Article:
Ke, Yuan, Lian, Heng and Zhang, Wenyang orcid.org/0000-0001-8391-1122 (2020) High Dimensional Dynamic Covariance Matrices with Homogeneous Structure. Journal of Business and Economic Statistics. pp. 1-16. ISSN 0735-0015

https://doi.org/10.1080/07350015.2020.1779079

High Dimensional Dynamic Covariance Matrices with Homogeneous Structure *

Yuan Ke
Department of Statistics
University of Georgia

Heng Lian
Department of Mathematics
City University of Hong Kong

Wenyang Zhang †
Department of Mathematics
The University of York, United Kingdom

June 3, 2020

Abstract

High dimensional covariance matrices appear in many disciplines. Much literature has devoted to the research in high dimensional constant covariance matrices. However, constant covariance matrices are not sufficient in applications, e.g. in portfolio allocation, dynamic covariance matrices would be more appropriate. As argued in this paper, there are two difficulties in the introduction of dynamic structures into covariance matrices: (1) simply assuming each entry of a covariance matrix is a function of time to introduce the dynamic needed would not work; (2) there is a risk of having too many unknowns to estimate due to the high dimensionality. In this paper, we propose a dynamic structure embedded with a homogeneous structure. We will demonstrate the proposed dynamic structure makes more sense in applications and avoids, in the meantime, too many unknown parameters/functions to estimate, due to the embedded homogeneous structure. An estimation procedure is also proposed to estimate the proposed high dimensional dynamic covariance matrices, and asymptotic properties are established to justify the proposed estimation procedure. Intensive simulation studies show the proposed estimation procedure works very well when the sample size is finite. Finally, we apply the proposed high dimensional dynamic covariance matrices to portfolio allocation. It is interesting to see the resulting portfolio yields much better returns than some commonly used ones.

*This research is supported by National Natural Science Foundation of China (Grant Number 11931014)

†The corresponding author, Department of Mathematics, University of York, York, YO10 5DD, United Kingdom, Email: wenyang.zhang@york.ac.uk.

KEY WORDS: B-Spline, high dimensional dynamic covariance matrices, homogeneous structure, portfolio allocation, single index models.

SHORT TITLE: Homogeneous Structure for HDCM.

1 Introduction

1.1 Motivation

Covariance matrices appear in many disciplines, such as economics, finance, engineering, psychology, and biology, to name but a few. The estimation of covariance matrices has a very long history. Traditionally, sample covariance matrices are used to estimate the covariance matrices. In many applications, we often come across the need for a function of a covariance matrix, e.g. in portfolio allocation, we need the inverse of the covariance matrix of the returns of the assets under consideration. When the dimension of the covariance matrix is big, the sample covariance matrix would not work well, this is because the estimation error accumulates very quickly to reach an unacceptable level when computing the function of the estimated covariance matrix.

During the past decades, there is much literature devoting to the research in the estimation of constant high dimensional covariance matrices. See, Wu and Pourahmadi (2003), Sun *et al.*(2007), Fan *et al.*(2008), Bickel and Levina (2008a, 2008b), El Karoui (2008), Rothman *et al.*(2009), Yuan (2010), Fan *et al.*(2011), Berthet and Rigollet (2013), Birnbaum *et al.*(2013), Fang *et al.*(2016), Guo *et al.*(2017), Avella-Medina *et al.*(2018), Fan *et al.*(2018), Ke *et al.*(2019), and the references therein. In many applications, constant covariance matrices are not suitable. For example, in portfolio allocation, we would expect that a good portfolio allocation should be dynamic, this is because an optimal portfolio allocation today may not be optimal tomorrow. Therefore, the covariance matrix used in forming a portfolio allocation has to be dynamic. To introduce a dynamic structure into a covariance matrix, simply assuming each entry of this covariance matrix is an unknown function of time would not work in many cases. For example, in portfolio allocation or risk management, the main purpose is for prediction. If we assume each entry of the covariance matrix used is an unknown function of time, we would not be able to estimate this unknown function well at time point $n+1$ when we have observations up to the time point n. This is because the unknown function can go smoothly either up or down at a time point $n+1$, and we do not have the information about which way the unknown function may go. Therefore, the resulting portfolio allocation or risk management would not work very well. Another commonly used approach to incorporate a dynamic pattern in a covariance matrix is to estimate the covariance matrix only based on the observations in a moving window. This approach is the same as that assuming each entry of the covariance matrix is an

unknown function of time and estimate it by the local constant estimation.

Research in dynamic covariance matrices has been attracting many scholars. Relevant literature includes Bollerslev *et al.*(1988), Harvey *et al.*(1994), Engle (2002, 2009), Ledoit and Wolf (2004, 2020), Bauwens *et al.*(2006), Asai *et al.*(2006), Yu and Meyer (2006), Silvennoinen and Teräsvirta (2009), Chib *et al.*(2009), Ledoit and Wolf (2012), Almeida *et al.*(2018), Francq and Zakoian (2019), Boudt *et al.*(2019), Engle *et al.*(2019), Kastner (2019), Pakel *et al.*(2020), and the references therein.

Taking different approach, based on the autoregressive idea and factor models, Guo *et al.*(2017) proposed the following models for the components of the vector $\mathbf{Y}_t$, $\mathbf{Y}_t = (y_{1,t},\ \cdots,\ y_{p_n,t})^{\mathrm{T}}$, to which we are interested in its covariance matrix

$$y_{i,t} = a_{i,0}(\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + \mathbf{X}_t^{\mathrm{T}}\mathbf{a}_i(\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + \epsilon_{i,t}, \quad i = 1,\ \cdots,\ p_n;\ \ t = 2,\ \cdots,\ n \tag{1.1}$$

where $\mathbf{X}_t$ is a q dimensional factor, and

$$\boldsymbol{\beta} = (\beta_1,\ \cdots,\ \beta_q)^{\mathrm{T}}, \quad \|\boldsymbol{\beta}\| = 1, \quad \beta_1 > 0, \quad \mathbf{a}_i(\cdot) = (a_{i,1}(\cdot),\ \cdots,\ a_{i,q}(\cdot))^{\mathrm{T}}$$

Based on (1.1), the covariance matrix of $\mathbf{Y}_t$ can be derived, and the covariance matrix comes with a nice dynamic structure.

The dynamic covariance matrices based on models (1.1) have two problems: (1) They may result in serious bias in some applications, this is because models (1.1), based on which the dynamic covariance matrices are introduced, do not appreciate the heterogeneity on the index $\boldsymbol{\beta}$ among the components of $\mathbf{Y}_t$ as the same $\boldsymbol{\beta}$ is used for all components. (2) There are $(q+1)p_n$ unknown functions involved, which is too many for high dimensional cases. This can cause serious trouble on the variance side and make the final estimators of the covariance matrices very unstable. In this paper, we are going to take a different approach to introduce a dynamic structure for high dimensional covariance matrices to overcome these problems.

1.2 The models

Suppose $(\mathbf{X}_t^{\mathrm{T}}, \mathbf{Y}_t^{\mathrm{T}})$, $t = 1,\ \cdots,\ n$, is a time series, where $\mathbf{Y}_t$ is a p_n dimensional vector and $\mathbf{X}_t$ is a q dimensional factor. An underlying assumption throughout this paper is that $p_n \longrightarrow \infty$ when $n \longrightarrow \infty$, and q is fixed. See Guo *et al.*(2017) for reasoning of this assumption. Also, we assume that $\{\mathbf{X}_t, t = 0,\ \cdots,\ n\}$ is a stationary Markov process. Let $\mathbf{Y}_t = (y_{1,t},\ \cdots,\ y_{p_n,t})^{\mathrm{T}}$, we assume the ith component of $\mathbf{Y}_t$, which is also called the ith individual throughout this paper, follows

$$y_{i,t} = a_{i,0}(\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_i) + \mathbf{X}_t^{\mathrm{T}}\mathbf{a}_i(\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_i) + \epsilon_{i,t}, \quad \|\boldsymbol{\beta}_i\| = 1, \quad \beta_{i,1} > 0, \tag{1.2}$$

where $\boldsymbol{\beta}_i = (\beta_{i,1},\ \cdots,\ \beta_{i,q})^{\mathrm{T}}$, $\mathbf{a}_i(\cdot) = (a_{i,1}(\cdot),\ \cdots,\ a_{i,q}(\cdot))^{\mathrm{T}}$ is a factor loading vector, and $\beta_{i,j}$s and $a_{i,j}(\cdot)$s have the following unknown homogeneous structure

$$\beta_{i,j} = \begin{cases} \beta_{(1)} & \text{when } (i,j) \in \mathcal{D}_1, \\ \vdots & \vdots \\ \beta_{(H)} & \text{when } (i,j) \in \mathcal{D}_H, \end{cases} \qquad a_{i,j}(\cdot) = \begin{cases} a_{(1)}(\cdot) & \text{when } (i,j) \in \mathcal{Q}_1, \\ \vdots & \vdots \\ a_{(\mathcal{N})}(\cdot) & \text{when } (i,j) \in \mathcal{Q}_{\mathcal{N}}, \end{cases} \tag{1.3}$$

$\{\mathcal{D}_k :\ k = 1,\ \cdots,\ H\}$ is an unknown partition of set $\{(i,j) :\ i = 1,\ \cdots,\ p_n;\ j = 2,\ \cdots,\ q\}$, $\{\mathcal{Q}_k :\ k = 1,\ \cdots,\ \mathcal{N}\}$ is an unknown partition of set $\{(i,j) :\ i = 1,\ \cdots,\ p_n;\ j = 0,\ \cdots,\ q\}$. $\{\boldsymbol{\epsilon}_t = (\epsilon_{1,t},\ \cdots,\ \epsilon_{p_n,t})^{\mathrm{T}},\ t = 1,\ \cdots,\ n\}$ are random errors which are independent of $\{\mathbf{X}_t,\ t = 0,\ \cdots,\ n\}$. Note that due to the unit norm constraint $\|\boldsymbol{\beta}_i\| = 1$, the value $\beta_{i,1}$ is determined by other components of $\boldsymbol{\beta}_i$ and thus we only specify partition for $\beta_{i,2}, \ldots, \beta_{i,q}$. We assume

$$E(\boldsymbol{\epsilon}_t | \{\boldsymbol{\epsilon}_l :\ l < t\}) = \mathbf{0}, \quad \mathrm{cov}(\boldsymbol{\epsilon}_t | \{\boldsymbol{\epsilon}_l :\ l < t\}) = \boldsymbol{\Sigma}_{0,t} = \mathrm{diag}\left(\sigma_{1,t}^2,\ \cdots,\ \sigma_{p_n,t}^2\right) \tag{1.4}$$

where

$$\sigma_{k,t}^2 = \alpha_{k,0} + \sum_{i=1}^{m} \alpha_{k,i} \epsilon_{k,t-i}^2 + \sum_{j=1}^{s} \gamma_{k,j} \sigma_{k,t-j}^2, \tag{1.5}$$

for each $k = 1,\ \cdots,\ p_n$ and for some integers m and s. $\alpha_{k,i}$s and $\gamma_{k,j}$s also enjoy the following unknown homogeneous structure

$$\alpha_{i,j} = \begin{cases} \alpha_{(1)} & \text{when } (i,j) \in \mathcal{A}_1, \\ \vdots & \vdots \\ \alpha_{(\varsigma)} & \text{when } (i,j) \in \mathcal{A}_\varsigma, \end{cases} \qquad \gamma_{i,j} = \begin{cases} \gamma_{(1)} & \text{when } (i,j) \in \Gamma_1, \\ \vdots & \vdots \\ \gamma_{(\tau)} & \text{when } (i,j) \in \Gamma_\tau, \end{cases} \tag{1.6}$$

$\{\mathcal{A}_k :\ k = 1,\ \cdots,\ \varsigma\}$ is an unknown partition of set $\{(i,j) :\ i = 1,\ \cdots,\ p_n;\ j = 0,\ \cdots,\ m\}$, $\{\Gamma_k :\ k = 1,\ \cdots,\ \tau\}$ is an unknown partition of set $\{(i,j) :\ i = 1,\ \cdots,\ p_n;\ j = 1,\ \cdots,\ s\}$.

The model (1.2) with (1.3), (1.4), (1.5) and (1.6) is the model this paper is going to address, in which, H, $\mathcal{N}$, ς, τ, $\beta_{(i)}$, $i = 1,\ \cdots,\ H$, $a_{(j)}(\cdot)$, $j = 1,\ \cdots,\ \mathcal{N}$, $\alpha_{(k)}$, $k = 1,\ \cdots,\ \varsigma$, $\gamma_{(l)}$, $l = 1,\ \cdots,\ \tau$, the partitions $\{\mathcal{D}_k :\ k = 1,\ \cdots,\ H\}$, $\{\mathcal{Q}_k :\ k = 1,\ \cdots,\ \mathcal{N}\}$, $\{\mathcal{A}_k :\ k = 1,\ \cdots,\ \varsigma\}$, and $\{\Gamma_k :\ k = 1,\ \cdots,\ \tau\}$ are all unknowns to be estimated.

Let $\mathcal{F}_t$ be the $\sigma-$algebra generated by $\{(\mathbf{X}_l^{\mathrm{T}},\ \boldsymbol{\epsilon}_l^{\mathrm{T}}) :\ l \le t\}$. The main focus of this paper is on the conditional covariance matrix

$$\mathrm{cov}(\mathbf{Y}_t | \mathcal{F}_{t-1}) = \begin{pmatrix} \mathbf{a}_1^{\mathrm{T}}(\mathbf{X}_{t-1}^{\mathrm{T}} \boldsymbol{\beta}_1) \\ \vdots \\ \mathbf{a}_{p_n}^{\mathrm{T}}(\mathbf{X}_{t-1}^{\mathrm{T}} \boldsymbol{\beta}_{p_n}) \end{pmatrix} \boldsymbol{\Sigma}_x(\mathbf{X}_{t-1}) \left(\mathbf{a}_1(\mathbf{X}_{t-1}^{\mathrm{T}} \boldsymbol{\beta}_1),\ \cdots,\ \mathbf{a}_{p_n}(\mathbf{X}_{t-1}^{\mathrm{T}} \boldsymbol{\beta}_{p_n})\right) + \boldsymbol{\Sigma}_{0,t} \tag{1.7}$$

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

Remark 1. The proposed dynamic structure for high dimensional covariance matrices is very different from those appear in the literature mentioned in the third paragraph of Section 1.1. The dynamic structure there is mainly based on the GARCH models. The proposed dynamic structure is based on the ideas of Fama-French common risk factors, single index varying coefficient models and homogeneity pursuit. Due to the homogeneity pursuit, the proposed dynamic structure is parsimonious, and our empirical studies show the proposed dynamic structure is appropriate in real life and works well in real data analysis.

Remark 2. In the dynamic structure introduced in Guo *et al.*(2017), they use the same index $\boldsymbol{\beta}$ for all components of the $\mathbf{Y}_t$, this may lead to serious misspecification, which would result in serious bias in the final estimator of the underlying true covariance matrix in real life application. In the proposed dynamic structure, we use different index for different component, namely $\boldsymbol{\beta}_i$ for the ith component of $\mathbf{Y}_t$, to make the modelling more flexible and avoid misspecification, in the meantime, we apply homogeneity pursuit to avoid overfitting, therefore, make the modelling parsimonious.

