A New Semiparametric Estimation Approach for Large Dynamic Covariance Matrices with Multiple Conditioning Variables*

Jia Chen† Degui Li‡ Oliver Linton§

Version: October 24, 2018

Abstract

This paper studies the estimation of large dynamic covariance matrices with multiple conditioning variables. We introduce an easy-to-implement semiparametric method to estimate each entry of the covariance matrix via model averaging marginal regression, and then apply a shrinkage technique to obtain the dynamic covariance matrix estimation. Under some regularity conditions, we derive the asymptotic properties for the proposed estimators including the uniform consistency with general convergence rates. We further consider extending our methodology to deal with the scenarios: (i) the number of conditioning variables is divergent as the sample size increases, and (ii) the large covariance matrix is conditionally sparse relative to contemporaneous market factors. We provide a simulation study that illustrates the finite-sample performance of the developed methodology. We also provide an application to financial portfolio choice from daily stock returns.

Keywords: Dynamic covariance matrix, MAMAR, Semiparametric estimation, Sparsity, Uniform consistency.

*The authors would like to thank the Guest Editor Professor Eric Ghysels and two anonymous referees for their comments and suggestions, which substantially improve an earlier version of the paper.

†Department of Economics and Related Studies, University of York, York, YO10 5DD, UK. E-mail: jia.chen@york.ac.uk

‡Department of Mathematics, University of York, York, YO10 5DD, UK. E-mail: degui.li@york.ac.uk.

§Faculty of Economics, University of Cambridge, Cambridge, CB3 9DD, UK. E-mail: obl20@cam.ac.uk.

1 Introduction

The classical theory of mean/variance portfolio choice is developed by Markowitz (1952), see Merton (1969) and Fama (1970) for some other important developments. More recently this topic has been at the centre of a lot of research, see Back (2010) and Brandt (2010) for some recent surveys. In practice, it is not uncommon that the dynamic portfolio choice depends on many conditioning (or forecasting) variables, reflecting the varying investment opportunities over the time. Generally speaking, there are two ways to describe the dependence of portfolio choice on the conditioning variables. One is to assume a parametric model that relates the returns of risky assets to the conditioning variables and then solve for an investor's portfolio choice using some traditional econometric approaches to estimate the unknown parameters. However, the assumed parametric models might be misspecified, which would lead to inconsistent estimation of the optimal portfolio and invalid inference. One way to avoid the possible model misspecification issue is to use some nonparametric methods such as the kernel estimation method to describe the dependence of the portfolio choice on conditioning variables. The latter method is introduced by Brandt (1999) in the case of a univariate conditioning variable. Aït-Sahalia and Brandt (2001) further develop a single-index strategy to handle multiple conditioning variables. This literature has worked with the case where the number of assets is fixed and relatively small. However, another literature has considered the case where there is no covariate but there are a large number of assets (c.f., Ledoit and Wolf, 2003, 2004, 2017; Kan and Zhou, 2007; Fan, Fan and Lv, 2008; DeMiguel *et al*, 2009; DeMiguel, Garlappi and Uppal, 2009; Pesaran and Zaffaroni, 2009; Frahm and Memmel, 2010; Tu and Zhou, 2011).

As seen from the aforementioned literature, accurate covariance matrix estimation plays a crucial role in portfolio choice problem. Suppose that the observations $X_t = (X_{t1}, \ldots, X_{tN})^\intercal$, $t = 1, \ldots, T$, are collected from an N-dimensional stationary process with covariance matrix $E\left\{[X_t - E(X_t)]\,[X_t - E(X_t)]^\intercal\right\} = \Sigma$, where the matrix Σ is invariant over time. There have been extensive studies on estimating such a *static* covariance matrix. For instance, when the dimension is fixed or significantly smaller than the sample size T, Σ can be consistently estimated by the sample covariance matrix (c.f. Anderson, 2003):

$$\overline{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}(X_t - \overline{X})(X_t - \overline{X})^\intercal, \quad \overline{X} = \frac{1}{T}\sum_{t=1}^{T} X_t. \tag{1.1}$$

However, the above conventional sample covariance matrix would fail when the dimension N is large and exceeds the sample size T. In the latter case, the matrix $\overline{\Sigma}$ becomes singular. In order to obtain a proper estimation of Σ when $N > T$, some structural assumptions such as sparsity and factor modelling are usually imposed in the literature, and then various regularisation techniques

are used to produce consistent and reliable estimates (c.f., Wu and Pourahmadi, 2003; Bickel and Levina, 2008a,b; Lam and Fan, 2009; Rothman, Levina and Zhu, 2009; Cai and Liu, 2011; Fan, Liao and Mincheva, 2013).

The aforementioned literature on large covariance matrix estimation assumes that the underlying covariance matrix is constant over time. Such an assumption is very restrictive and may be violated in many practical applications such as in dynamic optimal portfolio allocation (Guo, Box and Zhang, 2017). This motivates us to consider a dynamic large covariance matrix, whose entries may evolve over time. In recent years, there have been increasing interests in estimating dynamic covariance or correlation matrices and exploring their applications. For example, Engle (2002) uses the parametric multivariate GARCH modelling method to estimate dynamic conditional correlation; Guo, Box and Zhang (2017) combine semiparametric adaptive functional-coefficient and GARCH modelling approaches to estimate dynamic covariance structure with the dimension diverging at a polynomial rate of the sample size; Chen, Xu and Wu (2013) and Chen and Leng (2016) use the kernel smoothing method to nonparametrically estimate each entry in the dynamic covariance matrix and then apply the thresholding or generalised shrinkage technique when the dimension N can be divergent at an exponential rate but the conditioning variable is univariate; Engle, Ledoit and Wolf (2017) extends Engle (2002)'s dynamic conditional correlation models to large dimensional case using a nonlinear shrinkage technique derived from the random matrix theory.

Let $U_t = (U_{t1}, \ldots, U_{tp})^\tau$ be a p-dimensional vector of conditioning variables which are stationary over time. We consider the conditional covariance matrix of X_{t+1} given U_t:

$$\mathbf{\Sigma}_0(u) = \mathsf{E}\left(X_{t+1}X_{t+1}^\tau | U_t = u\right) - \left[\mathsf{E}(X_{t+1}|U_t = u)\right]\left[\mathsf{E}(X_{t+1}|U_t = u)\right]^\tau,$$

where $u = (u_1, \ldots, u_p)^\tau$ is a vector of fixed constants. To simplify notation, we let

$$\mathcal{C}_0(u) = \mathsf{E}\left(X_{t+1}X_{t+1}^\tau | U_t = u\right) \text{ and } \mathcal{M}_0(u) = \mathsf{E}(X_{t+1}|U_t = u),$$

and rewrite the conditional covariance matrix as

$$\mathbf{\Sigma}_0(u) = \mathcal{C}_0(u) - \mathcal{M}_0(u)\mathcal{M}_0^\tau(u). \tag{1.2}$$

In order to estimate $\mathbf{\Sigma}_0(u)$, one only needs to estimate $\mathcal{C}_0(u)$ and $\mathcal{M}_0(u)$. A natural way to estimate $\mathcal{C}_0(u)$ and $\mathcal{M}_0(u)$ is via nonparametric smoothing. However, although the nonparametric estimation is robust to model misspecification, its finite-sample performance is often poor when the dimension of conditioning variables U_t is moderately large (or even as small as three), owing to the "curse of dimensionality". Therefore, when U_t is a multivariate vector, a direct use of the

nonparametric kernel approach as in Chen, Xu and Wu (2013) or Chen and Leng (2016) would be inappropriate, and an alternative technique is needed. The conditioning variables may be chosen as the lagged terms of some elements of X_t or some low-dimensional factor variables which can be either observable or latent. In practice, many variables including momentum measures, seasonal dummy variables, past earnings and transaction volume, have been used to predict the mean and variance of stock returns.

Letting $\sigma^0_{ij}(u)$ and $c^0_{ij}(u)$ be the (i,j)-entry of the matrices $\mathbf{\Sigma}_0(u)$ and $\mathcal{C}_0(u)$, respectively, and $m^0_i(u)$ be the i-th element of $\mathcal{M}_0(u)$, it follows from (1.2) that

$$\sigma^0_{ij}(u) = c^0_{ij}(u) - m^0_i(u)m^0_j(u), \ \ 1 \leqslant i,j \leqslant N. \tag{1.3}$$

Instead of estimating $m^0_i(u)$ and $c^0_{ij}(u)$ directly via nonparametric smoothing, we approximate them using the Model Averaging MArginal Regression (MAMAR) (Li, Linton and Lu, 2015), i.e.,

$$m^0_i(u) \approx b_{i,0} + \sum_{k=1}^{p} b_{i,k}\mathsf{E}(X_{t+1,i}|U_{tk} = u_k) =: b_{i,0} + \sum_{k=1}^{p} b_{i,k}m_{i,k}(u_k), \ \ 1 \leqslant i \leqslant N, \tag{1.4}$$

where $b_{i,k}$ are unknown parameters which may be regarded as "weights" for marginal mean regression models; and similarly for $c^0_{ij}(u)$

$$c^0_{ij}(u) \approx a_{ij,0} + \sum_{k=1}^{p} a_{ij,k}\mathsf{E}(X_{t+1,i}X_{t+1,j}|U_{tk} = u_k) =: a_{ij,0} + \sum_{k=1}^{p} a_{ij,k}c_{ij,k}(u_k), \ \ 1 \leqslant i,j \leqslant N, \tag{1.5}$$

where $a_{ij,k}$ are unknown weighting parameters. In (1.4) and (1.5), both $m_{i,k}(u_k)$ and $c_{ij,k}(u_k)$ are univariate nonparametric functions and can be well estimated by commonly used nonparametric methods without incurring the curse of dimensionality. The MAMAR method provides an alternative way to estimate nonparametric joint mean regression with multiple regressors. The MAMAR approximation is introduced by Li, Linton and Lu (2015) in a semiparametric time series setting, and is applied to semiparametric dynamic portfolio choice by Chen *et al* (2016) and further generalised to the ultra-high dimensional time series setting by Chen *et al* (2018). A similar idea is also used by Fan *et al* (2016) in high-dimensional classification.

The accuracy of the MAMAR approximation to the joint regression functions relies on the choice of the weight parameters, i.e., $b_{i,k}$ and $a_{ij,k}$ in (1.4) and (1.5), respectively. Section 2.1 below derives the theoretically optimal weights and consequently obtains a proxy, $\mathbf{\Sigma}^*_A(u)$, of the true dynamic covariance matrix $\mathbf{\Sigma}_0(u)$. A two-stage semiparametric method is proposed to estimate each entry of $\mathbf{\Sigma}^*_A(u)$: in stage 1, the kernel smoothing method is used to estimate the marginal regression functions $m_{i,k}(u_k)$ and $c_{ij,k}(u_k)$; in stage 2, the least squares method is used to

estimate the optimal weights in the MAMAR approximation by replacing the marginal regression functions with their estimates obtained from stage 1 and then treating them as "regressors" in approximate linear models associated with $m_i^0(u)$ and $c_{ij}^0(u)$. Based on the above, an estimate of the optimal MAMAR approximation of $m_i^0(u)$ and $c_{ij}^0(u)$ can be constructed via (1.4) and (1.5), and subsequently the optimal MAMAR approximation of $\sigma_{ij}^0(u)$ can be estimated via (1.3). Finally, a generalised shrinkage technique is applied to the obtained covariance matrix to produce a non-degenerate estimate that has its small entries forced to be zero. Under some mild conditions and the assumption that $\mathbf{\Sigma}_A^\star(u)$ is approximately sparse, we derive the uniform consistency results for estimators of $\mathbf{\Sigma}_A^\star(u)$ and its inverse. These results also hold for the true covariance matrix $\mathbf{\Sigma}_0(u)$ as long as $\mathbf{\Sigma}_A^\star(u)$ and $\mathbf{\Sigma}_0(u)$ are sufficiently "close". The sparsity result for the semiparametric shrinkage estimator is established as well. In addition, we introduce two modification techniques in semiparametric large covariance matrix estimation to guarantee that the estimated covariance matrix is positive definite in finite samples.

Two interesting extensions of the developed semiparametric MAMAR covariance matrix estimation are explored in the present paper. The first one is to consider the case where the number of conditioning variables is divergent as either T or N increases. A penalised MAMAR method is introduced to simultaneously estimate each entry in the large dynamic covariance matrix and select significant conditioning variables. The second extension is to consider the more general setting that the high-dimensional time series random vector X_t is driven by some latent common factors and satisfy the approximate factor model structure. In this case, the approximate sparsity assumption on the covariance structure of X_t is weakened to the so-called conditional sparsity. The principal component method is applied to estimate the factors and factor loadings (up to an appropriate rotation) as well as the idiosyncratic errors. The semiparametric MAMAR method is used to estimate the dynamic covariance matrices for the low-dimensional estimated factor vector and high-dimensional estimated error vector, which are then combined with the generalised shrinkage technique to obtain the dynamic covariance matrix estimation of X_t.

The rest of the paper is organised as follows. Section 2 derives the optimal weights in the MAMAR approximation (1.4) and (1.5), and introduces the semiparametric shrinkage method to estimate the dynamic covariance matrix. Section 3 gives the limit theorems of the developed estimators. Section 4 introduces two modification techniques to guarantee the positive definiteness of the dynamic covariance matrix estimation, and discusses the choice of tuning parameter in the generalised shrinkage method. Section 5 reports finite-sample simulation studies of our methodology as well as an application to construct global minimum variance portfolios using the data consisted of 500 daily returns on stocks listed on New York Stock Exchange. Section 6 discusses some possible extensions and Section 7 concludes the paper. The proofs of the main results and some technical lemmas are given in the appendix. Throughout the paper, we use

$\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ to denote the minimum and maximum eigenvalues of a matrix; $\|\cdot\|_O$ to denote the operator (or spectral) norm defined as $\|\boldsymbol{\Delta}\|_O = \sup_{\boldsymbol{x}}\{\|\boldsymbol{\Delta}\boldsymbol{x}\| : \|\boldsymbol{x}\| = 1\}$ for a $q \times q$ matrix $\boldsymbol{\Delta} = (\delta_{ij})_{q\times q}$, where $\|\boldsymbol{x}\| = \left(\sum_{i=1}^{q} x_i^2\right)^{1/2}$ is the Euclidean norm; and $\|\cdot\|_F$ to denote the Frobenius norm defined as$\|\boldsymbol{\Delta}\|_F = \left(\sum_{i=1}^{q}\sum_{j=1}^{q}\delta_{ij}^2\right)^{1/2} = \left[\mathsf{Tr}(\boldsymbol{\Delta}\boldsymbol{\Delta}^\intercal)\right]^{1/2}$, where $\mathsf{Tr}(\cdot)$ denotes the trace of a matrix.

2 Estimation methodology

In this section we introduce an estimation method for the dynamic covariance matrix via the MAMAR approximation. It combines a semiparametric least squares method and the generalised shrinkage technique to produce reliable large covariance matrix estimation. We start with an introduction of the MAMAR approximation in our context and then derive the theoretically optimal weights in the approximation.

