{k'}(x_{t-1,j_2})x_{t,j_3}x_{t,j_3})^2] \le \mathcal{K}E[(B_k(x_{t-1,j_1}))^2] \le C\mathcal{K}.$$

Thus

$$E[|B_k(x_{t-1,j_1})B_{k'}(x_{t-1,j_2})x_{t,j_3}x_{t,j_4}|^r] \le C\mathcal{K}^{r-2} \cdot \mathcal{K},\ r = 3, 4, \ldots.$$

Using Theorem 2.19 of [12] (setting $q = n/(C \log n)$ in that theorem with large enough C), for any $a > 0$,

$$P\left(\left|n^{-1}\sum_t B_k(x_{t-1,j_1})B_{k'}(x_{t-1,j_2})x_{t,j_3}x_{t,j_4} - E[B_k(x_{t-1,j_1})B_{k'}(x_{t-1,j_2})x_{t,j_3}x_{t,j_4}]\right| > a\right)$$
$$\le C(1 + \log n + \mu(a))\exp\{-C\frac{n}{\log n}\mu(a)\} + Cn(1 + \mathcal{K}^C/a)n^{-C_1},$$

where $\mu(a) = a^2/(\mathcal{K} + \mathcal{K}a)$, and the constant C_1 can be arbitrarily large. Setting $a = \delta/\mathcal{K}$, we get

$$P(\max_{k,k',j_1,j_2,j_3,j_4}\left|n^{-1}\sum_t B_k(x_{t-1,j_1})B_{k'}(x_{t-1,j_2})x_{t,j_3}x_{t,j_4} - E[B_k(x_{t-1,j_1})B_{k'}(x_{t-1,j_2})x_{t,j_3}x_{t,j_4}]\right| > \delta/\mathcal{K})$$
$$= O(1/n^{C_1}), \tag{A.12}$$

if $\mathcal{K} = O(n^c)$ with $0 < c < 1/3$. Thus $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{op} = o_t(1; 1/n^{C_1})$ and the proof is complete. □

Lemma 2 *We show that*

$$\left\|\sum_t \left(\overline{\mathbf{X}}_t \otimes \mathbf{B}(\mathbf{X}_{t-1})\right)\epsilon_{it}\right\| = O_t(\sqrt{n\mathcal{K}\log n}; \mathcal{K}^d(\log n)^{3d}/n^{d-1}).$$

Proof of Lemma 2. Denote $V_{it} = x_{t,j}B_k(x_{t-1,j'})\epsilon_{it}$, $V'_{it} = V_{it}I\{|V_{it}| \le b_n\}$ for some sequence b_n to be chosen later and $V''_{it} = V_{it} - V'_{it}$. We have $E[V_{it}^2] \le C$. Applying Theorem 2.18 (ii) of [12], with the quantity q in their theorem set to be $n/(C \log n)$ with C sufficiently large (which makes it possible that C_1 in the bound below can be arbitrarily large), we get

$$P(|\sum_t V'_{it} - E[V'_{it}]| > na) \le C\exp\left\{-C\frac{a^2 n/\log n}{1/\log n + b_n a}\right\} + O(1/n^{C_1}).$$

Furthermore,

$$P(|\sum_t V''_{it}| > na) \le P(\exists t, |V_{it}| > b_n) \le nE[|V_{it}|^{2d}]/b_n^{2d} \le C\frac{n\mathcal{K}^{d-1}}{b_n^{2d}}.$$

Since by Hölder's inequality,

$$E[|V''_{it}|] \le E[|V_{it}|^{2d}]^{\frac{1}{2d}}P(|V_{it}| > b_n)^{\frac{2d-1}{2d}} \le C\frac{\mathcal{K}^{d-1}}{b_n^{2d-1}},$$

if $a > C\frac{\mathcal{K}^{d-1}}{b_n^{2d-1}}$ we will have,

$$P(|\sum_t EV_{it}''| > na) = 0.$$

Combining the above bounds, we get that if $a > C\frac{\mathcal{K}^{d-1}}{b_n^{2d-1}}$

$$P(|\sum_t V_{it} - E[V_{it}]| > na) \leq C\exp\left\{-C\frac{a^2 n/\log n}{1/\log n + b_n a}\right\} + O(1/n^{C_1}) + C\frac{n\mathcal{K}^{d-1}}{b_n^{2d}}.$$