Remark 3. The proposed homogeneous structure imposes a parsimonious yet flexible structure embedded in the high-dimensional parameter space. It is flexible in the sense that it includes many widely used low-dimensional structures as its special cases. For example, the sparsity structure can be considered as a dominant group of 0's plus several non-zero groups; the bi-clustering structure can be considered as a homogeneity structure with a known group number of two; and the tree structure can also be recovered by the proposed binary segmentation algorithm. In some modern data based economic and business applications, learning a flexible low-dimensional structure has become the primary objective over the coefficient estimation and inference. Besides the problem studied in this manuscript, the homogeneity structure has also been studied for large panel data analysis. Another hot topic in business analytic is to cluster the social media users across the topics and geological locations into homogeneity groups, such that further precision business actions can be applied to each group.

2 Estimation procedure

In this section, we introduce an estimation procedure for $\text{cov}(\mathbf{Y}_t|\mathcal{F}_{t-1})$. We will first estimate $\boldsymbol{\beta}_i$s, $\mathbf{a}_i(\cdot)$s, $\boldsymbol{\Sigma}_x(\cdot)$, $\alpha_{k,i}$s and $\gamma_{k,j}$s based on the model (1.2) together with (1.3), (1.4), (1.5) and (1.6), and denote the resulting estimators by $\widehat{\boldsymbol{\beta}}_i$, $\widehat{\mathbf{a}}_i(\cdot)$, $i = 1, \cdots, p_n$, $\widehat{\boldsymbol{\Sigma}}_x(\cdot)$, $\widehat{\alpha}_{k,i}$ and $\widehat{\gamma}_{k,j}$ for $i = 0, \cdots, m$ and $j = 1, \cdots, s$. Let $\widehat{\boldsymbol{\Sigma}}_{0,t}$ be $\boldsymbol{\Sigma}_{0,t}$ with $\alpha_{k,i}$ and $\gamma_{k,j}$ being replaced by $\widehat{\alpha}_{k,i}$ and $\widehat{\gamma}_{k,j}$ respectively.

We use

$$\widehat{\mathrm{cov}}(\mathbf{Y}_t|\mathcal{F}_{t-1}) = \begin{pmatrix} \widehat{\mathbf{a}}_1^{\mathrm{T}}(\mathbf{X}_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_1) \\ \vdots \\ \widehat{\mathbf{a}}_{p_n}^{\mathrm{T}}(\mathbf{X}_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{p_n}) \end{pmatrix} \widehat{\boldsymbol{\Sigma}}_x(\mathbf{X}_{t-1}) \left(\widehat{\mathbf{a}}_1(\mathbf{X}_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_1),\ \cdots,\ \widehat{\mathbf{a}}_{p_n}(\mathbf{X}_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{p_n})\right) + \widehat{\boldsymbol{\Sigma}}_{0,t} \quad (2.1)$$

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

2.1 Estimation of $a_{i,0}(\cdot)$s, $\boldsymbol{\beta}_i$s and $\mathbf{a}_i(\cdot)$s

Our approach to deal with the unknown functions $a_{i,j}(\cdot)$, $i = 1,\ \cdots,\ p_n$, $j = 0,\ \cdots,\ q$, in (1.2) is based on the B-Spline. The reason for us to use B-Spline rather than kernel smoothing is for the concern of homogeneity pursuit in $a_{i,j}(\cdot)$s, see Remark 4 at the end of this section for more details. To achieve the best result for the homogeneity pursuit, we have to decompose all $a_{i,j}(\cdot)$s by the same B-Spline basis, $\mathbf{B}(\cdot) = (B_1(\cdot),\ \cdots,\ B_K(\cdot))^{\mathrm{T}}$. For each i, $i = 1,\ \cdots,\ p_n$, let $\widetilde{\boldsymbol{\beta}}_i$ be the estimate of $\boldsymbol{\beta}_i$ obtained, based on the observations for the ith individual, by a standard estimation procedure for the varying coefficient single index models, and

$$a = \min_{1\le i\le p_n}\min_{0\le t\le n} \mathbf{X}_t^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_i, \quad b = \max_{1\le i\le p_n}\max_{0\le t\le n} \mathbf{X}_t^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_i.$$

We use the B-Spline basis of order 3 in this paper, and the basis, $\mathbf{B}(\cdot)$, is formed by the equally spaced knots, τ_k, $k = 0,\ \cdots,\ K-2$, on the interval $[a,\ b]$, with $\tau_0 = a$ and $\tau_{K-3+1} = b$. K can be selected by either cross-validation or BIC. Based on the basis $\mathbf{B}(\cdot)$, $a_{i,j}(\cdot)$ can be decomposed as

$$a_{i,j}(\cdot) \approx \mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\theta}_{ij}, \quad (2.2)$$

where $\boldsymbol{\theta}_{ij} = (\theta_{ij,1},\ \cdots,\ \theta_{ij,K})^{\mathrm{T}}$. So, to get the estimator of $a_{i,j}(\cdot)$, we only need to get the estimator of $\boldsymbol{\theta}_{ij}$.

The estimation procedure for $\boldsymbol{\beta}_i$s and $\boldsymbol{\theta}_{ij}$s consists of three stages: initial estimation, homogeneity pursuit and final estimation, which is detailed as follows:

Stage 1 (*Initial Estimation*). For each i, based on the observations for the ith individual, approximating $a_{i,j}(\cdot)$ by its decomposition (2.2) and applying the least squares estimation method, we have the following objective function

$$\sum_{t=2}^{n}\left(y_{it} - \boldsymbol{\theta}_{i0}^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_i) - \mathbf{B}(\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_i)^{\mathrm{T}}\boldsymbol{\Theta}_i\mathbf{X}_t\right)^2, \quad (2.3)$$

where $\boldsymbol{\Theta}_i = (\boldsymbol{\theta}_{i1},\ \cdots,\ \boldsymbol{\theta}_{iq})$. Minimise (2.3) with respect to $\boldsymbol{\beta}_i$, $\boldsymbol{\Theta}_i$ and $\boldsymbol{\theta}_{i0}$, and denote the resulting minimiser by $\widetilde{\boldsymbol{\beta}}_i$, $\widetilde{\boldsymbol{\Theta}}_i$ and $\widetilde{\boldsymbol{\theta}}_{i0}$. We will show how to conduct the minimisation in Section 2.4.

Stage 2 (*Homogeneity Pursuit*). Let $\widetilde{\beta}_{ij}$ be the jth component of $\widetilde{\boldsymbol{\beta}}_i$. We sort $\widetilde{\beta}_{ij}$, $i = 1, \cdots, p_n$, $j = 2, \cdots, q$, in ascending order, and denote them by

$$b_{(1)} \leq \cdots \leq b_{((q-1)p_n)}$$

We use R_{ij} to denote the rank of $\widetilde{\beta}_{ij}$. Identifying the homogeneity among $\widetilde{\beta}_{ij}$, $i = 1, \cdots, p_n$, $j = 2, \cdots, q$, is equivalent to detecting the change points among $b_{(l)}$, $l = 1, \cdots, (q-1)p_n$. To this end, we apply the Binary Segmentation algorithm as follows.

For any $1 \leq i < j \leq (q-1)p_n$, let

$$\Delta_{ij}(\kappa) = \sqrt{\frac{(j-\kappa)(\kappa-i+1)}{j-i+1}} \left| \frac{\sum_{l=\kappa+1}^{j} b_{(l)}}{j-\kappa} - \frac{\sum_{l=i}^{\kappa} b_{(l)}}{\kappa-i+1} \right|$$

Given a threshold δ, which can be selected by AIC or BIC in practice, the Binary Segmentation algorithm to detect the change points works as follows

(1) Find $\widehat{k}_1$ such that

$$\Delta_{1,(q-1)p_n}(\widehat{k}_1) = \max_{1 \leq \kappa < (q-1)p_n} \Delta_{1,(q-1)p_n}(\kappa).$$

If $\Delta_{1,(q-1)p_n}(\widehat{k}_1) \leq \delta$, there is no change point among $b_{(l)}$, $l = 1, \cdots, (q-1)p_n$, and the process of detection ends. Otherwise, add $\widehat{k}_1$ to the set of change points and divide the region $\{\kappa : \ 1 \leq \kappa \leq (q-1)p_n\}$ into two subregions: $\{\kappa : \ 1 \leq \kappa \leq \widehat{k}_1\}$ and $\{\kappa : \ \widehat{k}_1 + 1 \leq \kappa \leq (q-1)p_n\}$.

(2) Detect the change points in the two subregions obtained in (1), respectively. Let us deal with the region $\{\kappa : \ 1 \leq \kappa \leq \widehat{k}_1\}$ first. Find $\widehat{k}_2$ such that

$$\Delta_{1,\widehat{k}_1}(\widehat{k}_2) = \max_{1 \leq \kappa < \widehat{k}_1} \Delta_{1,\widehat{k}_1}(\kappa).$$

If $\Delta_{1,\widehat{k}_1}(\widehat{k}_2) \leq \delta$, there is no change point in the region $\{\kappa : \ 1 \leq \kappa \leq \widehat{k}_1\}$. Otherwise, add $\widehat{k}_2$ to the set of change points and divide the region $\{\kappa : \ 1 \leq \kappa \leq \widehat{k}_1\}$ into two subregions: $\{\kappa : \ 1 \leq \kappa \leq \widehat{k}_2\}$ and $\{\kappa : \ \widehat{k}_2 + 1 \leq \kappa \leq \widehat{k}_1\}$. For the region $\{\kappa : \ \widehat{k}_1 + 1 \leq \kappa \leq (q-1)p_n\}$, we find $\widehat{k}_3$ such that

$$\Delta_{\widehat{k}_1+1,(q-1)p_n}(\widehat{k}_3) = \max_{\widehat{k}_1+1 \leq \kappa < (q-1)p_n} \Delta_{\widehat{k}_1+1,(q-1)p_n}(\kappa).$$

If $\Delta_{\widehat{k}_1+1,(q-1)p_n}(\widehat{k}_3) \leq \delta$, there is no change point in the region $\{\kappa : \ \widehat{k}_1 + 1 \leq \kappa \leq (q-1)p_n\}$. Otherwise, add $\widehat{k}_3$ to the set of change points and divide the region $\{\kappa : \ \widehat{k}_1 + 1 \leq \kappa \leq (q-1)p_n\}$ into two subregions: $\{\kappa : \ \widehat{k}_1 + 1 \leq \kappa \leq \widehat{k}_3\}$ and $\{\kappa : \ \widehat{k}_3 + 1 \leq \kappa \leq (q-1)p_n\}$.

(3) For each subregion obtained in (2), we do exactly the same as that for the subregion $\{\kappa:\ 1 \le \kappa \le \widehat{k}_1\}$ or $\{\kappa:\ \widehat{k}_1 + 1 \le \kappa \le (q-1)p_n\}$ in (2), and keep doing so until there is no subregion containing any change point.

We sort the estimated change point locations in ascending order and denote them by

$$\widehat{k}_{(1)} < \widehat{k}_{(2)} < \cdots < \widehat{k}_{(\widehat{H}_{-1})},$$

where $\widehat{H}_{-1}$ is the number of change points detected. In addition, we denote $\widehat{k}_{(0)} = 0$, $\widehat{H} = \widehat{H}_{-1} + 1$, and $\widehat{k}_{(\widehat{H})} = (q-1)p_n$. We use $\widehat{H}$ to estimate H. Let

$$\widehat{\mathcal{D}}_\ell = \{(i,j):\ \widehat{k}_{(\ell-1)} < R_{ij} \le \widehat{k}_{(\ell)}\}, \quad 1 \le \ell \le \widehat{H},$$

we use $\left\{\widehat{\mathcal{D}}_\ell:\ 1 \le \ell \le \widehat{H}\right\}$ to estimate the partition $\{\mathcal{D}_\ell:\ 1 \le \ell \le H\}$. We consider all the $\beta_{i,j}$s with the subscript (i,j) in the same group of the estimated partition having the same value.

Let $\widetilde{\theta}_{ij,l}$ be the lth component of $\widetilde{\boldsymbol{\theta}}_{ij}$. Doing exactly the same to $\widetilde{\theta}_{ij,l}$, $i = 1, \cdots, p_n$, $j = 0, \cdots, q$, $l = 1, \cdots, K$, we get a partition $\{\mathcal{B}_1, \cdots, \mathcal{B}_N\}$ of $\{(i,j,l):\ i = 1, \cdots, p_n;\ j = 0, \cdots, q;\ l = 1, \cdots, K\}$. We consider all the $\theta_{ij,l}$s with the subscript (i,j,l) in the same group of the estimated partition having the same value.

Stage 3 (*Final Estimation*). Let $L(\xi_1, \cdots, \xi_{\widehat{H}}, \eta_1, \cdots, \eta_N)$ be

$$\sum_{i=1}^{p_n}\sum_{t=2}^{n}\left(y_{it} - \boldsymbol{\theta}_{i0}^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_i) - \mathbf{B}(\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_i)^{\mathrm{T}}\boldsymbol{\Theta}_i\mathbf{X}_t\right)^2,$$

with $\beta_{i,j}$, $i = 1, \cdots, p_n$, $j = 1, \cdots, q$, being replaced by ξ_k if $(i,j) \in \widehat{\mathcal{D}}_k$, and $\theta_{ij,l}$, $i = 1, \cdots, p_n$, $j = 0, \cdots, q$, $l = 1, \cdots, K$, being replaced by η_ℓ if $(i,j,l) \in \mathcal{B}_\ell$. Let $(\widehat{\xi}_1, \cdots, \widehat{\xi}_{\widehat{H}}, \widehat{\eta}_1, \cdots, \widehat{\eta}_N)$ minimise $L(\xi_1, \cdots, \xi_{\widehat{H}}, \eta_1, \cdots, \eta_N)$. The final estimator $\widehat{\beta}_{i,j}$ of $\beta_{i,j}$ is $\widehat{\xi}_k$ if $(i,j) \in \widehat{\mathcal{D}}_k$, and the final estimator $\widehat{\theta}_{ij,l}$ of $\theta_{ij,l}$ is $\widehat{\eta}_\ell$ if $(i,j,l) \in \mathcal{B}_\ell$. Once we have the estimator $\widehat{\theta}_{ij,l}$, let $\widehat{\boldsymbol{\theta}}_{ij} = (\widehat{\theta}_{ij,1}, \cdots, \widehat{\theta}_{ij,K})^{\mathrm{T}}$, the estimator $\widehat{a}_{ij}(\cdot)$ of $a_{ij}(\cdot)$ is taken to be $\mathbf{B}(\cdot)^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{ij}$.

Remark 4. When dealing with the unknown functions $a_{i,j}(\cdot)$, $i = 1, \cdots, p_n$, $j = 0, \cdots, q$, in the estimation procedure, instead of treating each unknown function as a single undivided unit to conduct homogeneity pursuit, we work on the coefficients of its B-Spline decomposition. This is because there may still be some kind of homogeneity between two functions even if they are different, e.g. some coefficients of the B-Spline decomposition of one function may be the same as some coefficients of the B-Spline decomposition of the other one, but not all the same. If we treat

each unknown function as a single undivided unit to conduct homogeneity pursuit, we would not be able to identify and use this kind of homogeneity, which would make our final estimators not as efficient as they should be.