2.1 Optimal weights in the MAMAR approximation

For each $k = 0, 1, \ldots, p$, let $\mathcal{A}_k = (a_{ij,k})_{N\times N}$ be a matrix consisting of the weights in (1.5) and $\mathcal{C}_k(u_k) = [c_{ij,k}(u_k)]_{N\times N}$ be a matrix consisting of the conditional means of $X_{t+1,i}X_{t+1,j}$ (for given $U_{tk} = u_k$) in (1.5). Then, the MAMAR approximation for $\mathcal{C}_0(u)$ can be written in matrix form as

$$\mathcal{C}_0(u) \approx \mathcal{A}_0 + \mathcal{A}_1 \odot \mathcal{C}_1(u_1) + \cdots + \mathcal{A}_p \odot \mathcal{C}_p(u_p) =: \mathcal{C}_A(u), \tag{2.1}$$

where $\odot$ denotes the Hadamard product. Similarly, we have the following MAMAR approximation for $\mathcal{M}_0(u)$

$$\mathcal{M}_0(u) \approx \mathcal{B}_0 + \mathcal{B}_1 \odot \mathcal{M}_1(u_1) + \cdots + \mathcal{B}_p \odot \mathcal{M}_p(u_p) =: \mathcal{M}_A(u), \tag{2.2}$$

where for $k = 0, 1, \ldots, p$, $\mathcal{B}_k = (b_{1,k}, b_{2,k}, \ldots, b_{N,k})^\intercal$ is a vector consisting of the weights in (1.4) and $\mathcal{M}_k(u_k) = [m_{1,k}(u_k), m_{2,k}(u_k), \ldots, m_{N,k}(u_k)]^\intercal$ is a vector consisting of the conditional means of $X_{t+1,i}$ (for given $U_{tk} = u_k$) in (1.4). Combining (2.1) and (2.2), we readily have the following MAMAR approximation for $\boldsymbol{\Sigma}_0(u)$

$$\begin{aligned}\boldsymbol{\Sigma}_0(u) &\approx \left[\mathcal{A}_0 + \sum_{k=1}^{p}\mathcal{A}_k \odot \mathcal{C}_k(u_k)\right] - \left[\mathcal{B}_0 + \sum_{k=1}^{p}\mathcal{B}_k \odot \mathcal{M}_k(u_k)\right]\left[\mathcal{B}_0 + \sum_{k=1}^{p}\mathcal{B}_k \odot \mathcal{M}_k(u_k)\right]^\intercal \\ &= \mathcal{C}_A(u) - \mathcal{M}_A(u)\mathcal{M}_A^\intercal(u) =: \boldsymbol{\Sigma}_A(u).\end{aligned} \tag{2.3}$$

The matrix $\boldsymbol{\Sigma}_A(u)$ on the right hand side of (2.3) can be viewed as a semiparametric approxi-

mation of $\mathbf{\Sigma}_0(u)$, in which the weights $a_{ij,k}$ and $b_{i,k}$ play an important role. These weights have to be appropriately chosen in order to achieve optimal MAMAR approximation. We next derive the theoretically optimal weights. For $1 \leqslant i, j \leqslant N$, we may choose the optimal weights $a^{\star}_{ij,k}$, $k = 0, 1, \ldots, p$, so that they minimise

$$\mathsf{E}\left[X_{t+1,i}X_{t+1,j} - a_{ij,0} - \sum_{k=1}^{p} a_{ij,k}\mathsf{E}(X_{t+1,i}X_{t+1,j}|U_{tk})\right]^2.$$

Following standard calculations (c.f., Li, Linton and Lu, 2015), we have the following solution for the theoretically optimal weights

$$\left(a^{\star}_{ij,1}, \ldots, a^{\star}_{ij,p}\right)^{\intercal} = \mathbf{\Omega}^{-1}_{XX,ij}\mathbf{V}_{XX,ij}, \;\; a^{\star}_{ij,0} = \left(1 - \sum_{k=1}^{p} a^{\star}_{ij,k}\right)\mathsf{E}(X_{ti}X_{tj}), \tag{2.4}$$

where $\mathbf{\Omega}_{XX,ij}$ is a $p \times p$ matrix with the (k, l)-entry being

$$\omega_{ij,kl} = \mathsf{Cov}\left[\mathsf{E}(X_{t+1,i}X_{t+1,j}|U_{tk}), \mathsf{E}(X_{t+1,i}X_{t+1,j}|U_{tl})\right] = \mathsf{Cov}\left[c_{ij,k}(U_{tk}), c_{ij,l}(U_{tl})\right],$$

and $\mathbf{V}_{XX,ij}$ is a p-dimensional column vector with the k-th element being

$$v_{ij,k} = \mathsf{Cov}\left[\mathsf{E}(X_{t+1,i}X_{t+1,j}|U_{tk}), X_{t+1,i}X_{t+1,j}\right] = \mathsf{Cov}\left[c_{ij,k}(U_{tk}), X_{t+1,i}X_{t+1,j}\right] = \mathsf{Var}\left[c_{ij,k}(U_{tk})\right].$$

We thus obtain the optimal weight matrix $\mathcal{A}^{\star}_{k}$ from $a^{\star}_{ij,k}$, $k = 0, 1, \ldots, p$, and subsequently the theoretically optimal MAMAR approximation to $\mathcal{C}_0(u)$:

$$\mathcal{C}^{\star}_{A}(u) := \mathcal{A}^{\star}_{0} + \sum_{k=1}^{p} \mathcal{A}^{\star}_{k} \odot \mathcal{C}_k(u_k). \tag{2.5}$$

Similarly, we can also derive the optimal weights $b^{\star}_{i,k}$ in the MAMAR approximation (1.4) and consequently obtain the optimal weight vector $\mathcal{B}^{\star}_{k}$ from $b^{\star}_{i,k}$, $k = 0, 1, \ldots, p$. The definitions of $b^{\star}_{i,k}$ are analogous to $a^{\star}_{ij,k}$ but with $X_{t+1,i}X_{t+1,j}$ and $c_{ij,k}(U_{tk})$ replaced by $X_{t+1,i}$ and $m_{i,k}(U_{tk})$, respectively. The optimal MAMAR approximation to $\mathcal{M}_0(u)$ is obtained via

$$\mathcal{M}^{\star}_{A}(u) := \mathcal{B}^{\star}_{0} + \sum_{k=1}^{p} \mathcal{B}^{\star}_{k} \odot \mathcal{M}_k(u_k). \tag{2.6}$$

Combining (2.3), (2.5) and (2.6), we obtain the optimal MAMAR approximation to $\mathbf{\Sigma}_0(u)$:

$$\mathbf{\Sigma}^{\star}_{A}(u) := \mathcal{C}^{\star}_{A}(u) - \mathcal{M}^{\star}_{A}(u)\left[\mathcal{M}^{\star}_{A}(u)\right]^{\intercal} \tag{2.7}$$

The matrix $\mathbf{\Sigma}_A^*(u)$ serves as a proxy for $\mathbf{\Sigma}_0(u)$. Throughout the paper, we assume that $\mathbf{\Sigma}_A^*(u)$ is positive definite uniformly over u. Our aim is to consistently estimate $\mathbf{\Sigma}_A^*(u)$. This will be done by a semiparametric shrinkage method.

2.2 Semiparametric shrinkage estimation

We next introduce a two-stage semiparametric method to estimate $m_i^0(u)$ and $c_{ij}^0(u)$, respectively.

STAGE 1. As both $m_{i,k}(u_k)$ and $c_{ij,k}(u_k)$ are univariate functions, they can be well estimated by the kernel method, i.e.,

$$\widehat{m}_{i,k}(u_k) = \left[\sum_{t=1}^{T-1} K\left(\frac{U_{tk}-u_k}{h_1}\right) X_{t+1,i}\right] / \left[\sum_{t=1}^{T-1} K\left(\frac{U_{tk}-u_k}{h_1}\right)\right], \ 1 \leqslant k \leqslant p, \ 1 \leqslant i \leqslant N,$$

and

$$\widehat{c}_{ij,k}(u_k) = \left[\sum_{t=1}^{T-1} K\left(\frac{U_{tk}-u_k}{h_2}\right) X_{t+1,i}X_{t+1,j}\right] / \left[\sum_{t=1}^{T-1} K\left(\frac{U_{tk}-u_k}{h_2}\right)\right], \ 1 \leqslant k \leqslant p, \ 1 \leqslant i,j \leqslant N,$$

where $K(\cdot)$ is a kernel function, h_1 and h_2 are two bandwidths. Other nonparametric estimation methods such as the local polynomial method (Fan and Gijbels, 1996) and the sieve method (Chen, 2007) are equally applicable here.

STAGE 2. With the kernel estimates in stage 1, we have the following approximate linear regression models:

$$X_{t+1,i} \approx b_{i,0} + \sum_{k=1}^{p} b_{i,k}\widehat{m}_{i,k}(U_{tk}), \ \ 1 \leqslant i \leqslant N, \tag{2.8}$$

and

$$X_{t+1,i}X_{t+1,j} \approx a_{ij,0} + \sum_{k=1}^{p} a_{ij,k}\widehat{c}_{ij,k}(U_{tk}), \ \ 1 \leqslant i,j \leqslant N. \tag{2.9}$$

Treating $\widehat{c}_{ij,k}(U_{tk})$ as "regressors" in (2.9) and using the ordinary least squares, we may obtain an estimate of the optimal weights defined in (2.4):

$$(\widehat{a}_{ij,1},\ldots,\widehat{a}_{ij,p})^\tau = \widehat{\mathbf{\Omega}}_{XX,ij}^{-1}\widehat{\mathbf{V}}_{XX,ij} \tag{2.10}$$

and

$$\widehat{a}_{ij,0} = \frac{1}{T-1}\sum_{t=1}^{T-1} X_{t+1,i}X_{t+1,j} - \sum_{k=1}^{p}\widehat{a}_{ij,k}\left(\frac{1}{T-1}\sum_{t=1}^{T-1}\widehat{c}_{ij,k}(U_{tk})\right), \tag{2.11}$$

where $\widehat{\mathbf{\Omega}}_{XX,ij}$ is a $p \times p$ matrix with the (k,l)-entry being

$$\widehat{\omega}_{ij,kl} = \frac{1}{T-1}\sum_{t=1}^{T-1} \widehat{c}^{c}_{ij,k}(U_{tk})\widehat{c}^{c}_{ij,l}(U_{tl}), \;\; \widehat{c}^{c}_{ij,k}(U_{tk}) = \widehat{c}_{ij,k}(U_{tk}) - \frac{1}{T-1}\sum_{s=1}^{T-1} \widehat{c}_{ij,k}(U_{sk}),$$

and $\widehat{\mathbf{V}}_{XX,ij}$ is a p-dimensional column vector with the k-th element being

$$\widehat{v}_{ij,k} = \frac{1}{T-1}\sum_{t=1}^{T-1} \widehat{c}^{c}_{ij,k}(U_{tk})X^{c}_{t+1,(i,j)}, \;\; X^{c}_{t+1,(i,j)} = X_{t+1,i}X_{t+1,j} - \frac{1}{T-1}\sum_{s=1}^{T-1} X_{s+1,i}X_{s+1,j}.$$

Analogously, treating $\widehat{m}_{i,k}(U_{tk})$ as "regressors" in (2.8) and then using the ordinary least squares, we may obtain estimates of the optimal weights $b^{\star}_{i,k}$, $k = 0, 1, \ldots, p$, defined in Section 2.1, and denote them by $\widehat{b}_{i,k}$, $k = 0, 1, \ldots, p$. The definitions of $\widehat{b}_{i,k}$ are similar to those of $\widehat{a}_{ij,k}$ in (2.10) and (2.11) but with $X_{t+1,i}X_{t+1,j}$ and $\widehat{c}_{ij,k}(U_{tk})$ replaced by $X_{t+1,i}$ and $\widehat{m}_{i,k}(U_{tk})$, respectively.

As a result, an estimate of $\sigma^{\star}_{ij}(u)$, the (i,j)-entry in $\mathbf{\Sigma}^{\star}_{\mathcal{A}}(u)$, can be obtained as

$$\widehat{\sigma}_{ij}(u) = \widehat{c}_{ij}(u) - \widehat{m}_{i}(u)\widehat{m}_{j}(u), \tag{2.12}$$

where

$$\widehat{c}_{ij}(u) = \widehat{a}_{ij,0} + \sum_{k=1}^{p} \widehat{a}_{ij,k}\widehat{c}_{ij,k}(u_k), \;\; \widehat{m}_{i}(u) = \widehat{b}_{i,0} + \sum_{k=1}^{p} \widehat{b}_{i,k}\widehat{m}_{i,k}(u_k).$$

A naive estimate, $\widehat{\mathbf{\Sigma}}(u)$, of $\mathbf{\Sigma}^{\star}_{\mathcal{A}}(u)$ uses $\widehat{\sigma}_{ij}(u)$ directly as its entries, i.e.,

$$\widehat{\mathbf{\Sigma}}(u) = [\widehat{\sigma}_{ij}(u)]_{N\times N}.$$

Unfortunately, this matrix gives a poor estimation of $\mathbf{\Sigma}_0(u)$ when the dimension N is ultra large. In this case, a commonly-used approach is to use a shrinkage method on $\widehat{\mathbf{\Sigma}}(u)$ so that very small values of $\widehat{\sigma}_{ij}(u)$ are forced to zero. We follow the same approach and denote $s_\rho(\cdot)$ a shrinkage function that satisfies the following three conditions: (i) $|s_\rho(z)| \leqslant |z|$ for $z \in \mathcal{R}$ (the real line); (ii) $s_\rho(z) = 0$ if $|z| \leqslant \rho$; (iii) $|s_\rho(z) - z| \leqslant \rho$, where ρ is a tuning parameter that controls the amount of shrinkage. It is easy to show that some commonly-used shrinkage methods including the hard thresholding, soft thresholding and SCAD satisfy the three conditions. To allow for different amount of shrinkage for different values of u, we will use a variable tuning parameter, $\rho(u)$ and define

$$\widetilde{\sigma}_{ij}(u) = s_{\rho(u)}\left(\widehat{\sigma}_{ij}(u)\right), \;\; 1 \leqslant i,j \leqslant N. \tag{2.13}$$

Then let

$$\widetilde{\mathbf{\Sigma}}(u) = [\widetilde{\sigma}_{ij}(u)]_{N\times N}, \tag{2.14}$$

be the shrinkage estimator of $\mathbf{\Sigma}_A^*(u)$. The asymptotic properties of $\widetilde{\mathbf{\Sigma}}(u)$ will be explored in Section 3 below. Since thus defined $\widetilde{\mathbf{\Sigma}}(u)$ is not necessarily positive definite in finite samples, in Section 4.1 we will introduce two methods to modify $\widetilde{\mathbf{\Sigma}}(u)$ to guarantee the positive definiteness of the resulting covariance matrix estimates.

3 Large-sample theory

In this section we first state the regularity conditions required for establishing the limit theorems of the large dynamic covariance matrix estimators developed in Section 2, and then present these theorems in Section 3.2.

3.1 Technical assumptions

Some of the assumptions presented below may not be the weakest possible, but they are imposed to facilitate proofs of our limit theorems and can be relaxed at the cost of more lengthy proofs.

ASSUMPTION 1. (i) The process $\{(X_t, U_t)\}_{t\geqslant 1}$ is stationary and α-mixing dependent with the mixing coefficient decaying to zero at a geometric rate, i.e., $\alpha_k \sim c_\alpha \gamma^k$ with $0 < \gamma < 1$ and c_α being a positive constant.

(ii) The variables X_{ti}, $1 \leqslant i \leqslant N$, satisfy the following moment condition:

$$\max_{1\leqslant i\leqslant N} \mathsf{E}\left[\exp\{sX_{ti}^2\}\right] \leqslant c_X, \quad 0 < s \leqslant s_0, \tag{3.1}$$

where c_X and s_0 are two positive constants.

(iii) The conditioning variables U_t have a compact support denoted by $\mathcal{U} = \prod_{k=1}^p \mathcal{U}_k$, where $\mathcal{U}_k = [a_k, b_k]$ is the support of the k-th conditional variable U_{tk}. The marginal density functions of U_{tk}, $f_k(\cdot)$, $1 \leqslant k \leqslant p$, are continuous and uniformly bounded away from zero on $\mathcal{U}_k$, i.e., there exists a positive constant c_f such that

$$\min_{1\leqslant k\leqslant p} \inf_{a_k\leqslant u_k\leqslant b_k} f_k(u_k) \geqslant c_f > 0.$$

In addition, the marginal density functions $f_k(\cdot)$, $1 \leqslant k \leqslant p$, have continuous derivatives up to the second order.

ASSUMPTION 2. (i) The marginal regression functions $c_{ij,k}(\cdot)$ and $m_{i,k}(\cdot)$ are continuous and uniformly bounded over $1 \leqslant i, j \leqslant N$ and $1 \leqslant k \leqslant p$. Furthermore, they have continuous and uniformly bounded derivatives up to the second order.

(ii) For $1 \leqslant i, j \leqslant N$, the $p \times p$ matrix $\mathbf{\Omega}_{XX,ij}$ defined in (2.4) is positive definite and satisfies

$$0 < \underline{c}_{\mathbf{\Omega}_{XX}} \leqslant \min_{1\leqslant i,j\leqslant N} \lambda_{\min}(\mathbf{\Omega}_{XX,ij}) \leqslant \max_{1\leqslant i,j\leqslant N} \lambda_{\max}(\mathbf{\Omega}_{XX,ij}) \leqslant \bar{c}_{\mathbf{\Omega}_{XX}} < \infty. \tag{3.2}$$

The analogous condition also holds for the matrix $\mathbf{\Omega}_{X,i}$ whose (k,l)-entry is defined as

$$\omega_{i,kl} = \mathsf{Cov}\left[\mathsf{E}(X_{t+1,i}|U_{tk}), \mathsf{E}(X_{t+1,i}|U_{tl})\right] = \mathsf{Cov}\left[m_{i,k}(U_{tk}), m_{i,l}(U_{tl})\right].$$

ASSUMPTION 3. (i) The kernel function $K(\cdot)$ is symmetric and Lipschitz continuous and has a compact support, say $[-1,1]$.