In particular, setting $a = C\sqrt{\log n/n}$, $b_n \sim \sqrt{n/\log^3 n}$, we get

$$\begin{aligned} &P(|\sum_t x_{t,j} B_k(x_{t-1,j'})\epsilon_{it}| > C\sqrt{n\log n}) \\ &\leq O(1/n^{C_1}) + O\left(\frac{\mathcal{K}^{d-1}(\log n)^{3d}}{n^{d-1}}\right). \end{aligned} \tag{A.13}$$

which implies the statement of the lemma. □

References

1. Avella-Madina M, Battey HS, Fan J, Li Q (2018) Robust estimation of high-dimensional covariance and precision matrices. Biometrika 105(2):271–284
2. Berthet Q, Rigollet P (2013) Optimal detection of sparse principal components in high dimension. Ann Statist 41(4):1780–1815
3. Bickel P, Levina E (2008) Covariance regularization by thresholding. Ann Statist 36(6):2577–2604
4. Bickel P, Levina E (2008) Regularized estimation of large covariance matrices. Ann Statist 36(1):199–227
5. Birnbaum A, Johnstone IM, Nadler B, Paul D (2013) Minimax bounds for sparse pca with noisy high-dimensional data. Ann Statist 41(3):1055–1084
6. Chen Z, Fan J, Li R (2018) Error variance estimation in ultrahigh dimensional additive models. J Amer Statist Assoc 113(521):315–327
7. EL Karoui N (2008) Operator norm consistent estimation of large-dimensional sparse covariance matrices. Ann Statist 36(6):2712–2756
8. Fama EF, French KR (1992) The cross-section of expected stock returns. J Finance 47(2):427–465
9. Fama EF, French KR (1993) Common risk factors in the returns on stocks and bonds. J Financ Econom 33(1):3–56
10. Fan J, Zhang W (1999) Statistical estimation in varying coefficient models. Ann Statist 27(5):1491–1518
11. Fan J, Zhang W (2000) Simultaneous confidence bands and hypothesis testing in varying-coefficient models. Scand J Statist 27(4):715–731
12. Fan J, Yao Q, Cai Z (2003) Adaptive varying-coefficient linear models. J R Stat Soc Ser B 65(1):57–80

13. Fan J, Fan Y, Lv J (2008) High dimensional covariance matrix estimation using a factor model. J Econometrics 147(1):186–197
14. Fan J, Liao Y, Mincheva M (2011) High-dimensional covariance matrix estimation in approximate factor models. Ann Statist 39(6):3320–3356
15. Fan J, Zhang J, Yu K (2012) Vast portfolio selection with gross-exposure constraints. J Amer Statist Assoc 107(498):592–606
16. Fan J, Liu H, Wang W (2018) Large covariance estimation through elliptical factor models. Ann Statist 46(4):1383–1414
17. Fang Y, Wang B, Feng Y (2016) Tuning-parameter selection in regularized estimations of large covariance matrices. Journal of Statistical Computation and Simulation 86:494–509
18. Francq J, Zakoïan JM (2010) GARCH Models: Struture, Statistical Inference and Financial Applications. John Wiley & Sons
19. Guo S, Box JL, Zhang W (2017) A dynamic structure for high dimensional covariance matrices and its application in portfolio allocation. J Amer Statist Assoc 112(517):235–253
20. Hastie H, Tibshirani R (1990) Generalized Additive Models. CRC Press
21. Hastie H, Tibshirani R (1993) Varying-coefficient models. Ann Statist 55(4):757–796
22. Kai B, Li R, Zou H (2011) New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. Ann Statist 39(1):305–332
23. Li J, Zhang W (2011) A semiparametric threshold model for censored longitudinal data analysis. J Amer Statist Assoc 106(494):685–696
24. Linton OB (1997) Miscellanea efficient estimation of additive nonparametric regression models. Biometrika 84(2):469–473
25. Rothman A, Levina E, Zhu J (2009) Generalized thresholding of large covariance matrices. J Amer Statist Assoc 104(485):177–186
26. Sun Y, Zhang W, Tong H (2007) Estimation of covariance matrix of random effects in longitudinal studies. Ann Statist 35(6):2795–2814
27. Sun Y, Yan H, Zhang W, Lu Z (2014) A semiparametric spatial dynamic model. Ann Statist 42(2):700–727
28. Wu WB, Pourahmadi M (2003) Nonparametric estimation of large covariance matrices of longitudinal data. Biometrika 90(4):831–884
29. Yuan M (2010) High dimensional inverse covariance matrix estimation via linear programming. J Mach Learn Res 11:2261–2286
30. Zhang W, Fan J, Sun Y (2009) A semiparametric model for cluster data. Ann Statist 37(5A):2377–2408