2.2 Estimation of $\mathbf{\Sigma}_x(\cdot)$

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

$$\widehat{E}(\mathbf{X}_t|\mathbf{X}_{t-1}=\mathbf{u})=\frac{\sum\limits_{t=2}^{n}\mathbf{X}_t K_h(\|\mathbf{X}_{t-1}-\mathbf{u}\|)}{\sum\limits_{t=2}^{n}K_h(\|\mathbf{X}_{t-1}-\mathbf{u}\|)}, \tag{2.4}$$

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

This gives us the following estimator of $\mathbf{\Sigma}_x(\mathbf{u})$

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_x(\mathbf{u}) &= \widehat{E}(\mathbf{X}_t\mathbf{X}_t^{\mathrm{T}}|\mathbf{X}_{t-1}=\mathbf{u})-\widehat{E}(\mathbf{X}_t|\mathbf{X}_{t-1}=\mathbf{u})\left\{\widehat{E}(\mathbf{X}_t|\mathbf{X}_{t-1}=\mathbf{u})\right\}^{\mathrm{T}}\\ &= \{\mathrm{tr}(\mathcal{W})\}^{-2}\mathbf{X}^{\mathrm{T}}\left\{\mathrm{tr}(\mathcal{W})\mathcal{W}-\mathcal{W}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathcal{W}\right\}\mathbf{X}\end{aligned} \tag{2.5}$$

where

$$\mathbf{X}=(\mathbf{X}_1,\ \cdots,\ \mathbf{X}_n)^{\mathrm{T}},\quad \mathcal{W}=\mathrm{diag}(K_h(\|\mathbf{X}_0-\mathbf{u}\|),\ \cdots,\ K_h(\|\mathbf{X}_{n-1}-\mathbf{u}\|)),$$

h is a bandwidth, $K_h(\cdot)=K(\cdot/h)/h$, and $K(\cdot)$ is a kernel function.

The reason for us to use kernel smoothing rather than B-Spline in the estimation of $\mathbf{\Sigma}_x(\mathbf{u})$ is for simplicity, because there is no homogeneity pursuit needed here and $\mathbf{u}$ is not a scalar.

2.3 Estimation of $\mathbf{\Sigma}_{0,t}$

For each k, $k=1,\ \cdots,\ p_n$, let

$$\widehat{\epsilon}_{k,t}=y_{k,t}-\widehat{a}_{k,0}(\mathbf{X}_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_k)-\mathbf{X}_t^{\mathrm{T}}\widehat{\mathbf{a}}_k(\mathbf{X}_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_k).$$

By (1.5), we have the following synthetic GARCH model

$$\sigma_{k,t}^2=\alpha_{k,0}+\sum_{i=1}^{m}\alpha_{k,i}\widehat{\epsilon}_{k,t-i}^2+\sum_{j=1}^{s}\gamma_{k,j}\sigma_{k,t-j}^2. \tag{2.6}$$

Once $\alpha_{k,i}$s and $\gamma_{k,j}$s have been estimated, by substituting them into (2.6) and with appropriate initial values we can obtain an estimator $\widehat{\sigma}_{k,t}^2$ of $\sigma_{k,t}^2$ and hence an estimator $\widehat{\mathbf{\Sigma}}_{0,t}$ of $\mathbf{\Sigma}_{0,t}$.

For each k, $k = 1, \cdots, p_n$, we apply a quasi-maximum likelihood approach to estimate $(\alpha_{k,0}, \cdots, \alpha_{k,m}, \gamma_{k,1}, \cdots, \gamma_{k,s})$ initially. We define the negative quasi log-likelihood function as

$$\mathcal{L}_k(\alpha_{k,0}, \cdots, \alpha_{k,m}, \gamma_{k,1}, \cdots, \gamma_{k,s}) = n^{-1} \sum_{t=1}^{n} \left\{ \frac{\widehat{\epsilon}_{k,t}^2}{\sigma_{k,t}^2} + \log \sigma_{k,t}^2 \right\}, \tag{2.7}$$

where $\sigma_{k,t}^2$ are recursively defined by (2.6) with initial values being either

$$\widehat{\epsilon}_{k,0}^2 = \cdots = \widehat{\epsilon}_{k,1-m}^2 = \sigma_{k,0}^2 = \cdots = \sigma_{k,1-s}^2 = \alpha_{k,0} \tag{2.8}$$

or

$$\widehat{\epsilon}_{k,0}^2 = \cdots = \widehat{\epsilon}_{k,1-m}^2 = \sigma_{k,0}^2 = \cdots = \sigma_{k,1-s}^2 = \widehat{\epsilon}_{k,1}^2. \tag{2.9}$$

By minimising (2.7) with respect to $(\alpha_{k,0}, \cdots, \alpha_{k,m}, \gamma_{k,1}, \cdots, \gamma_{k,s})$, we take the minimiser $(\widetilde{\alpha}_{k,0}, \cdots, \widetilde{\alpha}_{k,m}, \widetilde{\gamma}_{k,1}, \cdots, \widetilde{\gamma}_{k,s})$ as an initial estimate of $(\alpha_{k,0}, \cdots, \alpha_{k,m}, \gamma_{k,1}, \cdots, \gamma_{k,s})$.

Apply exactly the same homogeneity pursuit approach, stated in Stage 2 of the estimation procedure for $\boldsymbol{\beta}_i$ in Section 2.1, to the initial estimates $\widetilde{\alpha}_{k,j}$, $k = 1, \cdots, p_n$, $j = 0, \cdots, m$, and $\widetilde{\gamma}_{k,j}$, $k = 1, \cdots, p_n$, $j = 1, \cdots, s$, respectively. Denote the resulting partition for $\widetilde{\alpha}_{k,j}$s by $\{\widehat{\mathcal{A}}_k : k = 1, \cdots, \widehat{\varsigma}\}$, for $\widetilde{\gamma}_{k,j}$s by $\{\widehat{\Gamma}_k : k = 1, \cdots, \widehat{\tau}\}$.

Let $\mathcal{L}(\mu_1, \cdots, \mu_{\widehat{\varsigma}}, \nu_1, \cdots, \nu_{\widehat{\tau}})$ be

$$\sum_{k=1}^{p_n} \mathcal{L}_k(\alpha_{k,0}, \cdots, \alpha_{k,m}, \gamma_{k,1}, \cdots, \gamma_{k,s}),$$

with $\alpha_{i,j}$, $i = 1, \cdots, p_n$, $j = 0, \cdots, m$, being replaced by μ_k if $(i,j) \in \widehat{\mathcal{A}}_k$, and $\gamma_{i,j}$, $i = 1, \cdots, p_n$, $j = 1, \cdots, s$, being replaced by ν_ℓ if $(i,j) \in \widehat{\Gamma}_\ell$. Let $(\widehat{\mu}_1, \cdots, \widehat{\mu}_{\widehat{\varsigma}}, \widehat{\nu}_1, \cdots, \widehat{\nu}_{\widehat{\tau}})$ minimise $\mathcal{L}(\mu_1, \cdots, \mu_{\widehat{\varsigma}}, \nu_1, \cdots, \nu_{\widehat{\tau}})$. The final estimator $\widehat{\alpha}_{i,j}$ of $\alpha_{i,j}$ is $\widehat{\mu}_k$ if $(i,j) \in \widehat{\mathcal{A}}_k$, and the final estimator $\widehat{\gamma}_{i,j}$ of $\gamma_{i,j}$ is $\widehat{\nu}_\ell$ if $(i,j) \in \widehat{\Gamma}_\ell$.

2.4 Computational algorithm

In the estimation procedure described in Section 2.1, the minimiser of (2.3) does not have a closed form, neither does the minimiser of $L(\xi_1, \cdots, \xi_{\widehat{H}}, \eta_1, \cdots, \eta_N)$. To conduct the minimisation of either of the two objective functions, we appeal to the standard NLS algorithm, and use the `nls` of R to implement it. One can also use other NLS software, for example, the NLS routine `lsqnonlin()` from MATLAB and `PROC NLIN` from SAS. To use the `nls` of R, we first need to find an initial value. The initial value for minimising (2.3) can be obtained as follows:

(1) Apply the standard least squares estimation for the linear models to $(y_{it}, \mathbf{X}_t)$, $t = 1, \cdots, n$, and denote the resulting estimator by $\check{\boldsymbol{\beta}}_i$, the initial value for $\boldsymbol{\beta}_i$ is taken to be $\boldsymbol{\beta}_i^{(0)} = \check{\boldsymbol{\beta}}_i / \|\check{\boldsymbol{\beta}}_i\|$ if the first component of $\check{\boldsymbol{\beta}}_i$, $\check{\beta}_{i1}$, is positive, $\boldsymbol{\beta}_i^{(0)} = -\check{\boldsymbol{\beta}}_i / \|\check{\boldsymbol{\beta}}_i\|$ otherwise.

(2) Substitute $\boldsymbol{\beta}_i^{(0)}$ for $\boldsymbol{\beta}_i$ in (2.3), then minimise (2.3) with respect to $(\boldsymbol{\Theta}_i,\ \boldsymbol{\theta}_{i0})$, the minimiser $(\boldsymbol{\Theta}_i^{(0)},\ \boldsymbol{\theta}_{i0}^{(0)})$ is the initial value of $(\boldsymbol{\Theta}_i,\ \boldsymbol{\theta}_{i0})$.

Once we have $\boldsymbol{\beta}_i^{(0)}$, $\boldsymbol{\Theta}_i^{(0)}$ and $\boldsymbol{\theta}_{i0}^{(0)}$, the minimiser of (2.3) can be obtained by the `nls` of R straightforwardly.

For any set A, let $|A|$ be the number of elements in A. The initial value for minimising $L(\xi_1,\ \cdots,\ \xi_{\widehat{H}},\ \eta_1,\ \cdots,\ \eta_N)$ can be obtained through the initial estimates of $\boldsymbol{\beta}_i$ and $\boldsymbol{\theta}_i$, obtained in Stage 1 of the estimation procedure in Section 2.1, as follows:

$$\xi_k^{(0)} = \left(|\widehat{\mathcal{D}}_k|\right)^{-1} \sum_{(i,j)\in\widehat{\mathcal{D}}_k} \widetilde{\beta}_{ij}, \quad k = 1,\ \cdots,\ \widehat{H}$$

and

$$\eta_k^{(0)} = (|\mathcal{B}_k|)^{-1} \sum_{(i,j,l)\in\mathcal{B}_k} \widetilde{\theta}_{ij,l}, \quad k = 1,\ \cdots,\ N.$$

Once we have the initial value $(\xi_1^{(0)},\ \cdots,\ \xi_{\widehat{H}}^{(0)},\ \eta_1^{(0)},\ \cdots,\ \eta_N^{(0)})$, we can have the minimiser of $L(\xi_1,\ \cdots,\ \xi_{\widehat{H}},\ \eta_1,\ \cdots,\ \eta_N)$ by using the `nls` of R straightforwardly.

3 Asymptotic properties

We start with our definitions of the *overfitting*, *correct fitting*, and *misspecification* in this paper:

- *Overfitting*: the underlying homogeneity structure is ignored and the homogeneity pursuit is not conducted when $\widehat{\text{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ being constructed, namely $\widehat{\text{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ is constructed based on (2.1) and the initial estimators of $\boldsymbol{\beta}_i$s, $\mathbf{a}_i(\cdot)$s, $\alpha_{i,k}$s and $\gamma_{i,k}$s obtained in Stage 1 of the proposed estimation procedure in Section 2.1.

- *Correct fitting (which is our modelling together with the proposed estimation)*: the underlying homogeneity structure is appreciated and the homogeneity pursuit is conducted when $\widehat{\text{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ being constructed, namely $\widehat{\text{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ is constructed based on (2.1) and the final estimators of $\boldsymbol{\beta}_i$s, $\mathbf{a}_i(\cdot)$s, $\alpha_{i,k}$s and $\gamma_{i,k}$s in Section 2.1.

- *Misspecification*: $\widehat{\text{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ is constructed based on (2.1), and the estimators of $\boldsymbol{\beta}_i$s, $\mathbf{a}_i(\cdot)$s, $\alpha_{i,k}$s and $\gamma_{i,k}$s in (2.1) are obtained under the false assumption that $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_{p_n}$, by exactly the same estimation procedure stated in Section 2 but setting $\widehat{H} = q - 1$ and $\widehat{\mathcal{D}}_k = \{(i,\ k) :\ 1 \le i \le p_n\}$ in Stage 3 of the estimation for $\boldsymbol{\beta}_i$s and $\mathbf{a}_i(\cdot)$s in Section 2.1.

In this Section, we are going to discuss the asymptotic properties of the estimator of $\text{cov}(\mathbf{Y}_t|\mathcal{F}_{t-1})$ obtained by either *overfitting*, *correct fitting*, or *misspecification*.

To measure the accuracy of an estimator $\widehat{\mathbf{M}}$ of a positive definite matrix $\mathbf{M}$ of dimension $p_n \times p_n$, we use the entropy loss norm:

$$\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\mathbf{\Sigma}} = p_n^{-1/2}\|\mathbf{M}^{-1/2}(\widehat{\mathbf{M}} - \mathbf{M})\mathbf{M}^{-1/2}\|.$$

In the following, we are going to present three theorems to show theoretically the superiority of our modelling and the proposed estimation.

The relatively long list of assumptions is contained in the Appendix. However, it is worthwhile to mention here that we assume the number of groups H, $\mathcal{N}$, ς, τ are fixed, while $p_n = O(n^b)$ for some $b < 1$ (see as sumption (C13) for the exact constraints required on p_n). Furthermore, the sequence $\mathbf{Y}_t$ is assumed to be stationary and α-mixing while we allow some weak dependence across the components $i = 1, \cdots, p_n$.

Theorem 1 *(Overfitting case) Assume the number of spline basis function used is $K_1 \asymp n^{1/5}$ with bandwidth $h \asymp n^{1/(4+q)}$. For the convergence rate of $\widehat{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ constructed by overfitting, we have, under conditions (C1)-(C13) in the Appendix,*

$$\begin{aligned} & \|\widehat{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n) - cov(\mathbf{Y}_{n+1}|\mathcal{F}_n)\|^2_{\mathbf{\Sigma}} \\ \leq \;& Cp_n n^{-8/5}\log^2 n + Cn^{-4/5}\log n + Cp_n^{-1}n^{-4/(4+q)}\log n, \end{aligned}$$

with probability approaching one. When $p_n = O(n^{\frac{4(q-1)}{5(q+4)}})$, the 3rd term above is dominated by the second term and we have

$$\|\widehat{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n) - cov(\mathbf{Y}_{n+1}|\mathcal{F}_n)\|^2_{\mathbf{\Sigma}} \leq Cp_n n^{-8/5}\log^2 n + Cn^{-4/5}\log n,$$

with probability approaching one.

Theorem 2 *(Correct fitting case) Assume the number of spline basis function used is $K_2 \asymp (p_n n)^{1/5}$ with bandwidth $h \asymp n^{1/(4+q)}$. For the convergence rate of $\widehat{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ constructed by correct fitting, we have, under conditions (C1)-(C13) in the Appendix,*

$$\begin{aligned} & \|\widehat{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n) - cov(\mathbf{Y}_{n+1}|\mathcal{F}_n)\|^2_{\mathbf{\Sigma}} \\ \leq \;& Cp_n(p_n n)^{-8/5}\log^2 n + C(p_n n)^{-4/5}\log n + Cp_n^{-1}n^{-4/(4+q)}\log n, \end{aligned}$$

with probability approaching one. When $p_n = O(n^{\frac{4(q-1)}{q+4}})$, the 3rd term above is dominated by the second term and we have

$$\|\widehat{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n) - cov(\mathbf{Y}_{n+1}|\mathcal{F}_n)\|^2_{\mathbf{\Sigma}} \leq Cp_n(p_n n)^{-8/5}\log^2 n + C(p_n n)^{-4/5}\log n,$$

with probability approaching one.