(ii) The bandwidths h_1 and h_2 satisfy $h_1 \to 0$ and $h_2 \to 0$, and there exists $0 < \iota < 1/2$ so that

$$\frac{T^{1-\iota}h_1}{\log^3(N \vee T)} \to \infty, \quad \frac{T^{1-2\iota}h_2}{\log^3(N \vee T)} \to \infty, \tag{3.3}$$

where $x \vee y$ denotes the maximum of x and y.

(iii) The dimension, N, of X satisfies $(NT)\exp\{-sT^{\iota}\} = o(1)$ for some $0 < s < s_0$, where ι was defined in Assumption 3(ii) and s_0 was defined in Assumption 1(ii).

ASSUMPTION 4. The variable tuning parameter can be written as $\rho(u) = M_0(u)\tau_{T,N}$, where $M_0(u)$ is a positive function satisfying $c_M < \inf_{u\in\mathcal{U}} M_0(u) \leqslant \sup_{u\in\mathcal{U}} M_0(u) < \infty$ with c_M being a sufficiently large positive constant, and

$$\tau_{T,N} = \sqrt{\log(N \vee T)/(Th_1)} + \sqrt{\log(N \vee T)/(Th_2)} + h_1^2 + h_2^2.$$

Most of the above assumptions are commonly used and can be found in some existing literature. The stationarity and α-mixing dependence condition in Assumption 1(i) relaxes the restriction of independent observations usually imposed in the literature on high-dimensional covariance matrix estimation (c.f. Bickel and Levina, 2008a,b). For some classic vector time series processes such as vector auto-regressive processes, it is easy to verify Assumption 1(i) under some mild conditions. It is possible to allow the even more general setting of local stationarity, say Vogt (2012), which includes deterministic local trends, but for simplicity we have chosen not to go there. The moment condition (3.1) is similar to those in Bickel and Levina (2008a,b) and Chen and Leng (2016), and can be replaced by the weaker condition of $\mathsf{E}(|X_{ti}|^{\kappa})$ for $\kappa > 2$ sufficiently large if the dimension N diverges at a polynomial rate of T. The restriction of the conditioning variables U_t having a compact support in Assumption 1(iii) is imposed mainly in order to facilitate the proofs of our asymptotic results and can be removed by using an appropriate truncation technique on U_t (c.f., Remark 1 in Chen *et al*, 2018). The smoothness condition on the marginal

regression functions in Assumption 2(i) is commonly used in kernel smoothing, and it is relevant to the asymptotic bias of the kernel estimators (c.f., Wand and Jones, 1995). Assumption 2(ii) is crucial to the unique existence of optimal weights in the MAMAR approximation of $c_{ij}^0(\cdot)$ and $m_i^0(\cdot)$. Many commonly-used kernel functions, such as the uniform kernel and the Epanechnikov kernel, satisfy the conditions in Assumption 3(i). The conditions in Assumptions 3(ii) and (iii) indicate that the dimension N can be divergent at an exponential rate of T. For example, when h_1 and h_2 have the well-known optimal rate of $T^{-1/5}$, we may show that N can be divergent at a rate of $\exp\{T^{\zeta}\}$ with $0 < \zeta < 1/10$ while Assumptions 3(ii) and (iii) hold. Assumption 4 is critical to ensure the validity of the shrinkage method, and Section 4.2 below will discuss how to select $\rho(u)$ in numerical studies.

3.2 Asymptotic properties

In order to derive some sensible asymptotic results for the dynamic covariance matrix estimators defined in Section 2.2, we extend the sparsity assumption in Bickel and Levina (2008a), Rothman, Levina and Zhu (2009) and Cai and Liu (2011) and assume that $\mathbf{\Sigma}_A^\star(u)$ is *approximately sparse* uniformly over $u \in \mathcal{U}$. Specifically, this means that there exist c_N and $M_\star$ such that $\mathbf{\Sigma}_A^\star(\cdot) \in \mathcal{S}_A(q, c_N, M_\star, \mathcal{U})$, where

$$\mathcal{S}_A(q, c_N, M_\star, \mathcal{U}) = \left\{ \mathbf{\Sigma}(u), u \in \mathcal{U} \,\middle|\, \sup_{u\in\mathcal{U}} \sigma_{ii}(u) \leqslant M_\star < \infty, \sup_{u\in\mathcal{U}} \sum_{j=1}^{N} |\sigma_{ij}(u)|^q \leqslant c_N \; \forall\, 1 \leqslant i \leqslant N \right\} \tag{3.4}$$

with $0 \leqslant q < 1$. In particular, if $q = 0$, $\mathcal{S}_A(q, c_N, M_\star, \mathcal{U})$ becomes

$$\mathcal{S}_A(0, c_N, M_\star, \mathcal{U}) = \left\{ \mathbf{\Sigma}(u), u \in \mathcal{U} \,\middle|\, \sup_{u\in\mathcal{U}} \sigma_{ii}(u) \leqslant M_\star < \infty, \sup_{u\in\mathcal{U}} \sum_{j=1}^{N} I\left(|\sigma_{ij}(u)| \neq 0\right) \leqslant c_N \; \forall\, 1 \leqslant i \leqslant N \right\},$$

and we have $\mathbf{\Sigma}_A^\star(\cdot) \in \mathcal{S}_A(0, c_N, M_\star, \mathcal{U})$, the *exact sparsity* assumption, uniformly over $u \in \mathcal{U}$. The above assumption has been used by Chen and Leng (2016) for nonparametric estimation of large covariance matrices and would facilitate the asymptotic analysis of the developed semiparametric large covariance matrix estimation. However, the sparsity assumption may be violated in some empirical applications such as the covariance structure of vast stock returns. Such an issue will be addressed later in Section 6.2, where the semiparametric shrinkage method will be generalised to estimate large dynamic covariance matrices in the more general setting of conditional sparsity.

Define $\mathcal{U}_{h_\star} = \prod_{k=1}^{p} \mathcal{U}_{k,h_\star}$ with $\mathcal{U}_{k,h_\star} = [a_k + h_\star, b_k - h_\star]$ and $h_\star = h_1 \vee h_2$. Without loss of generality, we assume that, for each $1 \leqslant k \leqslant p$, all of the observations U_{tk}, $1 \leqslant t \leqslant T$, are located in the intervals $[a_k + h_\star, b_k - h_\star]$ (otherwise a truncation technique can be applied when

constructing the semiparametric estimators defined in Section 2.2). Theorem 1 below gives the uniform consistency of the semiparametric shrinkage estimator of $\mathbf{\Sigma}^*_A(u)$ and its inverse over $\mathcal{U}_{h_\star}$. The main reason for considering the uniform consistency only over the set $\mathcal{U}_{h_\star}$ rather than the whole support $\mathcal{U}$ of the conditioning variables is to avoid the boundary effect in kernel estimation (c.f., Fan and Gijbels, 1996).

THEOREM 1. *Suppose that Assumptions 1–4 are satisfied,* p *is fixed, and* $\mathbf{\Sigma}^*_A(\cdot) \in \mathcal{S}_A(q, c_N, M_\star, \mathcal{U})$.

(i) For $\widetilde{\mathbf{\Sigma}}(u)$*, we have*

$$\sup_{u\in\mathcal{U}_{h_\star}} \left\|\widetilde{\mathbf{\Sigma}}(u) - \mathbf{\Sigma}^*_A(u)\right\|_O = O_P\left(c_N \cdot \tau_{T,N}^{1-q}\right), \ 0 \leqslant q < 1, \tag{3.5}$$

where $\tau_{T,N}$ *was defined in Assumption 4 and* $\|\cdot\|_O$ *denotes the operator norm.*

(ii) If, in addition, $c_N\tau_{T,N}^{1-q} = o(1)$ *and there exists a positive constant* $c_{\mathbf{\Sigma}}$ *such that*

$$\inf_{u\in\mathcal{U}} \lambda_{\min}\left(\mathbf{\Sigma}^*_A(u)\right) \geqslant c_{\mathbf{\Sigma}} > 0, \tag{3.6}$$

then we have

$$\sup_{u\in\mathcal{U}_{h_\star}} \left\|\widetilde{\mathbf{\Sigma}}^{-1}(u) - \mathbf{\Sigma}^{*-1}_A(u)\right\|_O = O_P\left(c_N \cdot \tau_{T,N}^{1-q}\right), \ 0 \leqslant q < 1. \tag{3.7}$$

The uniform convergence rate in the above theorem is quite general. Its dependence on the sparsity structure of the matrix $\mathbf{\Sigma}^*_A(u)$ is shown through c_N, which controls the sparsity level in the covariance matrix and may be divergent to infinity. If we assume that $h_1 = h_2 = h$ and $h^2 = O\left(\sqrt{\log(N \vee T)/(Th)}\right)$, $\tau_{T,N}$ can be simplified to $\sqrt{\log(N \vee T)/(Th)}$. Then we may find that our uniform convergence rate is comparable to the rate derived by Bickel and Levina (2008a) and Rothman, Levina and Zhu (2009) if we treat Th as the "effective" sample size in nonparametric kernel-based estimation. In the special case of $q = 0$ and fixed N, $\log(N \vee T) = \log T$ and it would be reasonable to assume that c_N is fixed. Consequently, the rate in (3.5) and (3.7) reduces to

$$O_P(\tau_{T,N}) = O_P\left(\sqrt{\log T/(Th_1)} + \sqrt{\log T/(Th_2)} + h_1^2 + h_2^2\right),$$

the same as the uniform convergence rate for nonparametric kernel-based estimators (c.f., Bosq, 1998).

If we assume that the true dynamic covariance matrix $\mathbf{\Sigma}_0(\cdot)$ belongs to $\mathcal{S}_A(q, c_N, M_\star, \mathcal{U})$, and $\mathbf{\Sigma}^*_A(u)$ is sufficiently close to $\mathbf{\Sigma}_0(u)$ in the sense that there exists $b_{T,N}$ with $b_{T,N} \to 0$ such that $\sup_{u\in\mathcal{U}} \left\|\mathbf{\Sigma}^*_A(u) - \mathbf{\Sigma}_0(u)\right\|_O = O(b_{T,N})$ and $\max_{1\leqslant i,j\leqslant N} \sup_{u\in\mathcal{U}} \left|\sigma^*_{ij}(u) - \sigma^0_{ij}(u)\right| = O(\tau_{T,N})$, by The-

orem 1 and its proof in Appendix A, we may show that

$$\sup_{u\in\mathcal{U}_{h_\star}}\left\|\widetilde{\boldsymbol{\Sigma}}(u)-\boldsymbol{\Sigma}_0(u)\right\|_O \leqslant \sup_{u\in\mathcal{U}_{h_\star}}\left\|\widetilde{\boldsymbol{\Sigma}}(u)-\boldsymbol{\Sigma}_A^\star(u)\right\|_O + \sup_{u\in\mathcal{U}_{h_\star}}\left\|\boldsymbol{\Sigma}_A^\star(u)-\boldsymbol{\Sigma}_0(u)\right\|_O$$
$$= O_P\left(c_N\cdot\tau_{T,N}^{1-q}\right)+O\left(b_{T,N}\right)=O_P\left(c_N\cdot\tau_{T,N}^{1-q}+b_{T,N}\right). \tag{3.8}$$

Therefore, if the rate $b_{T,N}$ is dominated by $c_N\cdot\tau_{T,N}^{1-q}$, the uniform convergence rate in (3.5) still applies when $\boldsymbol{\Sigma}_A^\star(u)$ is replaced by $\boldsymbol{\Sigma}_0(u)$, the true dynamic covariance matrix. This shows that the accuracy of the proposed semiparametric MAMAR approximation and general shrinkage technique relies on the closeness of $\boldsymbol{\Sigma}_A^\star(u)$ to $\boldsymbol{\Sigma}_0(u)$. In fact, if the MAMAR approximation in the covariance structure is poor, $\boldsymbol{\Sigma}_A^\star(u)$ might not satisfy the sparsity structure even when the true dynamic covariance matrix $\boldsymbol{\Sigma}_0(u)$ is sparse, leading to poor numerical performance in finite samples.

The following theorem shows the sparsity property of the semiparametric shrinkage estimator defined in Section 2.2.

THEOREM 2. *Suppose that Assumptions 1–4 are satisfied and* p *is fixed. For any* $u\in\mathcal{U}_{h_\star}$ *and* $1\leqslant i,j\leqslant N$*, if* $\sigma_{ij}^\star(u)=0$*, we must have* $\widetilde{\sigma}_{ij}(u)=0$ *with probability approaching one.*

4 Modified covariance matrix estimation and variable tuning parameter selection

In this section we first modify the semiparametric shrinkage estimation developed in Section 2.2 to ensure the positive definiteness of the estimated dynamic covariance matrix in finite samples, and then discuss the choice of the variable tuning parameter $\rho(u)$.

4.1 Modified dynamic covariance matrix estimation

In practical application, the estimated covariance matrix $\widetilde{\boldsymbol{\Sigma}}(u)$ constructed in Section 2.2 is not necessarily uniformly positive definite on $\mathcal{U}$. To fix this problem, we next introduce two simple modification techniques for our semiparametric shrinkage estimation method.

TYPE I MODIFICATION. Let $\widetilde{\lambda}_1(u)\geqslant\widetilde{\lambda}_2(u)\geqslant\cdots\geqslant\widetilde{\lambda}_N(u)$ be the eigenvalues of $\widetilde{\boldsymbol{\Sigma}}(u)$ arranged in a non-increasing order and $m_{1,T}>0$ be a tuning parameter which tends to zero as the sample size T goes to infinity. When the smallest eigenvalue $\widetilde{\lambda}_N(u)$ is negative, the matrix $\widetilde{\boldsymbol{\Sigma}}(u)$ is clearly not positive definite. In this case, one way to modify $\widetilde{\boldsymbol{\Sigma}}(u)$ is to follow Chen and Leng (2016) and

construct

$$\widetilde{\mathbf{\Sigma}}_C(u) = \widetilde{\mathbf{\Sigma}}(u) + \left[m_{1,T} - \widetilde{\lambda}_N(u)\right] \cdot \mathbf{I}_{N\times N}, \tag{4.1}$$

where $\mathbf{I}_{N\times N}$ is the $N \times N$ identity matrix. This modification guarantees that the smallest eigenvalue of $\widetilde{\mathbf{\Sigma}}_C(u)$ is uniformly larger than zero, indicating that $\widetilde{\mathbf{\Sigma}}_C(u)$ is uniformly positive definite. In addition, $\widetilde{\mathbf{\Sigma}}_C(u)$ can retain the approximate sparsity structure of $\mathbf{\Sigma}^*_A(u)$ as described in (3.4). Hence, we can use $\widetilde{\mathbf{\Sigma}}_C(u)$ as an alternative estimate of $\mathbf{\Sigma}^*_A(u)$ when $\widetilde{\lambda}_N(u)$ is negative. Consequently, define the following modified version of $\widetilde{\mathbf{\Sigma}}(u)$:

$$\begin{aligned}\widetilde{\mathbf{\Sigma}}_{1,M}(u) &= \widetilde{\mathbf{\Sigma}}(u) \cdot I\left(\widetilde{\lambda}_N(u) > 0\right) + \widetilde{\mathbf{\Sigma}}_C(u) \cdot I\left(\widetilde{\lambda}_N(u) \leqslant 0\right) \\ &= \widetilde{\mathbf{\Sigma}}(u) + \left[m_{1,T} - \widetilde{\lambda}_N(u)\right] \mathbf{I}_{N\times N} \cdot I\left(\widetilde{\lambda}_N(u) \leqslant 0\right),\end{aligned} \tag{4.2}$$

where $I(\cdot)$ is an indicator function. Note that when $\widetilde{\lambda}_N(u) \leqslant 0$, by Weyl's inequality and Theorem 1, we have

$$\left|\widetilde{\lambda}_N(u)\right| \leqslant \left|\widetilde{\lambda}_N(u) - \lambda_{\min}(\mathbf{\Sigma}^*_A(u))\right| \leqslant \sup_{u\in\mathcal{U}_{h_\star}} \left\|\widetilde{\mathbf{\Sigma}}(u) - \mathbf{\Sigma}^*_A(u)\right\|_O = O_P\left(c_N \cdot \tau_{T,N}^{1-q}\right), \tag{4.3}$$

where $\lambda_{\min}(\mathbf{\Sigma}^*_A(u))$ represents the minimum eigenvalue of $\mathbf{\Sigma}^*_A(u)$. Hence,

$$\sup_{u\in\mathcal{U}_{h_\star}} \left\|\widetilde{\mathbf{\Sigma}}_{1,M}(u) - \mathbf{\Sigma}^*_A(u)\right\|_O \leqslant \sup_{u\in\mathcal{U}_{h_\star}} \left\|\widetilde{\mathbf{\Sigma}}(u) - \mathbf{\Sigma}^*_A(u)\right\|_O + O_P\left(c_N \cdot \tau_{T,N}^{1-q} + m_{1,T}\right). \tag{4.4}$$

By choosing $m_{1,T} = O(c_d \tau_{n,d}^{1-q})$ and using Theorem 1, we obtain the same uniform convergence rate for $\widetilde{\mathbf{\Sigma}}_{1,M}(u)$ as that for $\widetilde{\mathbf{\Sigma}}(u)$ in Theorem 1. Glad, Hjort and Ushakov (2003) consider a similar modification for density estimators that are not bona fide densities; and they show that the modification improves the performance of the density estimators as measured by the integrated mean squared error.