Theorem 3 *(Misspecification case) Under assumptions (C1)-(C16) in the Appendix, when* $\widehat{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ *is constructed by misspecification, we have*

$$\|\widehat{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n) - cov(\mathbf{Y}_{n+1}|\mathcal{F}_n)\|_{\mathbf{\Sigma}}^2 \geq Cp_n,$$

with probability bounded away from zero.

Theorems 1 and 2 show the error bar of the proposed estimator, which is the correct fitting case, is of a higher order of convergence rate than that of the estimator obtained without using the underlying homogeneity structure, therefore, the proposed estimator is more accurate. Theorem 3 shows the estimator, obtained under the assumption that there is no heterogeneity on the index $\boldsymbol{\beta}$ among the components of $\mathbf{Y}_t$, would not be consistent when heterogeneity exists.

In conclusion, theorems 1 - 3 show: (1) to assume every component of $\mathbf{Y}_t$ share the same index $\boldsymbol{\beta}$, when this is not the case, would result in a serious problem in the final estimator; (2) to ignore the homogeneity structure would lead to an inefficient final estimator, which is not good either; (3) our modelling together with the proposed estimation takes into account both homogeneity and heterogeneity, and therefore, results in the best estimator. Our simulation results in Section 5 will tell the same story.

4 Portfolio allocation

In this section, we are going to demonstrate how to apply the proposed dynamic covariance matrices to portfolio allocation, therefore, $\mathbf{Y}_t$ here is the vector of returns at time point t of assets concerned.

Our portfolio allocation is based on the mean-variance optimal portfolio proposed by Markowitz (1952, 1959). Specifically, we assume the conditional covariance matrix $\text{cov}(\mathbf{Y}_t|\mathcal{F}_{t-1})$ involved in the Markowitz's formula enjoys the proposed dynamic structure and estimate it by the proposed estimation procedure. Except $\text{cov}(\mathbf{Y}_t|\mathcal{F}_{t-1})$, in order to form a portfolio allocation based on the Markowitz's formula, we also need the conditional expectation $E(\mathbf{Y}_t|\mathcal{F}_{t-1})$, we therefore estimate it first.

By taking conditional expectation of (1.2), we have

$$E(y_{i,t}|\mathcal{F}_{t-1}) = a_{i0}(\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_i) + \mathbf{a}_i(\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_i)E(\mathbf{X}_t|\mathbf{X}_{t-1})$$

which leads to the following estimator of $E(y_{i,t}|\mathcal{F}_{t-1})$

$$\widehat{E}(y_{i,t}|\mathcal{F}_{t-1}) = \widehat{a}_{i0}(\mathbf{X}_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_i) + \widehat{\mathbf{a}}_i(\mathbf{X}_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_i)\widehat{E}(\mathbf{X}_t|\mathbf{X}_{t-1}) \tag{4.1}$$

where $\widehat{E}(\mathbf{X}_t|\mathbf{X}_{t-1})$ is defined in (2.4), and $\widehat{a}_{i0}(\cdot)$, $\widehat{\mathbf{a}}_i(\cdot)$ and $\widehat{\boldsymbol{\beta}}_i$ are the final estimators of $a_{i0}(\cdot)$, $\mathbf{a}_i(\cdot)$ and $\boldsymbol{\beta}_i$ obtained in Section 2.1.

Based on the Markowitz's formula, we define the vector of our portfolio weights of p_n risky assets, to be held between times $t-1$ and t, by

$$\widehat{\mathbf{w}}_{t-1} = \frac{c_3 - c_2\zeta}{c_1c_3 - c_2^2}\widehat{\mathrm{cov}}(\mathbf{Y}_t|\mathcal{F}_{t-1})^{-1}\mathbf{1}_{p_n} + \frac{c_1\zeta - c_2}{c_1c_3 - c_2^2}\widehat{\mathrm{cov}}(\mathbf{Y}_t|\mathcal{F}_{t-1})^{-1}\widehat{E}(\mathbf{Y}_t|\mathcal{F}_{t-1}), \tag{4.2}$$

where

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

ζ is the target return imposed on the portfolio.

5 Simulation study

In this section, we are going to use simulated examples to show how well the proposed estimation procedure for $\widehat{\mathrm{cov}}(\mathbf{Y}_t|\mathcal{F}_{t-1})$ works, and how well $\widehat{\boldsymbol{\Sigma}}_{0,t}$ works when applied to estimate $\boldsymbol{\Sigma}_{0,t}$. We will also compare the accuracy of $\widehat{\mathrm{cov}}(\mathbf{Y}_t|\mathcal{F}_{t-1})$s (and $\widehat{\boldsymbol{\Sigma}}_{0,t}$s) obtained by several competing methods, respectively, and the returns of the portfolios formed based on (4.2) with $\widehat{\mathrm{cov}}(\mathbf{Y}_t|\mathcal{F}_{t-1})$s obtained by competing methods.

5.1 Simulation Setting

We set $q = 4$, and $(n,\ p_n)$ to be either $(300,\ 120)$, $(500,\ 300)$, or $(1250,\ 1000)$. For each scenario, we do 100 simulations, and for each simulation, we generate data as follows:

We first generate $\mathbf{X}_t \in \mathbb{R}^4$, $t = 1,\ \cdots,\ n+1$, independently from an uniform distribution on $[-1,\ 1]^4$, then for each i, $i = 1,\ \cdots,\ p_n$, independently generate $\epsilon_{i,t}$, $t = 1,\ \cdots,\ n+1$, from the GARCH model (1.5) with

$$m = 1, \quad s = 1, \quad \alpha_{i,0} = 0.1,$$

and

$$\alpha_{i,1} = \begin{cases} 0.3 & \text{when } i \text{ is odd,} \\ 0.7 & \text{when } i \text{ is even,} \end{cases} \quad \text{and} \quad \gamma_{i,1} = \begin{cases} 0.6 & \text{when } i \text{ is odd,} \\ 0.2 & \text{when } i \text{ is even.} \end{cases}$$

After $\mathbf{X}_t$s and $\epsilon_{i,t}$s are generated, we generate $\mathbf{Y}_t = (y_{1,t},\ \cdots,\ y_{p_n,t})^{\mathrm{T}}$, $t = 1,\ \cdots,\ n+1$, based on model (1.2) with

$$\boldsymbol{\beta}_i = \begin{cases} (0.5,\ -0.5,\ 0.5,\ -0.5)^{\mathrm{T}} & \text{when } i \text{ is odd,} \\ (0.5,\ 0.5,\ -0.5,\ -0.5)^{\mathrm{T}} & \text{when } i \text{ is even,} \end{cases}$$

and

$$a_{i,0}(u) = \ln(1+u^2), \quad \mathbf{a}_i(u) = \begin{cases} \big(u,\ \cos(\pi u),\ u,\ 2-3\exp(-u^2)\big)^{\mathrm{T}} & \text{when } i \text{ is odd,} \\ \big(\cos(\pi u),\ u,\ 2-3\exp(-u^2),\ u\big)^{\mathrm{T}} & \text{when } i \text{ is even.} \end{cases}$$

Throughout this section, we use a B-Spline basis of order 3 with 10 equally spaced knots to approximate the unknown functions. The threshold δ in homogeneity pursuit is selected by BIC. The kernel function in the estimation of $\Sigma_x(\cdot)$ is taken to be the Epanechnikov kernel $K(u) = 0.75(1-u^2)_+$, and the bandwidth h is selected by 5-fold cross-validation.

5.2 Estimation of dynamic covariance matrix

Denote the proposed dynamic covariance matrix estimation method as Our Method. We compare Our Method with the Overfitting Method and the Misspecification Method defined in Section 3. Further, we include the DCC-NL Method*, proposed by Engle *et al.*(2019), in the comparison. In this subsection, we use $\{\mathbf{Y}_t,\ \mathbf{X}_t\}_{t=1}^n$ as the training sample to estimate the dynamic covariance matrix $\mathrm{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ as well as the residual covariance matrix $\mathbf{\Sigma}_{0,n+1}$ by each one of the above four methods. The estimation accuracy is measure by the scaled spectrum norms, which are defined as

$$p_n^{-1/2}\|\widehat{\mathrm{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n) - \mathrm{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)\| \quad \text{and} \quad p_n^{-1/2}\|\widehat{\mathbf{\Sigma}}_{0,n+1} - \mathbf{\Sigma}_{0,n+1}\|$$

where $\widehat{\mathrm{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ and $\widehat{\mathbf{\Sigma}}_{0,n+1}$ are estimates of $\mathrm{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ and $\mathbf{\Sigma}_{0,n+1}$ obtained by one of the above four methods, respectively. The sample mean and sample standard deviation of the scaled spectrum norm over 100 simulations for each method and each pair of $(n,\ p_n)$ are presented in Tables 1 and 2.

Tables 1 and 2 show Our Method performs best, and the Misspecification Method is the worst. The performance of the DCC-NL Method and the Overfitting Method are comparable since they both are tailored for dynamic covariance matrix estimation without homogeneity pursuit. This suggests that it is absolutely necessary to take into account both homogeneity and heterogeneity in the estimation of dynamic covariance matrices, which is in line with the asymptotic results in Section 3.

5.3 Performance of portfolio allocation

In this subsection, we are going to examine the performance of the portfolio allocation introduced in Section 4, and investigate the implication for the resulting portfolio allocation of Our Method,

*The DCC-NL package is available at `https://www.econ.uzh.ch/en/people/faculty/wolf/publications.html`

Table 1: Mean and Standard Deviation of: $p_n^{-1/2}\|\widehat{\text{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n) - \text{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)\|$

	Method	$n = 300$ $p_n = 120$	$n = 500$ $p_n = 300$	$n = 1250$ $p_n = 1000$
Mean	Our	0.2046	0.1983	0.1741
	DCC-NL	0.2551	0.2278	0.2055
	Overfitting	0.2998	0.2568	0.2158
	Misspecification	0.4178	0.3738	0.3612
SD	Our	0.0922	0.1154	0.1083
	DCC-NL	0.1508	0.1439	0.1201
	Overfitting	0.2822	0.2546	0.2144
	Misspecification	0.6510	0.4054	0.3406

Mean and SD stand for the sample mean and sample standard deviation of $p_n^{-1/2}\|\widehat{\text{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n) - \text{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)\|$ *over 100 simulations.*

Table 2: Mean and Standard Deviation of: $p_n^{-1/2}\|\widehat{\mathbf{\Sigma}}_{0,n+1} - \mathbf{\Sigma}_{0,n+1}\|$

	Method	$n = 300$ $p_n = 120$	$n = 500$ $p_n = 300$	$n = 1250$ $p_n = 1000$
Mean	Our	0.0762	0.0758	0.0724
	DCC-NL	0.0795	0.0802	0.0741
	Overfitting	0.1021	0.1035	0.0895
	Misspecification	0.1505	0.1476	0.1253
SD	Our	0.0922	0.0877	0.0520
	DCC-NL	0.1421	0.1350	0.1182
	Overfitting	0.1854	0.1756	0.1574
	Misspecification	0.3082	0.2561	0.2266

Mean and SD stand for the sample mean and sample standard deviation of $p_n^{-1/2}\|\widehat{\mathbf{\Sigma}}_{0,n+1} - \mathbf{\Sigma}_{0,n+1}\|$ *over 100 simulations.*

DCC-NL method, Overfitting Method and Misspecification Method in the estimation of the covariance matrix involved. We set the target return of every portfolio in this subsection to be 1%.

We form a portfolio allocation $\widehat{\mathbf{w}}_n$ at time point n based on the training sample $\{\mathbf{Y}_t,\ \mathbf{X}_t\}_{t=1}^n$. Specifically, $\widehat{\mathbf{w}}_n$ is formed by the formula (4.2) with $t-1$ being replaced by n, $\widehat{\text{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ being obtained by either of the four methods stated above and $\widehat{E}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ by formula (4.1). The return

yielded by the portfolio $\widehat{\mathbf{w}}_n$ at time point $n+1$ is

$$R(\widehat{\mathbf{w}}_n) = \widehat{\mathbf{w}}_n^{\mathrm{T}} \mathbf{Y}_{n+1}.$$

For each setting of $(n,\ p_n)$ in Subsection 5.1, we compute the sample mean (denoted by $\mathrm{Mean}\{R(\widehat{\mathbf{w}}_n)\}$) and sample standard deviation (denoted by $SD\{R(\widehat{\mathbf{w}}_n)\}$) of return, over the 100 simulations conducted, for each portfolio formed, and present them in Table 3. For each case, we also compute the Sharpe ratio, which is defined as

$$SR(\widehat{\mathbf{w}}_n) = \frac{\mathrm{Mean}\{R(\widehat{\mathbf{w}}_n\}}{SD\{R(\widehat{\mathbf{w}}_n)\}},$$

and still present it in Table 3.

Table 3: The Performance of Each Portfolio Allocation

	Method	$n = 300$ $p_n = 120$	$n = 500$ $p_n = 300$	$n = 1250$ $p_n = 1000$
$\mathrm{Mean}\{R(\widehat{\mathbf{w}}_n)\}$	Our	0.97	0.98	0.96
	DCC-NL	0.88	0.89	0.88
	Overfitting	0.83	0.86	0.88
	Misspecification	0.80	0.79	0.77
$SD\{R(\widehat{\mathbf{w}}_n)\}$	Our	0.52	0.56	0.58
	DCC-NL	0.65	0.60	0.64
	Overfitting	0.79	0.83	0.86
	Misspecification	0.92	0.99	0.98
$SR(\widehat{\mathbf{w}}_n)$	Our	1.87	1.75	1.66
	DCC-NL	1.35	1.48	1.38
	Overfitting	1.05	1.03	1.02
	Misspecification	0.87	0.79	0.78

Method column indicates the estimation methods of $\widehat{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ upon which the portfolios is formed. The entries corresponding to Mean$\{R(\widehat{\mathbf{w}}_n)\}$ are in percentage.

Table 3 shows the portfolio formed based on the $\widehat{\mathrm{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ obtained by Our Method performs best, with both sample mean and Sharpe ratio of the yielded return highest and sample standard deviation lowest for every case. This attributes to the homogeneity pursuit in Our Method, which significantly reduces the number of unknowns. The DCC-NL method is the runner up in the horse racing as it has the second-highest mean and the second-lowest standard deviation in every case. Again, the Misspecification method performs the worst in all cases.

The simulation results in this subsection shows the estimation of $\text{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ plays a key role in portfolio allocation, ignoring either homogeneity or heterogeneity in the modelling of $\text{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ would have serious consequences, which will eventually be hit on the yielded return.

6 Real data analysis

In this section, we are going to apply the proposed portfolio allocation (denoted by OUR), i.e. the portfolio formed by (4.2) with $\widehat{\text{cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ being obtained by correct fitting, to the 49 Industry Portfolios dataset[†], which has been analysed by Guo *et al.*(2017), and compare the yielded return with that of the portfolios formed by other methods. Specifically, the portfolio allocations under comparison are listed as follows:

(1) DCC-NL: The portfolio allocation based on the dynamic covariance matrix estimator proposed in Engle *et al.*(2019).

(2) FACE: The portfolio allocation based on the dynamic covariance matrix estimator proposed in Guo *et al.*(2017).

(3) FAN: The portfolio allocation based on the covariance matrix estimator proposed in Fan *et al.*(2008).

(4) SAM: The portfolio allocation based on the sample covariance matrix.

(5) MARKET: The market portfolio allocation. This is used as a benchmark.

We remark that the portfolio allocations in OUR, DCC-NL, FACE, FAN and SAM are all constructed out of the Markowitz's formula. The difference between them lies in the way to estimate the covariance matrix of return. Neither SAM nor FAN takes into account the dynamic feature of the covariance matrix in their estimation. Although DCC-NL and FACE estimate dynamic covariance matrices, they ignore the underlying homogeneous structure in $\mathbf{a}_i(\cdot)$s and heterogeneity on the index $\boldsymbol{\beta}$. It is worth mentioning that, OUR is the only one that uses a covariance matrix, in the modelling and estimation of which, both homogeneity and heterogeneity are taken into account.

We now give a brief description about the dataset: the response variable $\mathbf{Y}_t = (y_{1,t},\ \cdots,\ y_{49,t})^{\mathrm{T}}$ is the vector of daily returns of the 49 industry portfolios (value weighted) excess the risk-free rate. The observable factors $\mathbf{X}_t = (x_{1,t},\ x_{2,t},\ x_{3,t})^{\mathrm{T}}$ are the market, size and value factors in the Fama-French three-factor models. We refer to Guo *et al.*(2017) for more descriptive details of the dataset.

[†]The dataset is maintained by the Kenneth French's data library which is public available at `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

For example, the labeling along with a brief description of $\mathbf{Y}_t$ and $\mathbf{X}_t$ can be found in Tables 4 and 5 in Guo *et al.*(2017), respectively.

We assess the performance of a portfolio allocation with a year by year back-testing from 1995 to 2014. Throughout this section, in the implementation of OUR method, we use a B-Spline basis of order 3 with 10 equally spaced knots in the decomposition of unknown function $\mathbf{a}_i(\cdot)$, $i = 1, \cdots, 49$, the threshold δ in the homogeneity pursuit is selected by BIC, the kernel function involved is chosen as the Epanechnikov kernel, and the bandwidth is selected by 5-fold cross-validation.

To highlight the role of covariance matrix estimation, we adopt a simplified back-testing setup: we assume no transaction cost, allow for short selling and assume that all possible portfolio allocations are attainable. In each year, we start with an initial fund of 100 pounds and trade on a daily basis. The trading strategy consists of forming a portfolio allocation at the end of each trading day and holding it until the end of the next trading day. Between day $t-1$ and day t, we obtain the portfolio return

$$R(\widehat{\mathbf{w}}_{t-1}) = \widehat{\mathbf{w}}_{t-1}^{\mathrm{T}} \mathbf{Y}_t,$$

where $\widehat{\mathbf{w}}_{t-1}$ is obtained by the historical data of lag n, i.e. $(\mathbf{Y}_{t-j},\ \mathbf{X}_{t-j})$, $j = 1, \cdots, n$. In a year of T trading days, we calculate the annualized Sharpe ratio as

$$SR = \frac{\bar{R}}{SD(R)}\sqrt{T},$$

where

$$\bar{R} = \frac{1}{T}\sum_{t=1}^{T} R(\widehat{\mathbf{w}}_{t-1}) \quad \text{and} \quad SD(R) = \left[\frac{1}{T}\sum_{t=1}^{T}\{R(\widehat{\mathbf{w}}_{t-1}) - \bar{R}\}^2\right]^{1/2}. \tag{6.1}$$

For each year between 1995 and 2014, and each of the five portfolio allocations, we compute the balance at the end of the final trading day of the year and the annualized Sharpe ratio. We repeat this using $n = 100$ and 300, respectively. The end of year balances are presented in Table 4. In all years except 1998, OUR portfolio allocation produces the highest end of the year balance. In addition, OUR portfolio allocation is the only one that beats the market portfolio and gains positive profit (e.g. end of year balance > 100) in every year.

The annualized Sharpe ratios are presented in Figure 1. Again, OUR portfolio allocation has the highest Sharpe ratio in almost all scenarios. We also applied a one-tailed robust Sharpe ratio test[‡] (Ledoit and Wolf, 2008) to test the null hypothesis that the Sharpe ratio of OUR portfolio is no less than the Sharpe ratio of DCC-NL portfolio, versus the alternative hypothesis that the Sharpe

[‡]The robust Sharpe ratio test is implemented by the R code available at `https://www.econ.uzh.ch/en/people/faculty/wolf/publications.html`

ratio of OUR portfolio is less than the Sharpe ratio of DCC-NL portfolio. With a pre-specified significance level $\alpha = 0.05$, the testing results do not reject the null hypothesis.

Further, we choose the MARKET portfolio as the benchmark and report the information ratio for OUR, DCC-NL and FACE, which are three portfolios that can consistently outperform the MARKET portfolio. The information ratio has been widely studied to compare the performance of two portfolios, see Haugen and Baker (1991), Jagannathan and Ma (2003), Nielsen and Aylursubramanian (2008), among others. The information ratio is defined as

$$IR = \frac{\bar{R} - \bar{R}_M}{SD(R - R_M)},$$

where $R = R(\widehat{\mathbf{w}})$ is the return obtained by OUR, DCC-NL or FACE, and R_M is the return of MARKET portfolio. Further, $\bar{R}$, $\bar{R}_M$ and $SD(R - R_M)$ are defined similar as in (6.1). The results are presented in Figure 2. The results show that OUR portfolio has the highest information ratio in all scenarios. This highlights the effectiveness of OUR portfolio allocation in improving the balance as well as reducing the risk.

We would like to declare that the results obtained in this section are based on a simplified back-testing setup, and hence they should not be used as a guideline for any real-world investment activities.

7 Conclusion

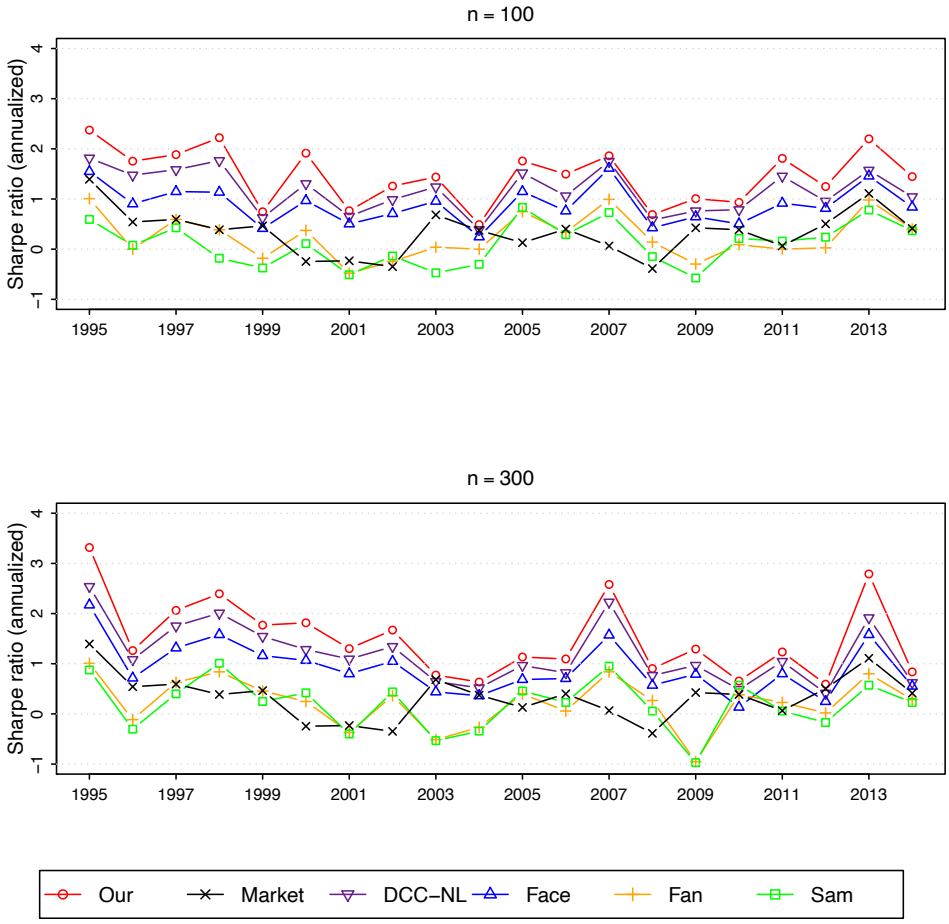
In this paper, we introduce a new dynamic structure for high dimensional covariance matrices, which combines common risk factors, single index varying coefficient modelling, and homogeneity pursuit to make the proposed structure more realistic, flexible but also parsimonious. Estimation procedure for the covariance matrices with the proposed dynamic structure is also established. The advantage of the proposed dynamic structure over that in Guo *et al.*(2017) is significant and has been clearly demonstrated by both simulation studies and empirical applications. The proposed dynamic structure is also different from the GARCH model based dynamic structure for high dimensional covariance matrices. Empirical applications show the proposed dynamic structure together with the proposed estimation procedure is superior to many existing methods.

Table 4: Balances of Trading Strategies

Year	Market	$n = 100$					$n = 300$				
		Our	DCC-NL	Face	Fan	Sam	Our	DCC-NL	Face	Fan	Sam
1995	137	278	260	224	164	216	**563**	543	541	277	347
1996	121	177	165	159	101	96	**194**	173	184	56	72
1997	131	216	194	179	138	155	**331**	304	303	146	207
1998	124	196	206	178	79	134	357	**364**	317	330	299
1999	126	144	137	121	61	78	**285**	276	260	117	175
2000	88	204	197	176	102	133	**272**	261	253	155	120
2001	89	151	140	129	53	60	**186**	141	167	49	49
2002	79	187	158	164	73	69	**239**	229	222	150	142
2003	132	**195**	178	161	57	97	150	166	134	40	45
2004	112	125	128	112	67	95	**143**	112	132	55	56
2005	106	**211**	202	179	194	166	205	193	184	157	151
2006	115	169	151	149	119	121	**206**	160	184	114	95
2007	106	273	261	233	185	231	**404**	381	376	305	217
2008	63	165	143	143	73	104	**216**	206	203	79	114
2009	128	172	167	147	48	66	**207**	196	188	9	5
2010	117	**158**	111	129	109	100	113	115	107	169	148
2011	100	**217**	214	177	107	93	209	201	192	88	120
2012	116	**198**	189	158	117	96	132	127	122	60	83
2013	135	294	224	232	200	226	**437**	390	412	180	275
2014	112	**194**	166	158	133	134	161	150	152	114	131

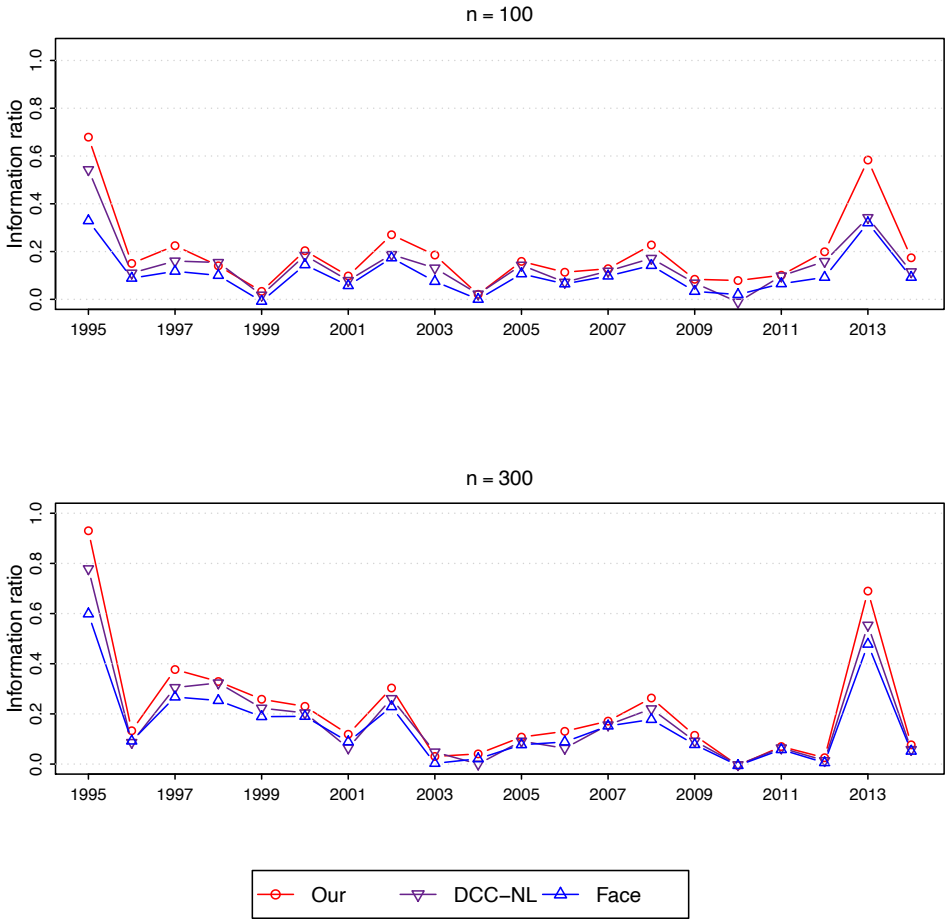
In this table, the first two columns show the year and the balance on the final trading day when investing in the market portfolio. The balances on the final trading day for Our, Face, Fan and Sam are grouped according to $n = 100$ *(columns 3-6),* $n = 300$ *(columns 7-10) and* $n = 500$ *(columns 11-14).* For each row, the bold figure indicates the highest balance of the year.

Figure 1: Annualized Sharpe Ratios



This figure shows the performance of the five portfolio allocations (Our, Market, DCC-NL, Face, Fan and Sam) in terms of the annualized Sharpe ratio, using different sample sizes $n = 100$ *and* $n = 300$.

Figure 2: Information Ratios



This figure shows the performance of the three portfolio allocations (Our, DCC-NL and Face) in terms of the annualized Sharpe ratio with respect to the market portfolio, using different sample sizes $n = 100$ *and* $n = 300$.