TYPE II MODIFICATION. The modification technique defined in (4.1) and (4.2) modifies all the eigenvalues of $\widetilde{\mathbf{\Sigma}}(u)$. The modification made on eigenvalues, which are significantly larger than zero, however, may be unnecessary and, if done, would affect the finite-sample performance of the resulting covariance matrix estimate. Hence, we next introduce a different modification technique. Again let $\widetilde{\lambda}_1(u) \geqslant \widetilde{\lambda}_2(u) \geqslant \cdots \geqslant \widetilde{\lambda}_N(u)$ be the eigenvalues of $\widetilde{\mathbf{\Sigma}}(u)$, and let $\widetilde{\mathbf{\phi}}_k(u)$ be the normalised eigenvector corresponding to the eigenvalue $\widetilde{\lambda}_k(u)$. Let $m_{2,T} > 0$ be a tuning parameter which tends to zero as the sample size T goes to infinity, and define $N_1 := N_1(m_{2,T}) = \max\{1 \leqslant k \leqslant N :$

$\widetilde{\lambda}_k(u) > m_{2,T}\}$. By standard eigen-decomposition, we have

$$\widetilde{\mathbf{\Sigma}}(u) = \sum_{k=1}^{N} \widetilde{\lambda}_k(u)\widetilde{\boldsymbol{\phi}}_k(u)\widetilde{\boldsymbol{\phi}}_k^{\tau}(u) = \sum_{k=1}^{N_1} \widetilde{\lambda}_k(u)\widetilde{\boldsymbol{\phi}}_k(u)\widetilde{\boldsymbol{\phi}}_k^{\tau}(u) + \sum_{k=N_1+1}^{N} \widetilde{\lambda}_k(u)\widetilde{\boldsymbol{\phi}}_k(u)\widetilde{\boldsymbol{\phi}}_k^{\tau}(u). \tag{4.5}$$

A natural way to achieve positive definiteness is to replace the nonpositive eigenvalues, $\widetilde{\lambda}_k(u)$, $k = N_1 + 1, \cdots, N$, by appropriate positive values in the eigen-decomposition (4.5). Define

$$\widetilde{\lambda}_{k,M}(u) = m_{2,T} \cdot I\left(\widetilde{\lambda}_k(u) \leqslant m_{2,T}\right), \ k = N_1 + 1, \cdots, N. \tag{4.6}$$

Another modified version of $\widetilde{\mathbf{\Sigma}}(u)$ is defined as

$$\widetilde{\mathbf{\Sigma}}_{2,M}(u) = \sum_{k=1}^{N_1} \widetilde{\lambda}_k(u)\widetilde{\boldsymbol{\phi}}_k(u)\widetilde{\boldsymbol{\phi}}_k^{\tau}(u) + \sum_{k=N_1+1}^{N} \widetilde{\lambda}_{k,M}(u)\widetilde{\boldsymbol{\phi}}_k(u)\widetilde{\boldsymbol{\phi}}_k^{\tau}(u). \tag{4.7}$$

With Theorem 1 in Section 3, we can also derive the uniform consistency for this modified semi-parametric shrinkage estimator. Note that when $\widetilde{\lambda}_k(u) \leqslant 0$, the convergence result in (4.3) also holds with $\widetilde{\lambda}_N(u)$ replaced by $\widetilde{\lambda}_k(u)$. This indicates that

$$\sup_{u \in \mathcal{U}_{h_\star}} \left\| \widetilde{\mathbf{\Sigma}}(u) - \widetilde{\mathbf{\Sigma}}_{2,M}(u) \right\|_O = O_P\left(c_N \cdot \tau_{T,N}^{1-q} + m_{2,T}\right).$$

This, together with Theorem 1, leads to

$$\begin{aligned} \sup_{u \in \mathcal{U}_{h_\star}} \left\| \widetilde{\mathbf{\Sigma}}_{2,M}(u) - \mathbf{\Sigma}_A^\star(u) \right\|_O &\leqslant \sup_{u \in \mathcal{U}_{h_\star}} \left\| \widetilde{\mathbf{\Sigma}}(u) - \mathbf{\Sigma}_A^\star(u) \right\|_O + \sup_{u \in \mathcal{U}_{h_\star}} \left\| \widetilde{\mathbf{\Sigma}}(u) - \widetilde{\mathbf{\Sigma}}_{2,M}(u) \right\|_O \\ &= O_P\left(c_N \cdot \tau_{T,N}^{1-q} + m_{2,T}\right). \end{aligned} \tag{4.8}$$

By choosing $m_{2,T} = O\left(c_N \tau_{T,N}^{1-q}\right)$, we obtain the same uniform convergence rate for $\widetilde{\mathbf{\Sigma}}_{2,M}(u)$ as that for $\widetilde{\mathbf{\Sigma}}(u)$ in Theorem 1. With $\widetilde{\mathbf{\Sigma}}_{2,M}(u)$, only the non-positive eigenvalues of $\widetilde{\mathbf{\Sigma}}(u)$ are modified. However, a disadvantage of this modification is that it may affect the sparsity structure of the estimated covariance matrix.

4.2 Choice of the variable tuning parameter

For any shrinkage method for covariance matrix estimation, it is essential to choose an appropriate tuning parameter. Since the variables (X_t, U_t) are allowed to be serially correlated over time, the tuning parameter selection criteria proposed in Bickel and Levina (2008b) or Chen and Leng (2016)

for independent data may no longer work well in our setting. We hence modify their method as follows.

STEP 1: For given $u \in \mathcal{U}$, use a rolling window of size $\lfloor T/2 \rfloor + K$ and split data within each window into two sub-samples of sizes $T_1 = \left\lfloor \frac{T}{2}\left(1 - \frac{1}{\log(T/2)}\right) \right\rfloor$ and $T_2 = \lfloor T/2 \rfloor - T_1$ by leaving out K observations in-between them, where T is the overall sample size and $\lfloor \cdot \rfloor$ denotes the floor function.

STEP 2: Obtain $\widetilde{\mathbf{\Sigma}}_{\rho(u),1,k}(u)$ (the semiparametric shrinkage estimate of the dynamic covariance matrix) from the first sub-sample of the k-th rolling window, and $\widehat{\mathbf{\Sigma}}_{2,k}(u)$ (the naive estimate without the generalised shrinkage applied) from the second sub-sample of the k-th rolling window, $k = 1, \ldots, M$ with $M = \lfloor T/(2K) \rfloor$.

STEP 3: Choose the tuning parameter $\rho(u)$ so that it minimises the sum of squared Frobenius norm:

$$\sum_{k=1}^{M} \left\| \widetilde{\mathbf{\Sigma}}_{\rho(u),1,k}(u) - \widehat{\mathbf{\Sigma}}_{2,k}(u) \right\|_F^2. \tag{4.9}$$

The purpose for leaving out K observations in between the two subsamples in each rolling window is to make these two subsamples have negligible correlation. The number of observations left out, i.e. K, can depend on the strength of serial correlation in X_t. The stronger the serial correlation, the larger the value of K can be. In the simulation studies in Section 5, we leave out 10 observations for a medium serial correlation of 0.5. The results show that the above selection method has reasonably good numerical performance. It remains unclear whether such a choice of variable tuning parameter is asymptotically optimal. This question will be left in our future research.

5 Numerical studies

In this section, we first provide three simulated examples to examine the small-sample performance of the proposed semiparametric covariance matrix estimation methods and then apply the methods to construct global minimum variance portfolios from 500 stocks listed on New York Stock Exchange. Throughout this section, we use the value 0.01 for the two parameters $m_{1,T}$ and $m_{2,T}$ in the two modification methods detailed in Section 4.1. In the nonparametric estimation of $m_{i,k}(u_k)$ and $c_{ij,k}(u_k)$ in Section 2.2, we use the probability density function of the standard normal distribution as the kernel function and the rule-of-thumb bandwidths as the smoothing parameters.

5.1 Simulation studies

In this section, we conduct some Monte-Carlo experiments to examine the finite-sample performance of the proposed methods for estimating large dynamic covariance matrix. In order to provide a full performance study, we consider three different sparsity patterns of the underlying covariance matrix, i.e., the dynamic banded structure, the dynamic AR(1) structure, and the varying-sparsity structure. These are the multivariate conditioning variables extension of the covariance models considered in Examples 1-3 of Chen and Leng (2016). To measure estimation accuracy, we consider the Frobenius losses, i.e., $\|\mathbf{\Sigma}_0(u) - \check{\mathbf{\Sigma}}(u)\|_F$, for an estimate $\check{\mathbf{\Sigma}}(u)$ at a point u. We compare the accuracy of our semiparametric shrinkage estimation, defined in (2.14), with that of the generalised thresholding of the sample covariance matrix, which treats the covariance matrix as static. The estimator defined in (2.14) sometimes produces covariance matrices that are not positive definite, in which case the modification in either (4.2) or (4.7) can be applied. Hence, we also report the accuracy of the modified dynamic covariance matrix estimates. Four commonly used shrinkage methods – the hard thresholding, soft thresholding, adaptive LASSO (A. LASSO) and Smoothly Clipped Absolute Deviation (SCAD) – are considered in the simulation. Throughout this section, the dimension of X_t, N, takes one of the values of 100, 300, and 500, and the dimension of the conditioning vector U_t is set to be $p = 3$. The sample size is fixed at $T = 300$.

SIMULATED EXAMPLE 5.1. (Dynamic banded covariance matrix) The conditioning variables $U_t = (U_{t1}, U_{t2}, U_{t3})^\tau$ are drawn from a VAR(1) process:

$$U_t = 0.5U_{t-1} + \nu_t, \quad t = 1, \ldots, T, \tag{5.1}$$

with $U_0 = \mathbf{0}$, where ν_t are i.i.d. three-dimensional random vectors following the $N(\mathbf{0}, \mathbf{I}_{3\times3})$ distribution. For each $t = 1, \ldots, T$, the N-dimensional vector X_t is generated from the multivariate Gaussian distribution $N(\mathbf{0},\ \mathbf{\Sigma}_0(U_t))$, where

$$\mathbf{\Sigma}_0(U_t) = \left\{\sigma_{ij}^0(U_t)\right\}_{d\times d} \text{ with } \sigma_{ij}^0(U_t) = 0.25 \times \sigma_{ij}\Big(\sum_{k=1}^{3} U_{t,k}\Big) \tag{5.2}$$

and

$$\sigma_{ij}(\nu) = \exp(\nu/4)\{I(i=j) + [\phi(\nu) + 0.1]I(|i-j| = 1) + \phi(\nu)I(|i-j| = 2)\}$$

for any $\nu \in \mathcal{R}$, in which $\phi(\nu)$ is the probability density function of the standard normal distribution. Note that such a conditional covariance function $\sigma_{ij}^0(u)$ is non-additive in the elements of u. This enables us to evaluate the performance of the MAMAR method for approximating entries of a conditional covariance matrix that are non-additive and nonlinear.

The dynamic covariance matrix is estimated at the grid points: $[-0.5, 0, 0.5]^3$ for U (in total

Table 5.1: Average Frobenius losses (standard error) for Example 5.1

Method		N = 100	N = 300	N = 500
Static	Hard	1.5300(0.2647)	3.1024(0.1531)	4.2396(0.3117)
	Soft	1.7453(0.0294)	3.4345(0.0662)	4.4264(0.0891)
	A. LASSO	1.6316(0.0409)	3.2965(0.0364)	4.2716(0.0459)
	SCAD	1.7658(0.0320)	3.4050(0.0313)	4.4140(0.0385)
Dynamic	Hard	1.7501(0.0555)	3.2473(0.0687)	4.2552(0.0779)
	Soft	1.5731(0.0403)	3.0388(0.0953)	4.0583(0.0959)
	A. LASSO	1.6149(0.0451)	3.0412(0.1054)	4.0708(0.1016)
	SCAD	1.6143(0.0484)	3.0325(0.0855)	4.0425(0.1090)
Modified Dynamic 1	Hard	1.7534(0.0544)	3.2669(0.0626)	4.2604(0.0768)
	Soft	1.5739(0.0406)	3.0686(0.1193)	4.0564(0.0975)
	A. LASSO	1.6150(0.0451)	3.0593(0.1028)	4.0713(0.1031)
	SCAD	1.6151(0.0481)	3.0628(0.1352)	4.0457(0.1105)
Modified Dynamic 2	Hard	1.7498(0.0556)	3.2460(0.0701)	4.2550(0.0780)
	Soft	1.5728(0.0403)	3.0373(0.0964)	4.0581(0.0959)
	A. LASSO	1.6149(0.0451)	3.0404(0.1058)	4.0707(0.1015)
	SCAD	1.6142(0.0485)	3.0315(0.0857)	4.0424(0.1090)

there are 27 grid points). The average Frobenius losses over 30 replications and their standard errors (in parentheses) are summarised in Table 5.1. In this tables, "Static" refers to the estimation by treating the underlying covariance matrix as static, "Dynamic" refers to the estimation by our semiparametric shrinkage method detailed in Section 2, "Modified Dynamic 1" refers to the modified dynamic covariance matrix estimation defined in (4.2), and "Modified Dynamic 2" refers to the modified estimation defined in (4.7). In addition, "Hard", "Soft", "A. LASSO" and "SCAD" in the tables represent thresholding by the methods of hard thresholding, soft thresholding, adaptive LASSO and the smoothly clipped absolute deviation, respectively.

The results in Table 5.1 reveal that our semiparametric dynamic covariance matrix estimation via the proposed MAMAR approximation and its modified versions outperform the static covariance matrix estimation in all the three cases for N and in all thresholding methods except the hard thresholding. The two modified dynamic covariance matrix estimators have similar performance to their non-modified version. In addition, as the second modified estimator only modifies the non-positive eigenvalues, it produces a bit more accurate estimation than the first modified one.

SIMULATED EXAMPLE 5.2. (Dynamic non-sparse covariance matrix) The specifications of the data generating process are the same as those in Example 5.1, except that the dynamic covariance matrix $\mathbf{\Sigma}_0(u_t)$ is non-sparse. Specifically, the function $\sigma_{ij}(\cdot)$ in (5.2) is assumed to follow the covariance pattern of an AR(1) process:

$$\sigma_{ij}(v) = \exp(v/4)\,[\phi(v)]^{|i-j|}$$

Table 5.2: Average Frobenius losses (standard error) for Example 5.2

Method		N = 100	N = 300	N = 500
Static	Hard	1.3980(0.0313)	2.4738(0.0799)	3.1337(0.0574)
	Soft	1.3126(0.0173)	2.3007(0.0329)	2.9927(0.1022)
	A. LASSO	1.3096(0.0507)	2.3694(0.0666)	3.0549(0.0800)
	SCAD	1.3460(0.0402)	2.4039(0.0827)	3.0587(0.0859)
Dynamic	Hard	1.2970(0.0159)	2.2666(0.0207)	2.9149(0.0318)
	Soft	1.2610(0.0308)	2.2911(0.0340)	3.0054(0.0566)
	A. LASSO	1.2515(0.0240)	2.2262(0.0247)	2.8799(0.0367)
	SCAD	1.2604(0.0288)	2.2567(0.0381)	2.9203(0.0663)
Modified Dynamic 1	Hard	1.2977(0.0161)	2.2666(0.0208)	2.9165(0.0336)
	Soft	1.2612(0.0312)	2.2922(0.0431)	3.0040(0.0505)
	A. LASSO	1.2518(0.0243)	2.2278(0.0277)	2.8770(0.0321)
	SCAD	1.2627(0.0321)	2.2590(0.0469)	2.9245(0.0696)
Modified Dynamic 2	Hard	1.2969(0.0158)	2.2662(0.0203)	2.9146(0.0314)
	Soft	1.2608(0.0304)	2.2906(0.0337)	3.0052(0.0563)
	A. LASSO	1.2513(0.0237)	2.2259(0.0244)	2.8798(0.0366)
	SCAD	1.2601(0.0284)	2.2564(0.0377)	2.9201(0.0661)

for any $v \in \mathcal{R}$. The dynamic covariance matrix is again estimated at the grid points $[-0.5, 0, 0.5]^3$, and the average Frobenius losses are summarised in Table 5.2. The proposed semiparametric method and its modified versions outperform the sample covariance matrix estimation in almost all the thresholding methods considered. The outperformance is more notable when the cross-sectional dimension N is 300 or 500. This shows that our methods can estimate the non-sparse dynamic covariance matrices with satisfactory accuracy.