Appendix A. Assumptions and notations

Let $\overline{\mathbf{X}}_t = (1, \mathbf{X}_t^{\mathrm{T}})^{\mathrm{T}}$, $\overline{\mathbf{a}}_i = (a_{i,0}, \mathbf{a}_i^{\mathrm{T}})^{\mathrm{T}}$. Due to the unit norm constraint for $\boldsymbol{\beta}_i$, its first component $\beta_{i,1}$ is a function of the remaining components $\boldsymbol{\beta}_i^{(-1)} = (\beta_{i,2}, \ldots, \beta_{i,q})^{\mathrm{T}}$. The Jacobian of the

transformation from $\boldsymbol{\beta}_i^{(-1)}$ to $\boldsymbol{\beta}_i$ is given by

$$\mathbf{M}_i = \frac{\partial \boldsymbol{\beta}_i}{\partial \boldsymbol{\beta}_i^{(-1)}} = \left(\begin{array}{c} -\frac{\boldsymbol{\beta}_i^{(-1)\mathrm{T}}}{(1-\|\boldsymbol{\beta}_i^{(-1)}\|^2)^{1/2}} \\ \mathbf{I}_{q-1} \end{array} \right) \Bigg|_{\boldsymbol{\beta}_i=\boldsymbol{\beta}_{0i}}.$$

Here and below we use subscript 0 to indicate the true value (when using multiple subscripts the true value is indicated by the zero appearing in the leading position). Let $\mathcal{M}_i = \{m : m(\mathbf{x}_1, \mathbf{x}_2) = \overline{\mathbf{x}}_1^{\mathrm{T}} \overline{\mathbf{b}}(\mathbf{x}_2^{\mathrm{T}} \boldsymbol{\beta}_{0i}) : E[m^2(\mathbf{X}_t, \mathbf{X}_{t-1})] < \infty, \overline{\mathbf{x}}_1 = (1, \mathbf{x}_1^{\mathrm{T}})^{\mathrm{T}} = (1, x_{11}, \ldots, x_{1q})^{\mathrm{T}}, \overline{\mathbf{b}}(.) = (b_0(.), \ldots, b_q(.))^{\mathrm{T}}\}$ be the space of functions taking the same form as for our mean model. Let $\overline{\mathbf{a}}_i' = (a_{i,0}', \ldots, a_{i,q}')^{\mathrm{T}}$ be the first derivative of the functions $a_{i,j}$. Define the projection of the random vector $\overline{\mathbf{X}}_t^{\mathrm{T}} \overline{\mathbf{a}}_i'(\mathbf{X}_{t-1}^{\mathrm{T}} \boldsymbol{\beta}_{0i}) \mathbf{X}_{t-1}$ on $\mathcal{M}_i$ as

$$E_{\mathcal{M}_i}[\overline{\mathbf{X}}_t^{\mathrm{T}} \overline{\mathbf{a}}_i'(\mathbf{X}_{t-1}^{\mathrm{T}} \boldsymbol{\beta}_{0i}) \mathbf{X}_{t-1}] = \mathbf{m}(\mathbf{X}_t, \mathbf{X}_{t-1}) = (m_1(\mathbf{X}_t, \mathbf{X}_{t-1}), \ldots, m_q(\mathbf{X}_t, \mathbf{X}_{t-1}))^{\mathrm{T}},$$

where $m_1, \ldots, m_q$ is the minimizer of

$$\min_{m_1, \ldots, m_q \in \mathcal{M}_i} E[\|\overline{\mathbf{X}}_t^{\mathrm{T}} \overline{\mathbf{a}}_i'(\mathbf{X}_{t-1}^{\mathrm{T}} \boldsymbol{\beta}_{0i}) \mathbf{X}_{t-1} - (m_1(\mathbf{X}_t, \mathbf{X}_{t-1}), \ldots, m_q(\mathbf{X}_t, \mathbf{X}_{t-1}))^{\mathrm{T}}\|^2]. \tag{A.1}$$

We impose the following assumptions.

(C1) $(y_t, \mathbf{X}_t, \boldsymbol{\epsilon}_t), t = 1, \ldots, n$ is stationary and α-mixing with mixing coefficient $\alpha(l) \leq \rho^l$ for some $\rho \in (0, 1)$. ϵ_{it} has mean zero and is independent of $\{\mathbf{X}_t\}$. The variables $X_{t,j}$ are uniformly bounded. The density of $\mathbf{X}_t^{\mathrm{T}} \boldsymbol{\beta}_{0i}$ is bounded and bounded away from zero on its support which is a closed interval.

(C2) Let $\sigma_{ii',l} = E[\epsilon_{it} \epsilon_{i't'}]$ with $|t - t'| = l$. We assume $\sum_{l=1}^{n} |\sigma_{ii',l}| \leq \tau_{ii'}$ for some $\tau_{ii'} > 0$ and $\max_i \sum_{i'} \tau_{ii'} \leq M$ for some constant M.

(C3) The functions $a_{i,j}, j = 0, \ldots, q$ are twice continuously differentiable. By definition of projection (A.1) we can write $E_{\mathcal{M}_i}[\overline{\mathbf{X}}_t^{\mathrm{T}} \overline{\mathbf{a}}_i'(\mathbf{X}_{t-1}^{\mathrm{T}} \boldsymbol{\beta}_{0i}) X_{t-1,j}] = \overline{\mathbf{X}}_t^{\mathrm{T}} \overline{\mathbf{b}}_{ij}(\mathbf{X}_{t-1}^{\mathrm{T}} \boldsymbol{\beta}_{0i})$ for some functions $\overline{\mathbf{b}}_{ij} = (b_{ij,0}, \ldots, b_{ij,q})^{\mathrm{T}}$. We also assume $b_{ij,k}$ are twice continuously differentiable.

(C4) $E[\mathbf{X}_t \mathbf{X}_t^{\mathrm{T}}]$ and $E[\mathbf{M}_i^{\mathrm{T}} (\overline{\mathbf{X}}_t^{\mathrm{T}} \overline{\mathbf{a}}_i'(\mathbf{X}_{t-1}^{\mathrm{T}} \boldsymbol{\beta}_{0i}) \mathbf{X}_{t-1} - E_{\mathcal{M}_i}[\overline{\mathbf{X}}_t^{\mathrm{T}} \overline{\mathbf{a}}_i'(\mathbf{X}_{t-1}^{\mathrm{T}} \boldsymbol{\beta}_{0i}) \mathbf{X}_{t-1}])^{\otimes 2} \mathbf{M}_i]$ have eigenvalues bounded and bounded away from zero, uniformly over i, where for any matrix $\mathbf{A}$, $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$.

(C5) $H, \mathcal{N}, \varsigma, \tau$ are fixed scalars (non-diverging with sample size). $|\mathcal{D}_j|/(p_n(q-1)) \to c_j$ for some constant $c_j \in (0, 1)$, $j = 1, \ldots, H$, where $|\mathcal{D}_j|$ is the cardinality of the set $\mathcal{D}_j$, and similarly for the other three partitions for $a_{i,j}$, $\alpha_{i,j}$ and $\gamma_{i,j}$.

(C6) Assume $\sqrt{p_n K \log n/n} << \delta << \lambda$, where K is equal to either K_1 or K_2, λ is the minimum jump size for any of the four sequences (defined in stage 2 of the estimation procedure) at the change points, and δ is the threshold used in the change point detection algorithm (we stop partitioning if the test statistic is below δ).

(C7) $\{\mathbf{X}_t, t = 1, \ldots, n\}$ is a stationary Markov chain. $E[X_{t,j}|\mathbf{X}_{t-1} = \mathbf{u}]$ and $E[X_{t,j}X_{t,j'}|\mathbf{X}_{t-1} = \mathbf{u}]$ are twice continuously differentiable in $\mathbf{u}$.

(C8) For each i, $(\epsilon_{it}, \sigma_{it}^2)$ is a strictly stationary and ergodic GARCH process with $\sup_i E[\sigma_{it}^{2d}] < \infty$ for some $d > 2$.

(C9) For each i, the innovations $\nu_{it} = \epsilon_{it}/\sigma_{it}$ are i.i.d. with a nondegenerate distribution, $E\nu_{it}^2 = 1$ and $\sup_i E[\nu_{it}^{2d}] < \infty$ with the same d as defined in (C8). Furthermore, ν_{it} are weakly correlated in the sense that there exists some $\delta > 0$ such that $(E[|(\nu_{it}^2 - 1)(\nu_{i't'}^2 - 1)|^{1+\delta}])^{\frac{1}{1+\delta}}$ satisfies the same condition assumed for $E[\epsilon_{it}\epsilon_{i't'}]$ in assumption (C2).

(C10) Let $\mathbf{\Omega}$ be a compact subset of $(0, \infty)^{m+s+1}$. $\sup_{(\boldsymbol{\alpha}_i, \boldsymbol{\gamma}_i) \in \mathbf{\Omega}} \sum_{j=1}^{s} \gamma_{i,j} < 1$, and $(\boldsymbol{\alpha}_{0i}, \boldsymbol{\gamma}_{0i})$ is an interior point of $\mathbf{\Omega}$.

(C11) Let $\mathcal{A}(z) = \sum_{j=1}^{m} \alpha_{0i,j} z^j$ and $\mathcal{B}(z) = 1 - \sum_{j=1}^{s} \gamma_{0i,j} z^j$. $\mathcal{A}(z)$ and $\mathcal{B}(z)$ have no common roots on the complex plane $\mathbb{C}$, $\mathcal{A}(1) \neq 0$, $\alpha_{0i,m} + \gamma_{0i,s} \neq 0$.

(C12) The kernel function $K(z)$ is a symmetric density function that is bounded on a bounded support and satisfies the Lipschitz condition. The bandwidth h satisfies $h \asymp n^{-c}$ with $0 < c < 1/(q+1)$.

(C13) The number of knots for splines satisfies $K_1 \asymp n^{1/5}$, $K_2 \asymp (p_n n)^{1/5}$. $p_n \lesssim n^b$ with $b < \min\{d/2 - 1, 1\}$.

Remark 5. (C1) contains some mild regularity assumptions. Assuming $X_{t.j}$ to be bounded is common in estimation with B-splines since the basis functions are constructed on a compact interval. (C2) roughly means the dependence across i is not too strong. If p_n is fixed, (C2) follows from the geometric mixing assumption. Assumptions similar to (C2) were also used in Bai (2003) to impose weak dependence among errors. Note Vogt and Linton (2017) made the stronger assumption that the data are independent across i which also easily implies (C2). (C3) contains smoothness conditions for some functions and (C4) contains some identifiability conditions usually assumed in models with a single-index structure and involves the projection one typically use to profile out the nonparametric part. Uniformity over i in various assumptions above is void if p_n is fixed.

(C5) is assumed for ease of exposition so that it can be clearly demonstrated that homogeneity pursuit can significantly improve the convergence rate. (C6) is used in showing that stage 2 of our estimation procedure can identify the true partition with probability approaching one. (C7) and (C12) are the same as assumed in Guo *et al.*(2017) for the estimation of $\mathbf{\Sigma}_x$. (C8)-(C9) are mild regularity assumptions for the GARCH model. Assumptions (C10) and (C11) imply that (1.5) admits a unique strictly stationary solution, and the parameters are identified. (C13) restricts the divergence rate of p_n.

For the misspecified case, we introduce the following notations and definitions before listing additional assumptions below. For any $\boldsymbol{\beta}$ with unit norm, let $\overline{\mathbf{a}}_i(x;\boldsymbol{\beta}) = (a_{i,0}(x;\boldsymbol{\beta}),\ldots,a_{i,q}(x;\boldsymbol{\beta}))^{\mathrm{T}} = \arg\min_{\mathbf{a}} E[(y_{it} - \overline{\mathbf{X}}_t^{\mathrm{T}}\mathbf{a})^2|\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta} = x]$ and $\boldsymbol{\beta}_0^{(m)} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{p_n} E[(y_{it} - \overline{\mathbf{X}}_t^{\mathrm{T}}\overline{\mathbf{a}}_i(\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}))^2]$. Here $\boldsymbol{\beta}_0^{(m)}$ and $\overline{\mathbf{a}}(x;\boldsymbol{\beta})$ can be regarded as the population limit of our estimator under misspecification. We generally use superscript (m) to denote the quantities under misspecification.

Misspecification in $\boldsymbol{\beta}$ also means the population limit of our estimator of $\boldsymbol{\alpha}_i = (\alpha_{i,0},\ldots,\alpha_{i,m})^{\mathrm{T}}$ and $\boldsymbol{\gamma}_i = (\gamma_{i,1},\ldots,\gamma_{i,s})^{\mathrm{T}}$ is no longer the true value in the GARCH model (1.5). Let $\epsilon_{it}^{(m)} = y_{it} - \overline{\mathbf{X}}_t^{\mathrm{T}}\overline{\mathbf{a}}_i(\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_0^{(m)};\boldsymbol{\beta}_0^{(m)})$ and define $\sigma_{it}^{(m)}(\boldsymbol{\vartheta}_i)$ with $\boldsymbol{\vartheta}_i = (\boldsymbol{\alpha}_i,\boldsymbol{\gamma}_i)$ by

$$(\sigma_{it}^{(m)})^2(\boldsymbol{\vartheta}_i) = \alpha_{k,0} + \sum_{j=1}^{m}\alpha_{i,j}(\epsilon_{i,t-j}^{(m)})^2 + \sum_{j=1}^{s}\gamma_{i,j}(\sigma_{i,t-j}^{(m)})^2(\boldsymbol{\vartheta}_i).$$

The quasi-likelihood assuming the knowledge of $\epsilon_{it}^{(m)}$ is $\mathcal{L}_i^{(m)}(\boldsymbol{\vartheta}_i) = (1/n)\sum_t(\epsilon_{it}^{(m)})^2/(\sigma_{it}^{(m)}(\boldsymbol{\vartheta}_i))^2 + \log(\sigma_{it}^{(m)}(\boldsymbol{\vartheta}_i))^2$.

Finally, we note that $\mathrm{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ can obviously be regarded as a random matrix-valued function of the parameters. Thus we can write, say, $\mathrm{cov}(\mathbf{Y}_{n+1}|\mathcal{F}_n) = \mathbf{R}(\{\boldsymbol{\beta}_i\},\{\overline{\mathbf{a}}_i\},\{\boldsymbol{\vartheta}_i\})$.

(C14) $a_{i,j}(x;\boldsymbol{\beta})$ is twice continuously differentiable in both x and $\boldsymbol{\beta}$, and $\sum_{i=1}^{p_n} E[(y_{it} - \overline{\mathbf{X}}_t^{\mathrm{T}}\overline{\mathbf{a}}_i(\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}))^2]$ has a unique minimizer so that $\boldsymbol{\beta}_0^{(m)}$ is well-defined.

(C15) For each i, $(\epsilon_{it}^{(m)}, \sigma_{it}^{(m)}(\boldsymbol{\vartheta}_i))$ is strictly stationary and ergodic with $\sup_{i,\boldsymbol{\theta}_i\in\mathbf{\Omega}} E[(\sigma_{it}^{(m)}(\boldsymbol{\vartheta}_i))^{2d}] < \infty$. For all i, $E[\mathcal{L}_i^{(m)}(\boldsymbol{\vartheta}_i)]$ has a unique minimizer in the interior of $\mathbf{\Omega}$, say $\boldsymbol{\vartheta}_{0i}^{(m)} = (\boldsymbol{\alpha}_{0i}^{(m)},\boldsymbol{\gamma}_{0i}^{(m)})$.

(C16) $\|\mathbf{R}(\{\boldsymbol{\beta}_0^{(m)}\},\{\overline{\mathbf{a}}_i^{(m)}\},\{\boldsymbol{\vartheta}_{0i}^{(m)}\}) - \mathbf{R}(\{\boldsymbol{\beta}_{0i}\},\{\overline{\mathbf{a}}_i\},\{\boldsymbol{\vartheta}_{0i}\})\|_{\mathbf{\Sigma}}^2 > c_1 p_n$ with probability at least c_2 for some positive constants c_1, c_2.

(C14) and (C15) are mild assumptions in the misspecified case. We do not impose those more stringent assumptions as in the overfitting and correct fitting case since we will only aim to establish convergence of the estimators in the misspecified case, rather than tight convergence rates. Assumption (C16) is a high-level assumption which guarantees that inconsistent estimate of the

parameters will result in inconsistent estimate of the covariance matrix. The scaling of p_n in the lower bound of (C16) is natural by the definition of the $\|.\|_{\mathbf{\Sigma}}$ norm.

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

Let

$$\overline{\boldsymbol{\theta}}_{0i} = (\boldsymbol{\theta}_{0i,0}^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_{0i,q}^{\mathrm{T}})^{\mathrm{T}} = \arg\min_{\overline{\boldsymbol{\theta}}} E[(\overline{\mathbf{X}}_t^{\mathrm{T}} \overline{\mathbf{a}}_i(\mathbf{X}_{t-1}^{\mathrm{T}} \boldsymbol{\beta}_{0i}) - (\overline{\mathbf{X}}_t \otimes \mathbf{B}(\mathbf{X}_{t-1}^{\mathrm{T}} \boldsymbol{\beta}_{0i}))^{\mathrm{T}} \overline{\boldsymbol{\theta}})^2], \tag{A.2}$$

and set $\overline{\mathbf{\Theta}}_{0i} = (\boldsymbol{\theta}_{0i,0}, \ldots, \boldsymbol{\theta}_{0i,q})$. The partition on functions $\{a_{i,j}\}$ induces a partition on the components of $\boldsymbol{\theta}_0 = (\overline{\boldsymbol{\theta}}_{01}^{\mathrm{T}}, \ldots, \overline{\boldsymbol{\theta}}_{0p_n}^{\mathrm{T}})^{\mathrm{T}}$. The induced partition on $\boldsymbol{\theta}_0$ has a constraint that each row of $(\boldsymbol{\theta}_{i,0}, \ldots, \boldsymbol{\theta}_{i,q})$ are partitioned in the same way. As mentioned previously in Remark **??**, we do not impose this constraint for flexibility and ease of modelling. Since the number of unique functions among $a_{i,j}$ is $\mathcal{N}$, the number of unique values in $\boldsymbol{\theta}_0$ is generally $\mathcal{N}K$. With abuse of notation, in the following we still use $\mathcal{Q}_1, \ldots, \mathcal{Q}_{\mathcal{N}K}$ to denote a partition of $\{(i, j, k) : 1 \le i \le p_n, 0 \le j \le q, 1 \le k \le K\}$ and assume naturally that $|\mathcal{Q}_j|/(p_n(q+1)K)$ has a limit in $(0, 1)$ for all $j = 1, \ldots, \mathcal{N}K$.

Assume the true partition of components of $\boldsymbol{\beta}_0^{(-1)} = (\boldsymbol{\beta}_{01}^{(-1)\mathrm{T}}, \ldots, \boldsymbol{\beta}_{0q}^{(-1)\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\theta}_0$ is given by $\mathcal{D}_h, h = 1, \ldots, H$ and $\mathcal{Q}_h, h = 1, \ldots, \mathcal{N}K$, respectively. The unique values of the components of $\boldsymbol{\beta}_0^{(-1)}$ and $\boldsymbol{\theta}_0$ are denoted by $\boldsymbol{\xi}_0 = (\xi_{01}, \ldots, \xi_{0H})^{\mathrm{T}}$ and $\boldsymbol{\eta}_0 = (\eta_{01}, \ldots, \eta_{0\mathcal{N}K})^{\mathrm{T}}$, respectively. Let $\mathbf{J}_i^{\mathcal{D}}$ be the $(q-1) \times H$ binary matrix whose (j, h) entry is 1 if $\beta_{0ij} = \xi_{0h}$ and 0 otherwise. We have $\boldsymbol{\beta}_{0i}^{(-1)} = \mathbf{J}_i^{\mathcal{D}} \boldsymbol{\xi}_0$. Similarly, we define $\mathbf{J}_i^{\mathcal{Q}}$ such that $\overline{\boldsymbol{\theta}}_{0i} = \mathbf{J}_i^{\mathcal{Q}} \boldsymbol{\eta}_0$. Finally, let $\mathbf{G}^{\mathcal{D}}$ and $\mathbf{G}^{\mathcal{Q}}$ be the diagonal matrices with entries $\sqrt{|\mathcal{D}_h|}, h = 1, \ldots, H$ and $\sqrt{|\mathcal{Q}_h|}, h = 1, \ldots, \mathcal{N}K$, respectively. For the parameters in the GARCH model, we write $\boldsymbol{\alpha}_i = (\alpha_{i,0}, \ldots, \alpha_{i,m})^{\mathrm{T}} = \mathbf{J}_i^{\mathcal{A}} \boldsymbol{\omega}$ and $\boldsymbol{\gamma}_i = (\gamma_{i,1}, \ldots, \gamma_{i,s})^{\mathrm{T}} = \mathbf{J}_i^{\Gamma} \boldsymbol{\pi}$, where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_\varsigma)^{\mathrm{T}}$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_\tau)^{\mathrm{T}}$ denote all the unique values of $\boldsymbol{\alpha}_i, i = 1, \ldots, p_n$ and $\boldsymbol{\gamma}_i, i = 1, \ldots, p_n$ respectively. We also define $\boldsymbol{\psi} = (\boldsymbol{\omega}^{\mathrm{T}}, \boldsymbol{\pi}^{\mathrm{T}})^{\mathrm{T}}$.

Appendix B. Sketch of Proofs for Theorem 1 and Theorem 2

The detailed proof is split into several subsections in the supplementary material. Here we just sketch the proof. In Section S.1, we consider the asymptotic property of the initial mean estimator $(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\beta}})$. Although the results are relatively standard, the challenge is to obtain bounds that hold uniformly over different responses $i = 1, \ldots, p_n$ in order to prove consistency of change point detection later. To this effect, tail bounds for the estimators are derived. In particular we show that $\max_i \|\widetilde{\overline{\boldsymbol{\theta}}}_i - \overline{\boldsymbol{\theta}}_{0i}\| + \|\widetilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{0i}\| = O_t(r'_n; p_n K^d (\log n)^{3d}/n^{d-1})$, where $r'_n = \sqrt{K \log n/n} + K^{-2}$. Although this rate is usable, we further refine it to $\max_i \|\widetilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{0i}\| = O_t(\sqrt{\log n/n}; p_n (\log n)^{3d}/n^{d-1} + p_n K^{-4d} + p_n K^d (\log n)^d/n^d)$ and $\max_i \|\widetilde{\overline{\boldsymbol{\theta}}}_i - \overline{\boldsymbol{\theta}}_{0i}\| = O_t(\sqrt{K \log n/n}; p_n K^d (\log n)^{3d}/n^{d-1} + p_n K^{-4d})$, in order to obtain more relaxed constraints regarding the order of p_n, K. These improved rates are obtained using the profiling strategy that orthogonalize the parametric part and the nonparametric part in the regression function. Such an strategy was often used in semiparametric models to show the improved rate of the parametric part, but as far as we know was not previously used for the nonparametric part as we do here.

With these bounds that are uniform over i, in Section S.2 we adapt the results of Venkatraman (1992) to show that the change point analysis can consistently identify the true partition for both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. This means we can then assume that the true partition is known when considering the asymptotic properties of the final estimator of the mean. Thus in Section S.3, we show that $\max_i \|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{0i}\| + \|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{0i}\| = O_p(\widetilde{r}_n)$ where $\widetilde{r}_n = \sqrt{K/(p_n n)} + K^{-2}$. Note that for the final estimator, the factor n in r'_n is replaced by $p_n n$ in $\widetilde{r}_n$, suggesting that for the correctly specified model, the sample size is effectively $p_n n$ as expected. The term K^{-2} comes from the bias of spline approximation and thus cannot be reduced even when the correct partition is known. The proof of the rate is based on the explicit parametrization of the model using $\boldsymbol{\beta}_i^{(-1)} = \mathbf{J}_i^{\mathcal{D}} \boldsymbol{\xi}$, $\bar{\boldsymbol{\theta}}_i = \mathbf{J}_i^{\mathcal{Q}} \boldsymbol{\eta}$ and carefully examining the structure and property of the binary matrices $\mathbf{J}_i^{\mathcal{D}}$ and $\mathbf{J}_i^{\mathcal{Q}}$. As done before for the initial estimator, by the profiling strategy we also obtain the improved rates $\max_i \|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{0i}\| = O_p(1/\sqrt{p_n n})$ and $\max_i \|\widehat{\overline{\boldsymbol{\theta}}}_i - \overline{\boldsymbol{\theta}}_{0i}\| = O_p(\sqrt{K/(p_n n)})$.

Now moving to the GARCH model for the error part. By the derived estimation error of the mean function, we can bound $\Delta := \max_{i,t} |\widehat{\epsilon}_{it} - \epsilon_{it}| = O_p\left(\sqrt{\frac{K}{p_n n}} + K^{-2}\right)$, where $\widehat{\epsilon}_{it}$ is the estimated residual. Let $\boldsymbol{\vartheta}_i = (\alpha_{i,0}, \alpha_{i,1}, \ldots, \alpha_{i,m}, \gamma_{i,1}, \ldots, \gamma_{i,s})^{\mathrm{T}}$ be all the parameters in the GARCH model, with true parameter value $\boldsymbol{\vartheta}_{0i} = (\alpha_{0i,0}, \alpha_{0i,1}, \ldots, \alpha_{0i,m}, \gamma_{0i,1}, \ldots, \gamma_{0i,s})^{\mathrm{T}}$. Let $\widetilde{\sigma}_{it}^2(\boldsymbol{\vartheta}_i)$ be defined iteratively by

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

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

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

both with the initial values given by (2.8) or (2.9). The negative quasi-log-likelihood is given by $\widehat{\mathcal{L}}_i(\boldsymbol{\vartheta}_i) = \frac{1}{n}\sum_{t=1}^{n}\widehat{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_i)$ with $\widehat{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_i) = \widehat{\epsilon}_{it}^2/\widehat{\sigma}_{it}^2(\boldsymbol{\vartheta}_i) + \log\widehat{\sigma}_{it}^2(\boldsymbol{\vartheta}_i)$. Similarly let $\widetilde{\mathcal{L}}_i(\boldsymbol{\vartheta}_i) = \frac{1}{n}\sum_{t=1}^{n}\widetilde{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_i)$ with $\widetilde{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_i) = \epsilon_{it}^2/\widetilde{\sigma}_{it}^2(\boldsymbol{\vartheta}_i) + \log\widetilde{\sigma}_{it}^2(\boldsymbol{\vartheta}_i)$. We further define $\sigma_{it}^2(\boldsymbol{\vartheta}_i)$ to be the unique strictly stationary solution of the GARCH model (1.5), and define $\mathcal{L}_i(\boldsymbol{\vartheta}_i) = \frac{1}{n}\sum_t \mathcal{L}_{it}(\boldsymbol{\vartheta}_i)$ with $\mathcal{L}_{it}(\boldsymbol{\vartheta}_i) = \epsilon_{it}^2/\sigma_{it}^2(\boldsymbol{\vartheta}_i) + \log\sigma_{it}^2(\boldsymbol{\vartheta}_i)$.

In Section S.4, we establish consistency of $\widetilde{\boldsymbol{\vartheta}}_i$ uniformly over i. Following the arguments of Theorem 7.1 in Francq and Zakoian (2019), we have

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

Thus we only need to show that

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

The strategy for establishing the latter is simply using the bound for Δ above, but involves detailed and technical derivations using the recursive form for the definition of $\widehat{\sigma}_{it}^2(\boldsymbol{\vartheta}_i)$.

In Section S.5, we show the convergence rate for the initial estimator

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

Since

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

where $\boldsymbol{\vartheta}_i^*$ lies between $\widetilde{\boldsymbol{\vartheta}}_i$ and $\boldsymbol{\vartheta}_{0i}$, we only need bounds for

$$\left\|\frac{1}{n}\sum_t \frac{\partial\widehat{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_{0i})}{\partial\boldsymbol{\vartheta}_i}\right\| \text{ and } \sup_{\boldsymbol{\vartheta}_i\in\mathcal{V}(\boldsymbol{\vartheta}_{0i})}\left\|\frac{1}{n}\sum_{t=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\vartheta}_i\partial\boldsymbol{\vartheta}_i^{\mathrm{T}}}\widehat{\mathcal{L}}_t(\boldsymbol{\vartheta}_i)\right\|$$

where $\mathcal{V}(\boldsymbol{\vartheta}_{0i})$ is an arbitrarily small neighborhood of $\boldsymbol{\vartheta}_{0i}$. Existing asymptotic theory for GARCH models has provided bounds when $\widehat{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_i)$ above is replaced by $\mathcal{L}_{it}(\boldsymbol{\vartheta}_i)$. In fact, adaptation of existing results yield

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

and

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

Furthermore, we will establish the bounds

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

and

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

These will complete our proof.

Lastly, in Section S.6, to establish the rate for the final estimator for the GARCH model parameters, we can assume the true partition is known (since it can be consistently estimated). First we establish the rate when ϵ_{it}^2 is observed with the estimator still denoted by $\widehat{\boldsymbol{\vartheta}}_i$. We write $\boldsymbol{\alpha}_i = (\alpha_{i,0}, \ldots, \alpha_{i,m})^{\mathrm{T}} = \mathbf{J}_i^{\mathcal{A}} \boldsymbol{\omega}$ and $\boldsymbol{\gamma}_i = (\gamma_{i,1}, \ldots, \gamma_{i,s})^{\mathrm{T}} = \mathbf{J}_i^{\Gamma} \boldsymbol{\pi}$, where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_\varsigma)^{\mathrm{T}}$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_\tau)^{\mathrm{T}}$ denotes all the unique values of $\boldsymbol{\alpha}_i, i = 1, \ldots, p_n$ and $\boldsymbol{\gamma}_i, i = 1, \ldots, p_n$ respectively. We also define $\boldsymbol{\psi} = (\boldsymbol{\omega}^{\mathrm{T}}, \boldsymbol{\pi}^{\mathrm{T}})^{\mathrm{T}}$. We can then write $\boldsymbol{\vartheta}_i = \mathbf{J}_i \boldsymbol{\psi}$ and $\widehat{\boldsymbol{\vartheta}}_i = \mathbf{J}_i \widehat{\boldsymbol{\psi}}$ with

$$\mathbf{J}_i = \begin{pmatrix} \mathbf{J}_i^{\mathcal{A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_i^{\Gamma} \end{pmatrix}.$$

The first order condition for the minimization yields

$$\begin{aligned} 0 &= \frac{1}{n} \sum_{i=1}^{p_n} \mathbf{J}_i^{\mathrm{T}} \sum_{t=1}^{n} \frac{\partial}{\partial \boldsymbol{\vartheta}_i} \mathcal{L}_{it}(\widehat{\boldsymbol{\vartheta}}_i) \\ &= \frac{1}{n} \sum_{i=1}^{p_n} \mathbf{J}_i^{\mathrm{T}} \sum_{t=1}^{n} \frac{\partial}{\partial \boldsymbol{\vartheta}_i} \mathcal{L}_{it}(\boldsymbol{\vartheta}_{0i}) + \frac{1}{n} \sum_{i=1}^{p_n} \left(\mathbf{J}_i^{\mathrm{T}} \sum_{t=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\vartheta}_i \partial \boldsymbol{\vartheta}_i^{\mathrm{T}}} \mathcal{L}_{it}(\boldsymbol{\vartheta}_i^*) \mathbf{J}_i \right) (\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0), \end{aligned}$$

where $\boldsymbol{\vartheta}_i^*$ lies between $\widehat{\boldsymbol{\vartheta}}_i$ and $\boldsymbol{\vartheta}_{0i}$. Let $\mathbf{G}^{\mathcal{A}}$ and $\mathbf{G}^{\Gamma}$ be the diagonal matrices with entries $\sqrt{|\mathcal{A}_h|}, h = 1, \ldots, \varsigma$ and $\sqrt{|\Gamma_h|}, h = 1, \ldots, \tau$, respectively and define

$$\begin{aligned} \mathbf{G} &:= \begin{pmatrix} \mathbf{G}^{\mathcal{A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{\Gamma} \end{pmatrix}, \\ \widetilde{\mathbf{O}} &:= \begin{pmatrix} \mathbf{J}_1^{\mathcal{A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_1^{\Gamma} \\ \vdots & \vdots \\ \mathbf{J}_{p_n}^{\mathcal{A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{p_n}^{\Gamma} \end{pmatrix} \begin{pmatrix} (\mathbf{G}^{\mathcal{A}})^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{G}^{\Gamma})^{-1} \end{pmatrix} = \begin{pmatrix} \mathbf{J}_1 \\ \vdots \\ \mathbf{J}_{p_n} \end{pmatrix} \mathbf{G}^{-1}. \end{aligned}$$