SIMULATED EXAMPLE 5.3. (Dynamic covariance matrix with varying sparsity) Data on U_t and X_t are generated in the same way as in Example 5.1 except that the dynamic covariance matrix $\mathbf{\Sigma}_0(U_t)$ has varying sparsity patterns. Specifically, $\mathbf{\Sigma}_0(U_t) = \left\{\sigma_{ij}^0(U_t)\right\}_{d\times d}$, where

$$\sigma_{ij}^0(U_t) = 0.4\sigma_{ij}(U_{t1}) + 0.3\sigma_{ij}(U_{t2}) + 0.3\sigma_{ij}(U_{t3})$$

and

$$\sigma_{ij}(v) = \exp(v/2)\left\{\mathrm{I}(i=j) + 0.5\exp\left[-\frac{(v-0.25)^2}{0.75^2-(v-0.25)^2}\right]\mathrm{I}(-0.49 \leqslant v \leqslant 0.99)\mathrm{I}(|i-j|=1)\right.$$
$$\left. +0.4\exp\left[-\frac{(v-0.65)^2}{0.35^2-(v-0.65)^2}\mathrm{I}(0.31 \leqslant v \leqslant 0.99)\mathrm{I}(|i-j|=2)\right]\right\}$$

for any $v \in \mathcal{R}$. The dynamic covariance matrix is again estimated at the grid points: $[-0.5, 0, 0.5]^3$, where the corresponding covariance matrix has different sparsity structures. For example, at point $(-0.5, -0.5, -0.5)$, the considered covariance matrix is diagonal; at point $(0, 0, 0)$, only entries on

Table 5.3: Average Frobenius losses (standard error) for Example 5.3

Method		N = 100	N = 300	N = 500
Static	Hard	5.9166(0.1800)	10.3160(0.2429)	13.2730(0.3309)
	Soft	5.2686(0.1919)	9.5356(0.1319)	12.3910(0.0886)
	A. LASSO	5.6566(0.1795)	9.9110(0.1905)	12.7320(0.2467)
	SCAD	5.8351(0.2540)	10.233(0.2380)	13.1670(0.3211)
Dynamic	Hard	5.5689(0.0592)	9.7132(0.1979)	12.5050(0.1156)
	Soft	5.3613(0.0985)	9.7916(0.1920)	12.7590(0.1960)
	A. LASSO	5.3986(0.0609)	9.5077(0.0942)	12.2770(0.0816)
	SCAD	5.4613(0.0729)	9.6936(0.1306)	12.5380(0.1332)
Modified Dynamic 1	Hard	5.5790(0.0986)	9.8320(0.1915)	12.5650(0.2671)
	Soft	5.3571(0.0877)	9.7923(0.2602)	12.7440(0.1911)
	A. LASSO	5.3983(0.0600)	9.5130(0.1098)	12.2820(0.0946)
	SCAD	5.4673(0.0933)	9.7292(0.2665)	12.5670(0.2061)
Modified Dynamic 2	Hard	5.5680(0.0571)	9.7086(0.1757)	12.5040(0.1141)
	Soft	5.3600(0.0947)	9.7887(0.1846)	12.7580(0.1946)
	A. LASSO	5.3984(0.0603)	9.5067(0.0915)	12.2760(0.0809)
	SCAD	5.4604(0.0709)	9.6917(0.1254)	12.5370(0.1325)

the main diagonal and those that are one row above or below the main diagonal are nonzero; at point $(0.5, 0.5, 0.5)$, entries on the main diagonal and those that are one or two rows above or below the main diagonal are nonzero. The average Frobenius losses are presented in Table 5.3, which shows that our semiparametric dynamic covariance matrix estimation and its modified versions outperform the static covariance matrix estimation in all the three cases for N and in all thresholding methods except the soft thresholding.

5.2 An empirical example

In this section, we consider an application to construction of global minimum variance portfolios. The data used are downloaded from Datastream and consist of daily returns (in %) on stocks listed on New York Stock Exchange over the period 11/12/2008 – 10/04/2018. We select 500 stocks which have largest capitalisation (measured at the end of the period) among those stocks that have complete trading history over this period. Also used, as the conditioning variables, are the one-day-before returns (i.e., over the period 10/12/2009 – 09/04/2018) on the Fama-French three factors, which are downloaded from Keneth French's data library website `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. The out-of-sample period is 11/12/2009 –10/04/2018.

Let X_{t+1} be the vector of returns for day $t+1$, U_t be the vector of returns on Fama-French three factors for day t, and $\mathbf{\Sigma}_t(u) = \mathsf{Cov}(X_{t+1}|U_t = u)$ be the covariance matrix of stock returns for day

Table 5.4: Out-of-sample performance of covariance matrix estimation

Method	Equal-weighted	Static	Dynamic	Modified Dynamic 1	Modified Dynamic 2
AVR(%)	11.1636	-1235.2540	-304.8696	6.6024	9.5256
STD(%)	16.5111	3965.4730	7678.0740	55.5814	11.5106
IR	0.6763	-0.3111	-0.0397	0.1191	0.8286

$t+1$ conditioning on the information on the three factors on day t. The global minimum variance portfolio for day $t+1$ given $U_t = u$ is obtained by solving

$$\arg\min_{w} w^\intercal \mathbf{\Sigma}_t(u) w \quad \text{subject to} \quad w^\intercal \mathbf{1} = 1,$$

where $\mathbf{1}$ is an N-dimensional column vector with each element being one, and $N = 500$. The analytical solution to the above problem can be written as

$$w_t^\star = \frac{\mathbf{\Sigma}_t^{-1}(u)\mathbf{1}}{\mathbf{1}^\intercal \mathbf{\Sigma}_t^{-1}(u)\mathbf{1}}.$$

Note that we allow short selling, i.e., elements of $w_t^\star$ can be negative, and we assume there is no transaction cost. The minimum variance portfolio is updated daily, i.e., for each trading day in the out-of-sample period, the 500×500 covariance matrix $\mathbf{\Sigma}_t(u)$ is estimated at the end of the previous trading day using data from most recent 250 trading days (which is roughly one year's data) and the minimum variance portfolio is constructed and held until the end of the current trading day.

After obtaining all the out-of-sample global minimum variance portfolio returns, we compute their annualised average return (AVR), annualised standard deviation (STD) and the annualised information ratio (IR), which is defined as the ratio of AVR to STD. The AVR, STD and IR will be used as measures of performance of portfolios constructed using the following covariance matrix estimates: Static (sample covariance matrix with SCAD thresholding), Dynamic (the proposed semiparametric method with SCAD thresholding), Modified Dynamic 1 (the Dynamic estimation with type I modification), and Modified Dynamic 2 (the Dynamic estimation with type II modification). As a benchmark, we also consider the equally-weighted portfolio (each stock has a weight of $1/500$). The results are summarised in Table 5.4.

As discussed in Engle, Ledoit and Wolf (2017), in the context of constructing global minimum variance portfolios, the primary measure that should be used to evaluate the performance of covariance matrix estimation methods is the STD. High AVR and IR are desirable but should be of secondary importance. In this respect, the dynamic covariance matrix estimation with type II modification outperforms the rest of the methods considered. We also note that the returns

on minimum variance portfolios constructed using covariance matrix estimates from the Static and Dynamic methods have large (in magnitude) average returns and standard deviations. This is likely to be caused by some covariance matrix estimates from these two methods being non-positive definite (as they are not corrected for non-positive definiteness as the last two methods in Table 5.4). Another possible cause of the large standard deviations of the returns from the Static and Dynamic methods is that the computation of the inverse of large (500×500) covariance matrices is unstable, which subsequently might also affected the performance of the two Modified Dynamic approaches. With the equally weighted portfolio, there is not such a computational issue and its performance is more stable.

6 Extensions

In this section, we consider extending our methodology to deal with the following two scenarios: (i) the number of conditioning variables is divergent as the sample size increases, and (ii) the large covariance matrix is conditionally sparse with X_t driven by some latent common factors and satisfying the approximate factor model structure.

6.1 Extension 1: the dimension of U_t is large

In the previous sections, we limit attention to the case where the number of conditioning variables is a fixed positive integer. However, it is often not uncommon to have a very large number of candidate conditioning variables in practice. In this latter case, a direct application of the MAMAR approximation and the semiparametric method proposed in Section 2.2 may result in poor and unstable matrix estimation results. Motivated by a recent paper by Chen *et al* (2018) on high-dimensional MAMAR method, we can circumvent this problem by assuming that the number of conditioning variables which make "significant" contribution to estimating joint regression functions, $m_i^0(u)$ and $c_{ij}^0(u)$, in (1.4) and (1.5) is relatively small, i.e., for each i and j, when p is divergent, the number of nonzero weights $b_{i,k}$ and $a_{ij,k}$, $1 \leqslant k \leqslant p$, is relatively small. This makes equations (1.4) and (1.5) fall into the classic sparsity framework commonly used in high-dimensional variable or feature selection literature. To remove the insignificant conditioning variables, we combine the penalisation and MAMAR techniques when estimating $m_i^0(u)$ and $c_{ij}^0(u)$. Specifically, for each $1 \leqslant i \leqslant N$, to estimate $\left(b_{i,1}^\star, \ldots, b_{i,p}^\star\right)^\intercal$, we define the penalised objective function:

$$Q_i(b_{i,1}, \ldots, b_{i,p}) = \sum_{t=1}^{T-1} \left[X_{t+1,i}^c - \sum_{k=1}^{p} b_{i,k} \widehat{m}_{i,k}^c(U_{tk}) \right]^2 + T \sum_{k=1}^{p} p_{\lambda_1}(|b_{i,k}|), \tag{6.1}$$

where $X^c_{t+1,i} = X_{t+1,i} - \frac{1}{T-1}\sum_{s=1}^{T-1} X_{s+1,i}$, $\widehat{m}^c_{i,k}(U_{tk}) = \widehat{m}_{i,k}(U_{tk}) - \frac{1}{T-1}\sum_{s=1}^{T-1}\widehat{m}_{i,k}(U_{sk})$, and $p_{\lambda_1}(\cdot)$ is a penalty function with a tuning parameter λ_1. The solution to the minimisation of $Q_i(b_{i,1},\ldots,b_{i,p})$ is the penalised estimator of the optimal weights and is denoted by $(\overline{b}_{i,1},\ldots,\overline{b}_{i,p})^\intercal$. The subsequent intercept estimate, denoted by $\overline{b}_{i,0}$, can be calculated similarly to $\widehat{a}_{ij,0}$ in (2.11). For each $1 \leqslant i,j \leqslant N$, to estimate $\left(a^\star_{ij,1},\ldots,a^\star_{ij,p}\right)^\intercal$, we define the penalised objective function:

$$Q_{ij}(a_{ij,1},\ldots,a_{ij,p}) = \sum_{t=1}^{T-1}\left[X^c_{t+1,(i,j)} - \sum_{k=1}^{p} a_{ij,k}\widehat{c}^c_{ij,k}(U_{tk})\right]^2 + T\sum_{k=1}^{p} p_{\lambda_2}(|a_{ij,k}|), \tag{6.2}$$

where $X^c_{t+1,(i,j)} = X_{t+1,i}X_{t+1,j} - \frac{1}{T-1}\sum_{s=1}^{T-1} X_{s+1,i}X_{s+1,j}$, $\widehat{c}^c_{ij,k}(U_{tk}) = \widehat{c}_{ij,k}(U_{tk}) - \frac{1}{T-1}\sum_{s=1}^{T-1}\widehat{c}_{ij,k}(U_{sk})$, and $p_{\lambda_2}(\cdot)$ is a penalty function with a tuning parameter λ_2. The solution to the minimisation of $Q_{ij}(a_{ij,1},\ldots,a_{ij,p})$ is denoted by $(\overline{a}_{ij,1},\ldots,\overline{a}_{ij,p})^\intercal$, and the intercept estimate, $\overline{a}_{ij,0}$, can be obtained accordingly by replacing $\widehat{a}_{ij,k}$ with $\overline{a}_{ij,k}$, $k = 1,\ldots,p$, on the right hand side of the equation for $\widehat{a}_{ij,0}$ in (2.11). By Theorem 2(ii) in Chen *et al* (2018), under the sparsity assumption and some technical conditions, the zero optimal weights can be estimated exactly as zeros with probability approaching one. After obtaining $\overline{b}_{i,k}$ and $\overline{a}_{ij,k}$, $0 \leqslant k \leqslant p$, we can calculate the penalised estimates of the optimal MAMAR approximation to $c^0_{ij}(u)$ and $m^0_i(u)$ as

$$\overline{c}_{ij}(u) = \overline{a}_{ij,0} + \sum_{k=1}^{p}\overline{a}_{ij,k}\widehat{c}_{ij,k}(u_k), \quad \overline{m}_i(u) = \overline{b}_{i,0} + \sum_{k=1}^{p}\overline{b}_{i,k}\widehat{m}_{i,k}(u_k),$$

and subsequently the penalised estimate of $\sigma^0_{ij}(u)$ as

$$\overline{\sigma}_{ij}(u) = \overline{c}_{ij}(u) - \overline{m}_i(u)\overline{m}_j(u). \tag{6.3}$$

Finally, we apply the shrinkage technique detailed in Section 2.2 to $\overline{\sigma}_{ij}(u)$ to obtain the dynamic covariance matrix estimators. Their asymptotic property and numerical performance will be explored in a separate project.

6.2 Extension 2: conditionally sparse covariance matrix estimation

In many practical applications, the approximate sparsity assumption on the dynamic covariance matrix may be too restrictive. For example, if there is very little ex ante predictability in stock returns, a zero in the conditional covariance matrix would imply a zero in the same location of the unconditional covariance matrix. To address this issue, we allow for common contemporaneous market factors that drive the covariance matrix of stock returns, and then impose sparseness on the residual covariance matrix. To address this problem fully, we need to substantially generalise

the methods developed in Section 2.2. Suppose that the high-dimensional X_t is generated via the approximate factor model

$$X_t = \Lambda F_t + \epsilon_t, \quad t = 1, \ldots, T, \tag{6.4}$$

where Λ is an $N \times r$ matrix of factor loadings, F_t is an r-dimensional vector of latent factors, and ϵ_t is an N-dimensional vector of idiosyncratic errors. For simplicity, we assume that r, the number of factors, is pre-specified in the following discussion. In practice, this number is usually unknown but can be determined by an information criterion (Bai and Ng, 2002) or a simple ratio method (Lam and Yao, 2012). As in the literature (c.f., Bai and Ng, 2002; Stock and Watson, 2002), we only consider the setting of large factor models with both N and T divergent to infinity, making it feasible to consistently estimate the latent factors and factor loadings (up to an appropriate rotation).

Throughout this section, we assume that the factor loading matrix Λ is non-random, and F_t and ϵ_t are orthogonal given the past conditioning variables $\mathcal{U}_s$, $s \leqslant t-1$. Without loss of generality, the conditional means of F_t and ϵ_t are assumed to be zeros. Therefore, the dynamic covariance matrix of X_t can be written as

$$\begin{aligned} \mathcal{C}_0(u) &= \mathsf{E}\left(X_{t+1}X_{t+1}^\intercal | \mathcal{U}_t = u\right) \\ &= \Lambda \mathsf{E}\left(F_{t+1}F_{t+1}^\intercal | \mathcal{U}_t = u\right)\Lambda^\intercal + \mathsf{E}\left(\epsilon_{t+1}\epsilon_{t+1}^\intercal | \mathcal{U}_t = u\right) \\ &=: \Lambda \mathcal{C}_F(u)\Lambda^\intercal + \mathcal{C}_\epsilon(u). \end{aligned} \tag{6.5}$$

It is easy to find that the dynamic covariance matrix for the factors F_t, $\mathcal{C}_F(u)$, is of size $r \times r$, whereas that for the errors ϵ_t, $\mathcal{C}_\epsilon(u)$, is of much larger size $N \times N$. For each entry of $\mathcal{C}_F(u)$ and $\mathcal{C}_\epsilon(u)$, we may approximate it via the MAMAR method with the optimal weights chosen as in Section 2.1, and then obtain the corresponding proxies: $\mathcal{C}^\star_{F,A}(u)$ and $\mathcal{C}^\star_{\epsilon,A}(u)$, respectively.

As neither F_t nor ϵ_t are observable, we need to estimate them first before using the two-stage semiparametric method introduced in Section 2.2. Let $\mathcal{X}_T = (X_1, \ldots, X_T)^\intercal$ be a $T \times N$ matrix of observations. We next use the principal component method to estimate the latent components in the factor model (6.4). With an eigenanalysis on the $T \times T$ matrix $\left(\mathcal{X}_T\mathcal{X}_T^\intercal\right)/(NT)$, we obtain a $T \times r$ matrix

$$\widehat{\mathcal{F}}_T := \left(\widehat{F}_1, \ldots, \widehat{F}_T\right)^\intercal$$

consisting of the r eigenvectors (multiplied by $\sqrt{T}$) corresponding to the r largest eigenvalues, and estimate the factor loading matrix Λ by $\widehat{\Lambda}_T$ which is defined by

$$\widehat{\Lambda}_T = \mathcal{X}_T^\intercal \widehat{\mathcal{F}}_T / T.$$

Note that $\widehat{F}_t$ and $\widehat{\Lambda}_T$ constructed above are the consistent estimators of the rotated latent factors HF_t and factor loading matrix ΛH^{-1}, respectively, rather than the factors and factor loading matrix themselves, where H is an $r \times r$ rotation matrix. The idiosyncratic errors ϵ_t can be approximated by $\widehat{\epsilon}_t$, which is defined by

$$\widehat{\epsilon}_t = X_t - \widehat{\Lambda}_T \widehat{F}_t, \quad t = 1, \ldots, T.$$

With $\widehat{F}_t$ and $\widehat{\epsilon}_t$ obtained above, we use the two-stage semiparametric estimation procedure in Section 2.2 to construct

$$\widehat{\mathcal{C}}_{\hat{F}}(u) = \left[\widehat{c}_{ij}^{\hat{F}}(u)\right]_{r\times r} \quad \text{and} \quad \widehat{\mathcal{C}}_{\hat{\epsilon}}(u) = \left[\widehat{c}_{ij}^{\hat{\epsilon}}(u)\right]_{N\times N},$$

where $\widehat{c}_{ij}^{\hat{F}}(u)$ and $\widehat{c}_{ij}^{\hat{\epsilon}}(u)$ are defined similarly to $\widehat{c}_{ij}(u)$ in (2.12). As the size of the estimated matrix $\widehat{\mathcal{C}}_{\hat{\epsilon}}(u)$ is large, we further apply the generalised shrinkage technique and consequently obtain

$$\widetilde{\mathcal{C}}_{\hat{\epsilon}}(u) = \left[\widetilde{c}_{ij}^{\hat{\epsilon}}(u)\right]_{N\times N} \quad \text{with} \quad \widetilde{c}_{ij}^{\hat{\epsilon}}(u) = s_{\rho(u)}\left(\widehat{c}_{ij}^{\hat{\epsilon}}(u)\right).$$

Then we use

$$\widetilde{\mathcal{C}}(u) := \widehat{\Lambda}_T \widehat{\mathcal{C}}_{\hat{F}}(u) \widehat{\Lambda}_T^{\tau} + \widetilde{\mathcal{C}}_{\hat{\epsilon}}(u) \tag{6.6}$$

to estimate

$$\mathcal{C}_A^{\star}(u) := \Lambda \mathcal{C}_{F,A}^{\star}(u) \Lambda^{\tau} + \mathcal{C}_{\epsilon,A}^{\star}(u),$$

the MAMAR approximation of $\mathcal{C}_0(u)$.

We finally sketch the asymptotic properties of the above estimators before concluding this section. Let $\widehat{\mathcal{C}}_F(u)$ and $\widetilde{\mathcal{C}}_\epsilon(u)$ be the infeasible covariance matrix estimates using the latent factors F_t and unobserved errors ϵ_t, respectively. Following the proofs of Proposition 1 and Theorem 1 in Appendix A, we can prove that $\widehat{\mathcal{C}}_F(u)$ and $\widetilde{\mathcal{C}}_\epsilon(u)$ converge to $\mathcal{C}_{F,A}^{\star}(u)$ and $\mathcal{C}_{\epsilon,A}^{\star}(u)$ uniformly over u measured by the matrix operator norm as in (3.5). Meanwhile, following the proofs of Propositions 3.1 and 3.2 in Li, Li and Fryzlewicz (2018), we may show that replacing HF_t and ϵ_t by $\widehat{F}_t$ and $\widehat{\epsilon}_t$ in the covariance matrix estimation has negligible effect in the asymptotic analysis. Combining the above arguments, we readily have that $\widehat{\mathcal{C}}_{\hat{F}}(u)$ and $\widetilde{\mathcal{C}}_{\hat{\epsilon}}(u)$ uniformly converge to $H\mathcal{C}_{F,A}^{\star}(u)H^{\tau}$ and $\mathcal{C}_{\epsilon,A}^{\star}(u)$, respectively. However, as the proxy covariance matrix $\mathcal{C}_A(u)$ has spiked eigenvalues, Theorem 1 cannot be directly extended to $\widetilde{\mathcal{C}}(u)$ defined in (6.6) (c.f., Fan, Liao and Mincheva, 2013).

7 Conclusion

In this paper we estimate the large dynamic covariance matrices for high-dimensional time series data where the conditioning random variables are multivariate. Through the semiparametric MAMAR approximation to each entry in the underlying dynamic covariance matrix, we successfully circumvent the curse of dimensionality problem in multivariate nonparametric estimation. The subsequent two-stage semiparametric estimation method, combined with the generalised shrinkage technique commonly used in high-dimensional data analysis, produces the dynamic covariance matrix estimation varying with the conditioning variables over time. Under some mild conditions such as the approximate sparsity assumption, the developed covariance matrix estimation is proved to be uniformly consistent with convergence rates comparable to those obtained in the literature. In addition, two easy-to-implement techniques are introduced to modify the semiparametric dynamic covariance matrix estimation to ensure that the estimated covariance matrix is positive definite. A new selection criterion to determine the optimal local tuning parameter is provided to implement the proposed semiparametric large covariance matrix estimation for high-dimensional weakly dependent time series data. Furthermore, two interesting extensions of our methodology are explored to deal with the two scenarios which are of empirical relevance. Simulation studies conducted in Section 5 show that the proposed approaches have reliable finite-sample performance.

A Proofs of the main limit theorems

In this appendix, we provide the detailed proofs of the main asymptotic theorems. We start with some technical lemmas whose proofs will be given in Appendix B.

LEMMA 1. *Suppose that Assumptions 1, 2(i) and 3 in Section 3.1 are satisfied. Then we have*

$$\max_{1\leqslant i\leqslant N}\max_{1\leqslant k\leqslant p}\sup_{a_k+h_\star\leqslant u_k\leqslant b_k-h_\star}|\widehat{m}_{i,k}(u_k)-m_{i,k}(u_k)|=O_P\left(\sqrt{\log(N\vee T)/(Th_1)}+h_1^2\right), \tag{A.1}$$

and

$$\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\sup_{a_k+h_\star\leqslant u_k\leqslant b_k-h_\star}|\widehat{c}_{ij,k}(u_k)-c_{ij,k}(u_k)|=O_P\left(\sqrt{\log(N\vee T)/(Th_2)}+h_2^2\right), \tag{A.2}$$

where $h_\star=h_1\vee h_2$.

LEMMA 2. *Suppose that Assumptions 1–3 in Section 3.1 are satisfied. Then we have*

$$\max_{1\leqslant i,j\leqslant N}\sum_{k=0}^{p}\left|\widehat{a}_{ij,k}-a^{\star}_{ij,k}\right| = O_P\left(\sqrt{\log(N\vee T)/T}+\sqrt{\log(N\vee T)/(Th_2)}+h_2^2\right), \tag{A.3}$$

and

$$\max_{1\leqslant i\leqslant N}\sum_{k=0}^{p}\left|\widehat{b}_{i,k}-b^{\star}_{i,k}\right| = O_P\left(\sqrt{\log(N\vee T)/T}+\sqrt{\log(N\vee T)/(Th_1)}+h_1^2\right). \tag{A.4}$$

The following proposition gives an uniform consistency (with convergence rates) for the nonparametric conditional covariance function estimation via the MAMAR approximation.

PROPOSITION 1. *Suppose that Assumptions 1–3 in Section 3.1 are satisfied. Then we have*

$$\max_{1\leqslant i,j\leqslant N}\sup_{u\in\mathcal{U}}\left|\widehat{\sigma}_{ij}(u)-\sigma^{\star}_{ij}(u)\right| = O_P(\tau_{T,N}), \tag{A.5}$$

where $\tau_{T,N}$ was defined in Assumption 4, and $\sigma^{\star}_{ij}(u)=c^{\star}_{ij}(u)-m^{\star}_i(u)m^{\star}_j(u)$, $c^{\star}_{ij}(u)$ is the (i,j)-entry of $\mathcal{C}^{\star}_A(u)$ and $m^{\star}_i(u)$ is the i-th element of $\mathcal{M}^{\star}_A(u)$, $\mathcal{C}^{\star}_A(u)$ and $\mathcal{M}^{\star}_A(u)$ were defined in Section 2.1.

PROOF OF PROPOSITION 1. By (A.2) and (A.3), we have

$$\begin{aligned}
\widehat{c}_{ij}(u)-c^{\star}_{ij}(u) &= \left[\widehat{a}_{ij,0}+\sum_{k=1}^{p}\widehat{a}_{ij,k}\widehat{c}_{ij,k}(u_k)\right]-\left[a^{\star}_{ij,0}+\sum_{k=1}^{p}a^{\star}_{ij,k}c_{ij,k}(u_k)\right]\\
&= \left(\widehat{a}_{ij,0}-a^{\star}_{ij,0}\right)+\sum_{k=1}^{p}\left(\widehat{a}_{ij,k}-a^{\star}_{ij,k}\right)c_{ij,k}(u_k)+\sum_{k=1}^{p}a^{\star}_{ij,k}\left[\widehat{c}_{ij,k}(u_k)-c_{ij,k}(u_k)\right]+\\
&\quad\sum_{k=1}^{p}\left(\widehat{a}_{ij,k}-a^{\star}_{ij,k}\right)\left[\widehat{c}_{ij,k}(u_k)-c_{ij,k}(u_k)\right]\\
&= O_P\left(\sqrt{\log(N\vee T)/(Th_2)}+h_2^2\right)
\end{aligned} \tag{A.6}$$

uniformly for $1\leqslant i,j\leqslant N$ and $u\in\mathcal{U}_{h_\star}$. On the other hand, note that

$$\begin{aligned}
\widehat{m}_i(u)\widehat{m}_j(u)-m^{\star}_i(u)m^{\star}_j(u) = {}& \left[\widehat{m}_i(u)-m^{\star}_i(u)\right]m^{\star}_j(u)+m^{\star}_i(u)\left[\widehat{m}_j(u)-m^{\star}_j(u)\right]+\\
&\left[\widehat{m}_i(u)-m^{\star}_i(u)\right]\left[\widehat{m}_j(u)-m^{\star}_j(u)\right]
\end{aligned} \tag{A.7}$$

with

$$\begin{aligned}\widehat{m}_i(u) - m_i^\star(u) = &\left(\widehat{b}_{i,0} - b_{i,0}^\star\right) + \sum_{k=1}^{p}\left(\widehat{b}_{i,k} - b_{ij,k}^\star\right) m_{i,k}(u_k) + \sum_{k=1}^{p} b_{i,k}^\star \left[\widehat{m}_{i,k}(u_k) - m_{i,k}(u_k)\right] + \\ &\sum_{k=1}^{p}\left(\widehat{b}_{i,k} - b_{i,k}^\star\right)\left[\widehat{m}_{i,k}(u_k) - m_{i,k}(u_k)\right] \\ = &\, O_P\left(\sqrt{\log(N \vee T)/(Th_1)} + h_1^2\right) \end{aligned} \tag{A.8}$$

uniformly for $1 \leqslant i \leqslant N$ and $u \in \mathcal{U}_{h_\star}$, where (A.1) and (A.4) have been used.

Therefore, by (A.6)–(A.8), we have

$$\begin{aligned}&\max_{1\leqslant i,j\leqslant N} \sup_{u\in\mathcal{U}_{h_\star}} \left|\widehat{\sigma}_{ij}(u) - \sigma_{ij}^\star(u)\right| \\ = &\max_{1\leqslant i,j\leqslant N} \sup_{u\in\mathcal{U}_{h_\star}} \left|\widehat{c}_{ij}(u) - c_{ij}^\star(u)\right| + \max_{1\leqslant i,j\leqslant N} \sup_{u\in\mathcal{U}_{h_\star}} \left|\widehat{m}_i(u)\widehat{m}_j(u) - m_i^\star(u)m_j^\star(u)\right| \\ = &\, O_P\left(\sqrt{\log(N \vee T)/(Th_1)} + \sqrt{\log(N \vee T)/(Th_2)} + h_1^2 + h_2^2\right), \end{aligned} \tag{A.9}$$

completing the proof of Proposition 1. □

PROOF OF THEOREM 1. From the definition of $\widetilde{\mathbf{\Sigma}}(u)$ and $\widetilde{\sigma}_{ij}(u)$, we have

$$\begin{aligned}
\sup_{u\in\mathcal{U}_{h_\star}} \left\|\widetilde{\mathbf{\Sigma}}(u) - \mathbf{\Sigma}_A^\star(u)\right\|_O &\leqslant \sup_{u\in\mathcal{U}} \max_{1\leqslant i\leqslant N} \sum_{j=1}^N \left|\widetilde{\sigma}_{ij}(u) - \sigma_{ij}^\star(u)\right| \\
&= \sup_{u\in\mathcal{U}_{h_\star}} \max_{1\leqslant i\leqslant N} \sum_{j=1}^N \left|s_{\rho(u)}\left(\widehat{\sigma}_{ij}(u)\right) I\left(|\widehat{\sigma}_{ij}(u)| > \rho(u)\right) - \sigma_{ij}^\star(u)\right| \\
&= \sup_{u\in\mathcal{U}_{h_\star}} \max_{1\leqslant i\leqslant N} \sum_{j=1}^N \Big|s_{\rho(u)}\left(\widehat{\sigma}_{ij}(u)\right) I\left(|\widehat{\sigma}_{ij}(u)| > \rho(u)\right) - \\
&\quad \sigma_{ij}^\star(u) I\left(|\widehat{\sigma}_{ij}(u)| > \rho(u)\right) - \sigma_{ij}^\star(u) I\left(|\widehat{\sigma}_{ij}(u)| \leqslant \rho(u)\right)\Big| \\
&\leqslant \sup_{u\in\mathcal{U}_{h_\star}} \max_{1\leqslant i\leqslant N} \sum_{j=1}^N \left|s_{\rho(u)}\left(\widehat{\sigma}_{ij}(u)\right) - \widehat{\sigma}_{ij}(u)\right| I\left(|\widehat{\sigma}_{ij}(u)| > \rho(u)\right) + \\
&\quad \sup_{u\in\mathcal{U}_{h_\star}} \max_{1\leqslant i\leqslant N} \sum_{j=1}^N \left|\widehat{\sigma}_{ij}(u) - \sigma_{ij}^\star(u)\right| I\left(|\widehat{\sigma}_{ij}(u)| > \rho(u)\right) + \\
&\quad \sup_{u\in\mathcal{U}_{h_\star}} \max_{1\leqslant i\leqslant N} \sum_{j=1}^N \left|\sigma_{ij}^\star(u)\right| I\left(|\widehat{\sigma}_{ij}(u)| \leqslant \rho(u)\right) \\
&=: I_1 + I_2 + I_3. \qquad (A.10)
\end{aligned}$$

From Proposition 1, we define an event

$$\mathcal{E} = \left\{ \max_{1\leqslant i,j\leqslant N} \sup_{u\in\mathcal{U}_{h_\star}} \left|\widehat{\sigma}_{ij}(u) - \sigma_{ij}^\star(u)\right| \leqslant M_1\tau_{T,N} \right\},$$

where M_1 is a positive constant such that $\mathsf{P}(\mathcal{E}) \geqslant 1-\epsilon$ with $\epsilon > 0$ being arbitrarily small. By property (iii) of the shrinkage function and Proposition 1, we readily have

$$I_1 \leqslant \sup_{u\in\mathcal{U}_{h_\star}} \rho(u) \left[\max_{1\leqslant i\leqslant N} \sum_{j=1}^N I\left(|\widehat{\sigma}_{ij}(u)| > \rho(u)\right) \right] \qquad (A.11)$$

and

$$I_2 \leqslant M_1\tau_{T,N} \sup_{u\in\mathcal{U}_{h_\star}} \max_{1\leqslant i\leqslant N} \sum_{j=1}^N I\left(|\widehat{\sigma}_{ij}(u)| > \rho(u)\right) \qquad (A.12)$$

conditional on the event $\mathcal{E}$. Note that on $\mathcal{E}$,

$$|\widehat{\sigma}_{ij}(u)| \leqslant |\sigma_{ij}^\star(u)| + |\widehat{\sigma}_{ij}(u) - \sigma_{ij}^\star(u)| \leqslant |\sigma_{ij}^\star(u)| + M_1\tau_{T,N}.$$

Recall that $\rho(u) = M_0(u)\tau_{T,N}$ in Assumption 4 and choose $M_0(u)$ such that $\inf_{u\in\mathcal{U}} M_0(u) = 2M_1$. Then, it is easy to see the event $\{|\widehat{\sigma}_{ij}(u)| > \rho(u)\}$ indicates that $\{|\sigma^\star_{ij}(u)| > M_1\tau_{T,N}\}$ holds. As $\mathbf{\Sigma}^\star_A(\cdot) \in \mathcal{S}(q, c_N, M_\star, \mathcal{U})$ defined in (3.4), we may show that

$$\begin{aligned} I_1 + I_2 &\leqslant \tau_{T,N}\left[\sup_{u\in\mathcal{U}} M_0(u) + M_1\right]\left[\sup_{u\in\mathcal{U}_{h_\star}} \max_{1\leqslant i\leqslant N} \sum_{j=1}^N I\left(|\widehat{\sigma}_{ij}(u)| > M_1\tau_{T,N}\right)\right] \\ &\leqslant \tau_{T,N}\left[\sup_{u\in\mathcal{U}} M_0(u) + M_1\right]\left[\sup_{u\in\mathcal{U}} \max_{1\leqslant i\leqslant N} \sum_{j=1}^N \frac{|\sigma^\star_{ij}(u)|^q}{M_1^q\tau_{T,N}^q}\right] \\ &= O\left(c_N \cdot \tau_{T,N}^{1-q}\right) \end{aligned} \tag{A.13}$$

on the event $\mathcal{E}$.

On the other hand, by the triangle inequality, we have for any $u \in \mathcal{U}_{h_\star}$,

$$|\widehat{\sigma}_{ij}(u)| \geqslant |\sigma^\star_{ij}(u)| - |\widehat{\sigma}_{ij}(u) - \sigma^\star_{ij}(u)| \geqslant |\sigma^\star_{ij}(u)| - M_1\tau_{T,N}$$

on the event $\mathcal{E}$. Hence, we readily show that $\{|\widehat{\sigma}_{ij}(u)| \leqslant \rho(u)\}$ indicates

$$\left\{|\sigma^\star_{ij}(u)| \leqslant \left(\sup_{u\in\mathcal{U}} M_0(u) + M_1\right)\tau_{T,N}\right\}.$$

Then, for I_3, by Assumption 4 and the definition of $\mathcal{S}(q, c_N, M_\star, \mathcal{U})$, we have

$$\begin{aligned} I_3 &\leqslant \sup_{u\in\mathcal{U}_{h_\star}} \max_{1\leqslant i\leqslant N} \sum_{j=1}^N \left|\sigma^\star_{ij}(u)\right| I\left(|\sigma^\star_{ij}(u)| \leqslant (\sup_{u\in\mathcal{U}} M_0(u) + M_1)\tau_{T,N}\right) \\ &\leqslant (\sup_{u\in\mathcal{U}} M_0(u) + M_1)^{1-q}\tau_{T,N}^{1-q} \sup_{u\in\mathcal{U}} \max_{1\leqslant i\leqslant N} \sum_{j=1}^N \left|\sigma^\star_{ij}(u)\right|^q \\ &= O_P\left(c_N \cdot \tau_{T,N}^{1-q}\right). \end{aligned} \tag{A.14}$$

The proof of (3.5) in Theorem 1(i) can be completed by (A.10), (A.13) and (A.14).

Note that

$$\begin{aligned} \sup_{u\in\mathcal{U}_{h_\star}} \left\|\widetilde{\mathbf{\Sigma}}^{-1}(u) - \mathbf{\Sigma}_A^{\star-1}(u)\right\|_O &= \sup_{u\in\mathcal{U}_{h_\star}} \left\|\widetilde{\mathbf{\Sigma}}^{-1}(u)\mathbf{\Sigma}^\star_A(u)\mathbf{\Sigma}_A^{\star-1}(u) - \widetilde{\mathbf{\Sigma}}^{-1}(u)\widetilde{\mathbf{\Sigma}}(u)\mathbf{\Sigma}_A^{\star-1}(u)\right\|_O \\ &\leqslant \sup_{u\in\mathcal{U}_{h_\star}} \left\|\widetilde{\mathbf{\Sigma}}^{-1}(u)\right\|_O \sup_{u\in\mathcal{U}_{h_\star}} \left\|\widetilde{\mathbf{\Sigma}}(u) - \mathbf{\Sigma}^\star_A(u)\right\|_O \sup_{u\in\mathcal{U}_{h_\star}} \left\|\mathbf{\Sigma}_A^{\star-1}(u)\right\|_O. \end{aligned}$$

It is easy to prove (3.7) in Theorem 1(ii) from (3.6) and Theorem 1(i). □

PROOF OF THEOREM 2. By the definition of $s_{\rho(u)}(\cdot)$, it is easy to show that $\{\widetilde{\sigma}_{ij}(u) = s_{\rho(u)}(\widehat{\sigma}_{ij}(u)) \neq 0\}$ is equivalent to $\{|\widehat{\sigma}_{ij}(u)| > \rho(u)\}$ for any $u \in \mathcal{U}_{h_\star}$ and $1 \leqslant i, j \leqslant N$. Hence, $\{\widetilde{\sigma}_{ij}(u) \neq 0$ and $\sigma^\star_{ij}(u) = 0\}$ indicates that

$$\left|\widehat{\sigma}_{ij}(u) - \sigma^\star_{ij}(u)\right| > \rho(u). \tag{A.15}$$

Note that $\rho(u) = M_0(u)\tau_{T,N}$ with $\inf_{u\in\mathcal{U}} M_0(u) \geqslant c_M > 0$. From (A.15) and Proposition 1 above, taking $c_M > 0$ sufficiently large, we have

$$\begin{aligned}
&P\left(\widetilde{\sigma}_{ij}(u) \neq 0 \text{ and } \sigma^\star_{ij}(u) = 0 \text{ for } u \in \mathcal{U}_{h_\star} \text{ and } 1 \leqslant i, j \leqslant N\right) \\
&\leqslant P\left(\max_{1\leqslant i,j\leqslant N} \sup_{u\in\mathcal{U}_{h_\star}} \left|\widehat{\sigma}_{ij}(u) - \sigma^\star_{ij}(u)\right| > c_M\tau_{T,N}\right) \to 0,
\end{aligned}$$

completing the proof of Theorem 2. □

B Proofs of the technical lemmas

We next give the detailed proofs of the lemmas used in Appendix A to prove the main results.

PROOF OF LEMMA 1. We next only give a detailed proof of (A.2) as the proof of (A.1) is similar. By the definitions of $\widehat{c}_{ij,k}(u_k)$ and $c_{ij,k}(u_k)$, we have

$$\begin{aligned}
\widehat{c}_{ij,k}(u_k) - c_{ij,k}(u_k) &= \left\{\sum_{t=1}^{T-1} K\left(\frac{U_{tk}-u_k}{h_2}\right)[X_{t+1,i}X_{t+1,j} - c_{ij,k}(u_k)]\right\} \Big/ \left\{\sum_{t=1}^{T-1} K\left(\frac{U_{tk}-u_k}{h_2}\right)\right\} \\
&= \left\{\sum_{t=1}^{T-1} K\left(\frac{U_{tk}-u_k}{h_2}\right)\xi_{t+1,ij,k}\right\} \Big/ \left\{\sum_{t=1}^{T-1} K\left(\frac{U_{tk}-u_k}{h_2}\right)\right\} + \\
&\quad \left\{\sum_{t=1}^{T-1} K\left(\frac{U_{tk}-u_k}{h_2}\right)[c_{ij,k}(U_{tk}) - c_{ij,k}(u_k)]\right\} \Big/ \left\{\sum_{t=1}^{T-1} K\left(\frac{U_{tk}-u_k}{h_2}\right)\right\} \\
&=: I^{(1)}_{ij,k}(u_k) + I^{(2)}_{ij,k}(u_k),
\end{aligned} \tag{B.1}$$

where $\xi_{t+1,ij,k} = X_{t+1,i}X_{t+1,j} - c_{ij,k}(U_{tk})$.

To simplify the notation, we let $\nu_{T,N} = \sqrt{\log(N \vee T)/(Th_2)}$. First consider $I^{(1)}_{ij,k}(u_k)$ and prove that

$$\max_{1\leqslant i,j\leqslant N} \max_{1\leqslant k\leqslant p} \sup_{a_k+h_\star\leqslant u_k\leqslant b_k-h_\star} \left|\frac{1}{Th_2}\sum_{t=1}^{T-1} K\left(\frac{U_{tk}-u_k}{h_2}\right)\xi_{t+1,ij,k}\right| = O_P\left(\nu_{T,N}\right), \tag{B.2}$$

and

$$\max_{1\leqslant k\leqslant p}\sup_{a_k+h_\star\leqslant u_k\leqslant b_k-h_\star}\left|\frac{1}{Th_2}\sum_{t=1}^{T-1}K\left(\frac{U_{tk}-u_k}{h_2}\right)-f_k(u_k)\right|=O_P\left(h_2^2+\sqrt{\log T/(Th_2)}\right). \tag{B.3}$$

In fact, by (B.2) and (B.3) and noting that $f_k(\cdot)$ is positive and uniformly bounded away from zero in Assumption 1(iii), we readily have

$$\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\sup_{a_k+h_\star\leqslant u_k\leqslant b_k-h_\star}\left|I^{(1)}_{ij,k}(u_k)\right|=O_P\left(\nu_{T,N}\right). \tag{B.4}$$

We next only prove (B.2) as (B.3) can be proved in a similar (and simpler) way. Define

$$\xi^*_{t+1,ij,k}=\xi_{t+1,ij,k}I\left(|\xi_{t+1,ij,k}|\leqslant T^\iota\right),\quad \xi^\diamond_{t+1,ij,k}=\xi_{t+1,ij,k}-\xi^*_{t+1,ij,k}, \tag{B.5}$$

where ι was defined in Assumption 3(ii). Observe that

$$\begin{aligned}\frac{1}{Th_2}\sum_{t=1}^{T-1}K\left(\frac{U_{tk}-u_k}{h_2}\right)\xi_{t+1,ij,k}&=\frac{1}{Th_2}\sum_{t=1}^{T-1}K\left(\frac{U_{tk}-u_k}{h_2}\right)\xi^*_{t+1,ij,k}+\frac{1}{Th_2}\sum_{t=1}^{T-1}K\left(\frac{U_{tk}-u_k}{h_2}\right)\xi^\diamond_{t+1,ij,k}\\&=\frac{1}{Th_2}\sum_{t=1}^{T-1}K\left(\frac{U_{tk}-u_k}{h_2}\right)\left[\xi^*_{t+1,ij,k}-E(\xi^*_{t+1,ij,k})\right]+\\&\quad\frac{1}{Th_2}\sum_{t=1}^{T-1}K\left(\frac{U_{tk}-u_k}{h_2}\right)\left[\xi^\diamond_{t+1,ij,k}-E(\xi^\diamond_{t+1,ij,k})\right]\end{aligned} \tag{B.6}$$

as $E(\xi_{t+1,ij,k})=E(\xi^*_{t+1,ij,k})+E(\xi^\diamond_{t+1,ij,k})=0$.

By the moment condition (3.1) in Assumption 1(ii), we have

$$E\left(\left|\xi^\diamond_{t+1,ij,k}\right|\right)=E\left[|\xi_{t+1,ij,k}|I\left(|\xi_{t+1,ij,k}|>T^\iota\right)\right]=O\left(T^{-\iota M_2}\right), \tag{B.7}$$

where M_2 can be arbitrarily large. Then, by (B.7), Assumptions 1(ii), 2(i) and 3(iii), we have that for

any $\bar{M} > 0$,

$$
\begin{aligned}
&P\left(\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\sup_{a_k+h_\star\leqslant u_k\leqslant b_k-h_\star}\left|\frac{1}{Th_2}\sum_{t=1}^{T-1}K\left(\frac{U_{tk}-u_k}{h_2}\right)\left[\xi^{\diamond}_{t+1,ij,k}-E(\xi^{\diamond}_{t+1,ij,k})\right]\right|>\bar{M}v_{T,N}\right)\\
&\leqslant P\left(\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\sup_{a_k+h_\star\leqslant u_k\leqslant b_k-h_\star}\left|\frac{1}{Th_2}\sum_{t=1}^{T-1}K\left(\frac{U_{tk}-u_k}{h_2}\right)\xi^{\diamond}_{t+1,ij,k}\right|>\frac{1}{2}\bar{M}v_{T,N}\right)\\
&\leqslant P\left(\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\max_{1\leqslant t\leqslant T-1}\left|\xi^{\diamond}_{t+1,ij,k}\right|>0\right)\leqslant P\left(\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\max_{1\leqslant t\leqslant T-1}|\xi_{t+1,ij,k}|>T^{\iota}\right)\\
&\leqslant P\left(\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant t\leqslant T-1}|X_{t+1,i}X_{t+1,j}|>T^{\iota}-\bar{c}\right)\leqslant P\left(\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant t\leqslant T-1}\left(X^2_{t+1,i}+X^2_{t+1,j}\right)>2(T^{\iota}-\bar{c})\right)\\
&\leqslant 2P\left(\max_{1\leqslant i\leqslant N}\max_{1\leqslant t\leqslant T-1}X^2_{t+1,i}>T^{\iota}-\bar{c}\right)\leqslant 2\sum_{i=1}^{N}\sum_{t=1}^{T-1}P\left(X^2_{t+1,i}>T^{\iota}-\bar{c}\right)\\
&=O_P\left(NT\exp\{-sT^{\iota}\}\max_{1\leqslant i\leqslant N}E\left[\exp\left\{sX^2_{ti}\right\}\right]\right)=o(1) \qquad \text{(B.8)}
\end{aligned}
$$

for $0 < s < s_0$, where $\bar{c} = \max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\sup_{a_k\leqslant u_k\leqslant b_k}|c_{ij,k}(u_k)|$ is bounded by Assumption 2(i).

We next consider covering the set $\mathcal{U}_k$ by some disjoint intervals $\mathcal{U}_{k,l}$, $l = 1,\ldots,q$, with the center $u_{k,l}$ and length $h_2^2T^{-\iota}v_{T,N}$. It is easy to find that q is of order $T^{\iota}h_2^{-2}v_{T,N}^{-1}$. Note that

$$
\begin{aligned}
&\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\sup_{a_k+h_\star\leqslant u_k\leqslant b_k-h_\star}\left|\frac{1}{Th_2}\sum_{t=1}^{T-1}K\left(\frac{U_{tk}-u_k}{h_2}\right)\left[\xi^{*}_{t+1,ij,k}-E(\xi^{*}_{t+1,ij,k})\right]\right|\\
&\leqslant \max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\max_{1\leqslant l\leqslant q}\left|\frac{1}{Th_2}\sum_{t=1}^{T-1}K\left(\frac{U_{tk}-u_{k,l}}{h_2}\right)\left[\xi^{*}_{t+1,ij,k}-E(\xi^{*}_{t+1,ij,k})\right]\right|+\\
&\quad\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\sup_{u_k\in\mathcal{U}_{k,l}}\left|\frac{1}{Th_2}\sum_{t=1}^{T-1}\left[K\left(\frac{U_{tk}-u_k}{h_2}\right)-K\left(\frac{U_{tk}-u_{k,l}}{h_2}\right)\right]\left[\xi^{*}_{t+1,ij,k}-E(\xi^{*}_{t+1,ij,k})\right]\right|\\
&\leqslant \max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\max_{1\leqslant l\leqslant q}\left|\frac{1}{Th_2}\sum_{t=1}^{T-1}K\left(\frac{U_{tk}-u_{k,l}}{h_2}\right)\left[\xi^{*}_{t+1,ij,k}-E(\xi^{*}_{t+1,ij,k})\right]\right|+\\
&\quad\max_{1\leqslant k\leqslant p}\sup_{u_k\in\mathcal{U}_{k,l}}\frac{2T^{\iota}}{Th_2}\sum_{t=1}^{T-1}\left|K\left(\frac{U_{tk}-u_k}{h_2}\right)-K\left(\frac{U_{tk}-u_{k,l}}{h_2}\right)\right|\\
&\leqslant \max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\max_{1\leqslant l\leqslant q}\left|\frac{1}{Th_2}\sum_{t=1}^{T-1}K\left(\frac{U_{tk}-u_{k,l}}{h_2}\right)\left[\xi^{*}_{t+1,ij,k}-E(\xi^{*}_{t+1,ij,k})\right]\right|+O_P\left(v_{T,N}\right),
\end{aligned}
$$

where Assumption 3(i) and the definition of $\xi^{*}_{t+1,ij,k}$ in (B.5) are used.

By the exponential inequality for the α-mixing dependent sequence such as Theorem 1.3 in

Bosq (1998), we may show that

$$
\begin{aligned}
&P\left(\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\sup_{a_k+h_\star\leqslant u_k\leqslant b_k-h_\star}\left|\frac{1}{Th_2}\sum_{t=1}^{T-1}K\left(\frac{U_{tk}-u_k}{h_2}\right)\left[\xi^*_{t+1,ij,k}-E(\xi^*_{t+1,ij,k})\right]\right|>\bar{M}v_{T,N}\right)\\
&\leqslant\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{p}\sum_{l=1}^{q}P\left(\left|\frac{1}{Th_2}\sum_{t=1}^{T-1}K\left(\frac{U_{tk}-u_{k,l}}{h_2}\right)\left[\xi^*_{t+1,ij,k}-E(\xi^*_{t+1,ij,k})\right]\right|>\bar{M}v_{T,N}\right)\\
&=O\left(N^2pq\exp\{-M_*\log(N\vee T)\}\right)+O\left(N^2pq\left[T^{3+6\iota}\log(N\vee T)/h_2^3\right]^{1/4}\gamma^{M_\diamond\log(N\vee T)}\right)\\
&=o_P(1),
\end{aligned}
$$

where M_* and $M_\diamond$ are two positive constants which could be sufficiently large (by choosing $\bar{M}$ large enough), and $0<\gamma<1$ was defined in Assumption 1(i). Therefore, we have

$$
P\left(\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\sup_{a_k+h_\star\leqslant u_k\leqslant b_k-h_\star}\left|\frac{1}{Th_2}\sum_{t=1}^{T-1}K\left(\frac{U_{tk}-u_k}{h_2}\right)\left[\xi^*_{t+1,ij,k}-E(\xi^*_{t+1,ij,k})\right]\right|>\bar{M}v_{T,N}\right)=o(1) \tag{B.9}
$$

when $\bar{M}$ is sufficiently large. By (B.8) and (B.9), we can complete the proof of (B.2).

Similarly, we can also show that

$$
\begin{aligned}
&\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\sup_{a_k+h_\star\leqslant u_k\leqslant b_k-h_\star}\left|\frac{1}{Th_2}\sum_{t=1}^{T-1}\left\{K\left(\frac{U_{tk}-u_k}{h_2}\right)c_{ij,k}(U_{tk})-E\left[K\left(\frac{U_{tk}-u_k}{h_2}\right)c_{ij,k}(U_{tk})\right]\right\}\right|\\
&=O_P\left(v_{T,N}\right), \qquad\qquad (B.10)
\end{aligned}
$$

and by Taylor's expansion for $c_{ij,k}(\cdot)$ and $f_k(\cdot)$

$$
\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\sup_{a_k+h_\star\leqslant u_k\leqslant b_k-h_\star}\left|\frac{1}{h_2}E\left[K\left(\frac{U_{tk}-u_k}{h_2}\right)c_{ij,k}(U_{tk})\right]-c_{ij,k}(u_k)f_k(u_k)\right|=O_P\left(h_2^2\right). \tag{B.11}
$$

By (B.3), (B.10) and (B.11), we have

$$
\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\sup_{a_k+h_\star\leqslant u_k\leqslant b_k-h_\star}\left|I^{(2)}_{ij,k}(u_k)\right|=O_P\left(v_{T,N}+h_2^2\right). \tag{B.12}
$$

Then the proof of (A.2) is completed in view of (B.1), (B.4) and (B.12). □

PROOF OF LEMMA 2. From the definition of $(\widehat{a}_{ij,1},\ldots,\widehat{a}_{ij,p})^\intercal$ in (2.10), we have

$$
(\widehat{a}_{ij,1},\ldots,\widehat{a}_{ij,p})^\intercal=\widehat{\mathbf{\Omega}}_{ij}^{-1}\widehat{\mathbf{V}}_{ij}=\left[\widetilde{\mathbf{\Omega}}_{ij}+\left(\widehat{\mathbf{\Omega}}_{ij}-\widetilde{\mathbf{\Omega}}_{ij}\right)\right]^{-1}\left[\widetilde{\mathbf{V}}_{ij}+\left(\widehat{\mathbf{V}}_{ij}-\widetilde{\mathbf{V}}_{ij}\right)\right], \tag{B.13}
$$

where $\widehat{\mathbf{\Omega}}_{ij} := \widehat{\mathbf{\Omega}}_{XX,ij}$, $\widehat{\mathbf{V}}_{ij} := \widehat{\mathbf{V}}_{XX,ij}$, $\widetilde{\mathbf{\Omega}}_{ij}$ is a $p \times p$ matrix with the (k,l)-entry being

$$\widetilde{\omega}_{ij,kl} = \frac{1}{T-1}\sum_{t=1}^{T-1} c^c_{ij,k}(U_{tk})c^c_{ij,l}(U_{tl}), \;\; c^c_{ij,k}(U_{tk}) = c_{ij,k}(U_{tk}) - \mathsf{E}\left[c_{ij,k}(U_{tk})\right],$$

and $\widetilde{\mathbf{V}}_{ij}$ is a p-dimensional column vector with the k-th element being

$$\widetilde{v}_{ij,k} = \frac{1}{T-1}\sum_{t=1}^{T-1} c^c_{ij,k}(U_{tk})X^*_{t+1,(i,j)}, \;\; X^*_{t+1,(i,j)} = X_{t+1,i}X_{t+1,j} - \mathsf{E}\left[X_{t+1,i}X_{t+1,j}\right].$$

Following the proof of (B.2) above, we may show that

$$\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\left|\frac{1}{T-1}\sum_{t=1}^{T-1} c_{ij,k}(U_{tk}) - \mathsf{E}\left[c_{ij,k}(U_{tk})\right]\right| = O_P\left(\sqrt{\log(N\vee T)/T}\right) \tag{B.14}$$

and

$$\max_{1\leqslant i,j\leqslant N}\max_{1\leqslant k\leqslant p}\left|\frac{1}{T-1}\sum_{t=1}^{T-1} X_{t+1,i}X_{t+1,j} - \mathsf{E}\left[X_{t+1,i}X_{t+1,j}\right]\right| = O_P\left(\sqrt{\log(N\vee T)/T}\right). \tag{B.15}$$

By (B.14), (B.15) and Lemma 1, we readily have

$$\max_{1\leqslant i,j\leqslant N}\left\|\widehat{\mathbf{\Omega}}_{ij} - \widetilde{\mathbf{\Omega}}_{ij}\right\|_F = O_P\left(\sqrt{\log(N\vee T)/T} + \sqrt{\log(N\vee T)/(Th_2)} + h_2^2\right) \tag{B.16}$$

and

$$\max_{1\leqslant i,j\leqslant N}\left\|\widehat{\mathbf{V}}_{ij} - \widetilde{\mathbf{V}}_{ij}\right\| = O_P\left(\sqrt{\log(N\vee T)/T} + \sqrt{\log(N\vee T)/(Th_2)} + h_2^2\right). \tag{B.17}$$

By (B.13), (B.16) and (B.17), we have

$$(\widehat{a}_{ij,1},\ldots,\widehat{a}_{ij,p})^\intercal = \widetilde{\mathbf{\Omega}}_{ij}^{-1}\widetilde{\mathbf{V}}_{ij} + O_P\left(\sqrt{\log(N\vee T)/T} + \sqrt{\log(N\vee T)/(Th_2)} + h_2^2\right). \tag{B.18}$$

Again, following the proof of (B.2), we can easily show that

$$\max_{1\leqslant i,j\leqslant N}\left\|\widetilde{\mathbf{\Omega}}_{ij} - \mathbf{\Omega}_{ij}\right\|_F = O_P\left(\sqrt{\log(N\vee T)/T}\right) \tag{B.19}$$

and

$$\max_{1\leqslant i,j\leqslant N}\left\|\widetilde{\mathbf{V}}_{ij} - \mathbf{V}_{ij}\right\| = O_P\left(\sqrt{\log(N\vee T)/T}\right), \tag{B.20}$$

which together with (B.18), indicates that

$$\max_{1\leqslant i,j\leqslant N}\sum_{k=1}^{p}\left|\widehat{a}_{ij,k}-a^{\star}_{ij,k}\right|=O_P\left(\sqrt{\log(N\vee T)/T}+\sqrt{\log(N\vee T)/(Th_2)}+h_2^2\right). \tag{B.21}$$

We finally consider $\widehat{a}_{ij,0}$. Note that uniformly for $1\leqslant i,j\leqslant N$,

$$\begin{aligned}
\widehat{a}_{ij,0}&=\frac{1}{T-1}\sum_{t=1}^{T-1}X_{t+1,i}X_{t+1,j}-\sum_{k=1}^{p}\widehat{a}_{ij,k}\left[\frac{1}{T-1}\sum_{t=1}^{T-1}\widehat{c}_{ij,k}(U_{tk})\right]\\
&=\frac{1}{T-1}\sum_{t=1}^{T-1}X_{t+1,i}X_{t+1,j}-\sum_{k=1}^{p}\widehat{a}_{ij,k}\left[\frac{1}{T-1}\sum_{t=1}^{T-1}c_{ij,k}(U_{tk})+O_P\left(\sqrt{\log(N\vee T)/(Th_2)}+h_2^2\right)\right]\\
&=E\left(X_{ti}X_{tj}\right)+O_P\left(\sqrt{\log(N\vee T)/T}\right)-\sum_{k=1}^{p}\widehat{a}_{ij,k}\left[E\left(X_{ti}X_{tj}\right)+O_P\left(\sqrt{\log(N\vee T)/(Th_2)}+h_2^2\right)\right]\\
&=\left(1-\sum_{k=1}^{p}\widehat{a}_{ij,k}\right)E\left(X_{ti}X_{tj}\right)+O_P\left(\sqrt{\log(N\vee T)/T}+\sqrt{\log(N\vee T)/(Th_2)}+h_2^2\right)\\
&=\left(1-\sum_{k=1}^{p}a^{\star}_{ij,k}\right)E\left(X_{ti}X_{tj}\right)+O_P\left(\sqrt{\log(N\vee T)/T}+\sqrt{\log(N\vee T)/(Th_2)}+h_2^2\right)\\
&=a^{\star}_{ij,0}+O_P\left(\sqrt{\log(N\vee T)/T}+\sqrt{\log(N\vee T)/(Th_2)}+h_2^2\right),
\end{aligned} \tag{B.22}$$

where (B.14), (B.15) and (B.21) have been used.

From (B.21) and (B.22), we can complete the proof of (B.3). The proof of (B.4) is similar, so details are omitted here. The proof of Lemma 2 has been completed. □

References

AÏT-SAHALIA, Y. AND BRANDT, M. W. (2001). Variable selection for portfolio choice. *Journal of Finance* **56**, 1297–1351.

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis (3rd Edition)*. Wiley Series in Probability and Statistics.

BACK, K. E. (2010). *Asset Pricing and Portfolio Choice Theory*. Oxford University Press.

BAI, J. AND NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191–221.

BICKEL, P. AND LEVINA, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics* **36**, 2577–2604.

BICKEL, P. AND LEVINA, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36**, 199–227.

BOSQ, D. (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. Springer.

BRANDT, M. W. (1999). Estimating portfolio and consumption choice: a conditional Euler equations approach. *Journal of Finance* **54**, 1609–1645.

BRANDT, M. W. (2010). Portfolio choice problems. *Handbook of Financial Econometrics, Volume 1: Tools and Techniques* (Editors: Aït-Sahalia Y. and Hansen L. P.), 269-336.

CAI, T. T. AND LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106**, 672–684.

CHEN, J., LI, D., LINTON, O. AND LU, Z. (2016). Semiparametric dynamic portfolio choice with multiple conditioning variables. *Journal of Econometrics* **194**, 309–318.

CHEN, J., LI, D., LINTON, O. AND LU, Z. (2018). Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *Journal of the American Statistical Association* **113**, 919–932 .

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, **76**, North Holland, Amsterdam.

CHEN, X., XU, M. AND WU, W. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Annals of Statistics* **41**, 2994–3021.

CHEN, Z. AND LENG, C. (2016). Dynamic covariance models. *Journal of the American Statistical Association* **111**, 1196–1207.

DEMIGUEL, V., GARLAPPI, L., NOGALES, F. J. AND UPPAL, R. (2009). A generalized approach to portfolio optimization: improving performance by constraining portfolio norms. *Management Science* **55**, 798–812.

DEMIGUEL, V., GARLAPPI, L. AND UPPAL, R. (2009). Optimal versus naive diversification: how inefficient is the 1/N portfolio strategy? *Review of Financial Studies* **22**, 1915–1953.

ENGLE, R. F. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business and Economic Statistics* **20**, 339–350.

ENGLE, R. F., LEDOIT, O. AND WOLF, M. (2017). Large dynamic covariance matrices. Forthcoming in *Journal of Business and Economic Statistics*.

FAMA, E. F. (1970). Multiperiod consumption-investment decisions. *American Economic Review* **60**, 163–174.

FAN, J., FAN, Y. AND LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* **147**, 186-197.

FAN, J., FENG, Y., JIANG, J. AND TONG, X. (2016). Feature augmentation via nonparametrics and selection (FANS) in high dimensional classification. *Journal of American Statistical Association* **111**, 275–287.

FAN, J. AND GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

FAN, J., LIAO, Y. AND MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society, Series B* **75**, 603–680.

FRAHM, G. AND MEMMEL, C. (2010). Dominating estimators for minimum-variance portfolios. *Journal of Econometrics* **159**, 289–302.

GLAD, I., HJORT, N. L. AND USHAKOV, N. G. (2003). Correction of Density Estimators that are not Densities. *Scandinavian Journal of Statistics* **30** 415–427.

GUO, S., BOX, J. AND ZHANG, W. (2017). A dynamic structure for high dimensional covariance matrices and its application in portfolio allocation. *Journal of the American Statistical Association* **112**, 235–253.

KAN, J. AND ZHOU, G. (2007). Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis* **42**, 621–656.

LAM, C. AND FAN, J. (2009). Sparsity and rates of convergence in large covariance matrix estimation. *Annals of Statistics* **37**, 4254–4278.

LAM, C. AND YAO, Q. (2012). Factor modelling for high-dimensional time series: inference for the number of factors. *Annals of Statistics*, **40**, 694–726.

LEDOIT, O. AND WOLF, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* **10**, 603–621.

LEDOIT, O. AND WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88**, 365–411.

LEDOIT, O. AND WOLF, M. (2017). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *Review of Financial Studies* **30**, 4349–4388.

LI, D., LINTON, O. AND LU, Z. (2015). A flexible semiparametric forecasting model for time series. *Journal of Econometrics* **187**, 345–357.

LI, Y., LI, D. AND FRYZLEWICZ, P. (2018). Detection of multiple structural breaks in large covariance matrix. *Working Paper*.

MARKOWITZ, H. M. (1952). Portfolio selection. *Journal of Finance* **7**, 77–91.

MERTON, R. C. (1969). Lifetime portfolio selection under uncertainty: the continuous time case. *Review of Economics and Statistics* **51**, 247–257.

PESARAN, M. H. AND ZAFFARONI, P. (2009). Optimality and diversifiability of mean variance and arbitrage pricing portfolios. CESifo Working Paper Series No. 2857.

ROTHMAN, A. J., LEVINA, E. AND ZHU, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* **104**, 177–186.

STOCK, J. H. AND WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* **97** 1167–1179.

TU, J. AND ZHOU, G. (2011). Markowitz meets Talmud: a combination of sophisticated and naive diversification strategies. *Journal of Financial Economics* **99**, 204–215.

VOGT, M. (2012). Nonparametric regression for locally stationary time series. *The Annals of Statistics* **40**, 2601–2633.

WAND, M. P. AND JONES, M. C. (1995). *Kernel Smoothing*. Chapman and Hall.

WU, W. AND POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90**, 831–844.