Then the first order condition is rewritten as

$$0 = \frac{1}{n}\sum_{t=1}^{n}\widetilde{\mathbf{O}}^{\mathrm{T}}\frac{\partial}{\partial\boldsymbol{\vartheta}}\mathcal{L}_t(\boldsymbol{\vartheta}_0)+\frac{1}{n}\sum_{t=1}^{n}\widetilde{\mathbf{O}}^{\mathrm{T}}\frac{\partial^2}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^{\mathrm{T}}}\mathcal{L}_t(\boldsymbol{\vartheta}^*)\widetilde{\mathbf{O}}\mathbf{G}(\widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0),$$

where

$$\frac{\partial}{\partial\boldsymbol{\vartheta}}\mathcal{L}_t(\boldsymbol{\vartheta})=\left(\frac{\partial}{\partial\boldsymbol{\vartheta}_1^{\mathrm{T}}}\mathcal{L}_{1t}(\boldsymbol{\vartheta}_1),\ldots,\frac{\partial}{\partial\boldsymbol{\vartheta}_{p_n}^{\mathrm{T}}}\mathcal{L}_{p_nt}(\boldsymbol{\vartheta}_{p_n})\right)^{\mathrm{T}},$$

and

$$\frac{\partial^2}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^{\mathrm{T}}}\mathcal{L}_t(\boldsymbol{\vartheta})=\mathrm{diag}\left\{\frac{\partial^2}{\partial\boldsymbol{\vartheta}_1\partial\boldsymbol{\vartheta}_1^{\mathrm{T}}}\mathcal{L}_{1t}(\boldsymbol{\vartheta}_1),\ldots,\frac{\partial^2}{\partial\boldsymbol{\vartheta}_{p_n}\partial\boldsymbol{\vartheta}_{p_n}^{\mathrm{T}}}\mathcal{L}_{p_nt}(\boldsymbol{\vartheta}_{p_n})\right\}.$$

Using assumption (C9) and similar to the proof of Lemma 4, we can show

$$\|\sum_{t=1}^{n}\widetilde{\mathbf{O}}^{\mathrm{T}}\frac{\partial}{\partial\boldsymbol{\vartheta}}\mathcal{L}_t(\boldsymbol{\vartheta}_0)\|=O_p(\sqrt{n}).$$

Thus we get $\|\mathbf{G}(\widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}_0)\|=O_p(1/\sqrt{n})$. By assumption (C5), this implies $\|\boldsymbol{\psi}-\boldsymbol{\psi}_0\|=O_p(1/\sqrt{p_nn})$.

When using quasi-likelihood $\widehat{\mathcal{L}}_{it}(\boldsymbol{\vartheta}_i)$ based on the estimated residuals $\widehat{\epsilon}_{it}$, as before the proof mainly consists of bounding the difference between $\widehat{\mathcal{L}}_t(\boldsymbol{\vartheta})$ and $\mathcal{L}_t(\boldsymbol{\vartheta})$ (as well as their derivatives), taking into account the effect of $\widetilde{\mathbf{O}}$. Such bounds combined with the first order condition above eventually yield the desired conclusion $\|\boldsymbol{\psi}-\boldsymbol{\psi}_0\|=O_p(\sqrt{K/(p_nn)}+K^{-2})$ and $\|\widehat{\boldsymbol{\vartheta}}_i-\boldsymbol{\vartheta}_{0i}\|=O_p(\sqrt{K/(p_nn)}+K^{-2})$.

Given the rates obtained for estimators of $\boldsymbol{\beta}_i$, $\mathbf{a}_i$, $\boldsymbol{\alpha}_i$, $\boldsymbol{\gamma}_i$, and $\boldsymbol{\Sigma}_x$ (proved in Lemma D.1 of Guo *et al.*(2017)), the proof of convergence rate for $\widehat{\mathrm{Cov}}(\mathbf{Y}_{n+1}|\mathcal{F}_n)$ is exactly as the proof of Theorem 2 in Guo *et al.*(2017).

Appendix C: Proof of Theorem 3

Let $\widehat{\boldsymbol{\beta}}_1^{(m)}=\cdots=\widehat{\boldsymbol{\beta}}_{p_n}^{(m)}=\widehat{\boldsymbol{\beta}}^{(m)}$ and $\widehat{\overline{\boldsymbol{\Theta}}}_i^{(m)}, i=1,\ldots,p_n$ be the minimizer of $\sum_i\sum_t(y_{it}-\mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_i)\overline{\boldsymbol{\Theta}}_i\overline{\mathbf{X}}_t)^2$ with the constraint $\boldsymbol{\beta}_1=\cdots=\boldsymbol{\beta}_{p_n}$. For any $\boldsymbol{\beta}$, define

$$\widehat{\overline{\boldsymbol{\Theta}}}_i^{(m)}(\boldsymbol{\beta})=\arg\min_{\boldsymbol{\Theta}}\sum_t(y_{it}-\mathbf{B}^{\mathrm{T}}(\mathbf{X}_{t-1}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\Theta}\overline{\mathbf{X}}_t)^2.$$

Obviously we have $\widehat{\overline{\boldsymbol{\Theta}}}_i^{(m)}=\widehat{\overline{\boldsymbol{\Theta}}}_i^{(m)}(\widehat{\boldsymbol{\beta}}^{(m)})$. Let $\widehat{\overline{\mathbf{a}}}_i(x;\boldsymbol{\beta})=(\widehat{\overline{\boldsymbol{\Theta}}}_i^{(m)}(\boldsymbol{\beta}))^{\mathrm{T}}\mathbf{B}(x)$. Since $\boldsymbol{\beta}$ is in a compact set, based on standard results for the spline estimator, it can be shown that

$$\sup_{i,x,\boldsymbol{\beta}}\|\widehat{\overline{\mathbf{a}}}_i(x;\boldsymbol{\beta})-\overline{\mathbf{a}}_i(x;\boldsymbol{\beta})\|=o_p(1), \tag{C.1}$$

and

$$\sup_{i,x,\boldsymbol{\beta}}\|\widehat{\overline{\mathbf{a}}'}_i(x;\boldsymbol{\beta})-\overline{\mathbf{a}}'_i(x;\boldsymbol{\beta})\|=o_p(1), \tag{C.2}$$

where $\overline{\mathbf{a}}'_i(x;\boldsymbol{\beta})$ denotes the derivative with respect to x. Similar results for the even more complicated quantile regression case has been established in Proposition 1 of Zhao *et al.*(2018), for example. Then we proceed similarly as in Ichimura (1993) to establish convergence of $\widehat{\boldsymbol{\beta}}^{(m)}$ to $\boldsymbol{\beta}_0^{(m)}$ as follows.

Write $J_n(\boldsymbol{\beta}) = (np_n)^{-1}\sum_{i,t}(y_{it}-\overline{\mathbf{X}}_{it}^{\mathrm{T}}\widehat{\overline{\mathbf{a}}}_i(\mathbf{X}_{it-1}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}))^2$, $\widetilde{J}_n(\boldsymbol{\beta}) = (np_n)^{-1}\sum_{i,t}(y_{it}-\overline{\mathbf{X}}_{it}^{\mathrm{T}}\overline{\mathbf{a}}_i(\mathbf{X}_{it-1}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}))^2$ and $J(\boldsymbol{\beta}) = (np_n)^{-1}E\left[\sum_{i,t}(y_{it}-\overline{\mathbf{X}}_{it}^{\mathrm{T}}\overline{\mathbf{a}}_i(\mathbf{X}_{it-1}^{\mathrm{T}}\boldsymbol{\beta};\boldsymbol{\beta}))^2\right]$. Then we have, as in the proof of Theorem 5.1 in Ichimura (1993),

$$\begin{aligned}
& P(\|\widehat{\boldsymbol{\beta}}^{(m)} - \boldsymbol{\beta}_0^{(m)}\| > \delta) \\
\leq\ & P\left(\inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0^{(m)}\|>\delta} J_n(\boldsymbol{\beta}) < J_n(\boldsymbol{\beta}_0^{(m)})\right) \\
\leq\ & P\left(2\sup_{\boldsymbol{\beta}}|J_n(\boldsymbol{\beta}) - \widetilde{J}_n(\boldsymbol{\beta})| + 2\sup_{\boldsymbol{\beta}}|\widetilde{J}_n(\boldsymbol{\beta}) - J(\boldsymbol{\beta})| > \inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0^{(m)}\|>\delta} J(\boldsymbol{\beta}) - J(\boldsymbol{\beta}_0^{(m)})\right).
\end{aligned}$$

Using (C.1), we have $\sup_{\boldsymbol{\beta}}|J_n(\boldsymbol{\beta}) - \widetilde{J}_n(\boldsymbol{\beta})| = o_p(1)$. The uniform law of large numbers result $\sup_{\boldsymbol{\beta}}|\widetilde{J}_n(\boldsymbol{\beta}) - J(\boldsymbol{\beta})| = o_p(1)$ can be established using Theorem 2.18 (ii) of Fan and Yao (2003) (similar to the way applied in the proof of Lemma 6). Finally, assumption (C14) means that $\inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0^{(m)}\|>\delta} J(\boldsymbol{\beta}) - J(\boldsymbol{\beta}_0^{(m)})$ is bounded away from zero. Thus $\|\widehat{\boldsymbol{\beta}}^{(m)} - \boldsymbol{\beta}_0^{(m)}\| = o_p(1)$.

Let $\widehat{\epsilon}_{it}^{(m)} = y_{it} - \overline{\mathbf{X}}_t^{\mathrm{T}}\widehat{\overline{\mathbf{a}}}_i(\mathbf{X}_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(m)};\widehat{\boldsymbol{\beta}}^{(m)})$ and define $\widehat{\sigma}_{it}^{(m)}(\boldsymbol{\vartheta}_i)$ by

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

and let $\widehat{\mathcal{L}}_i^{(m)}(\boldsymbol{\vartheta}_i) = (1/n)\sum_t(\widehat{\epsilon}_{it}^{(m)})^2/(\widehat{\sigma}_{it}^{(m)}(\boldsymbol{\vartheta}_i))^2 + \log(\widehat{\sigma}_{it}^{(m)}(\boldsymbol{\vartheta}_i))^2$.

Using the convergence of $\widehat{\boldsymbol{\beta}}^{(m)}$ to $\boldsymbol{\beta}_0^{(m)}$ as well as (C.1) and (C.2), we have

$$\Delta^{(m)} := \max_{i,t}|\widehat{\epsilon}_{it}^{(m)} - \epsilon_{it}^{(m)}| = o_p(1).$$

Following the proof of consistency of $\boldsymbol{\vartheta}_i$ in Section S.4, we get $\max_i\|\widehat{\boldsymbol{\vartheta}}_i^{(m)} - \boldsymbol{\vartheta}_i^{(m)}\| = o_p(1)$, where $\widehat{\boldsymbol{\vartheta}}_i^{(m)}$ is the minimizer of $\widehat{\mathcal{L}}_i^{(m)}(\boldsymbol{\vartheta}_i)$.

Using the convergence of all the parameters/functions even in the misspecificed case, exactly as in the proof of Theorem 2 of Guo *et al.*(2017), we can show the $\|.\|_{\boldsymbol{\Sigma}}$-norm of the difference between the estimated conditional covariance matrix of $\mathbf{Y}_{n+1}$ and $\mathbf{R}(\{\boldsymbol{\beta}_0^{(m)}\},\{\overline{\mathbf{a}}_i^{(m)}\},\{\boldsymbol{\vartheta}_{0i}^{(m)}\})$ is of order $o_p(p_n)$, and the proof is complete by condition (C16). □

Supplementary Materials

The supplementary material contains detailed proofs of Theorems 1 and 2 in Section 3 and additional simulation results in Section 5.

Acknowledgements

The authors sincerely thank the Editor Professor Jianqing Fan, the Associate Editor and three anonymous reviewers for their insightful comments that significantly improve the paper.

References

Almeida, D., Hotta, L. and Ruiz, E. (2018). MGARCH models: Trade-off between feasibility and flexibility. *International Journal of Forecasting*, **34**, 45-63.

Asai, M., McAleer, M. and Yu, J. (2006). Multivariate stochastic volatility: a review. *Econometric Reviews*, **25**, 145-175.

Avella-Medina, M., Battey, H., Fan, J., and Li, Q. (2018). Robust estimation of high dimensional covariance and precision matrices. *Biometrika*, **105**, 271-284.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, **71**, 135–171.

Bauwens, L., Laurent, S. and Rombouts, J. (2006). Multivariate GARCH models: a survey. *J. Appl. Econ.*, **21**, 79-109.

Berthet, Q. and Rigollet, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.*, **41**, 1780-1815.

Bickel, P. and Levina, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.*, **36**, 2577-2604.

Bickel, P. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.*, **36**, 199-227.

Birnbaum, A., Johnstone, I. M., Nadler, B. and Paul, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.*, **41**, 1055-1084.

Bollerslev, T., Engle, R. and Wooldridge, J. (1988). A capital asset pricing model with time-varying covariances. *Journal of Political Economy* , 116-31.

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

Boudt, K., Galanos, A., Payseur, S. and Zivot, E. (2019). Multivariate GARCH models for large-scale applications: A survey. *Conceptual econometrics using R. In Handbook of statistics 41.* 193-242.

Chib, S., Omori, V. and Asai, M. (2009). Multivariate stochastic volatility. *Handbook of Financial Time Series.* 365-400.

El Karoui, N. (2008). Operator norm consistent estimation of a large dimensional sparse covariance matrices. *Ann. Statist.*, **36**, 2717-2756.

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models *Journal of Business & Economic Statistics*. **20**, 339-350.

Engle, R. (2009). *Anticipating Correlations: A New Paradigm for Risk Management*, Princeton University Press.

Engle, R. and Sheppard, K (2001). Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH. *National Bureau of Economic Research*.

Engle, R., Ledoit, O. and Wolf, M. (2019). Large dynamic covariance matrices. *Journal of Business & Economic Statistics*, **37**, 363-375.

Fama, E. and French, K. (1992). The cross-section of expected stock returns. *J. Finance* **47**, 427-465.

Fama, E. and French, K. (1993). Common risk factors in the returns on stocks and bonds. *J. Financ. Econom.* **33**, 3-56.

Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics*, **147**, 186-197.

Fan, J., Liao, Y., and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.*, **39(6)**, 3320 - 3356.

Fan, J., Liu, H., and Wang, W. (2018). Large covariance estimation through elliptical factor models. *Annals of Statistics*, **46**, 1383-1414.

Fan, J., and Yao, Q. (2003). Nonlinear time series: nonparametric and parametric methods. Springer Verlag.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.*, **27**, 1491-1518.

Fang, Y., Wang, B. and Feng, Y. (2016). Tuning-parameter selection in regularized estimations of large covariance matrices. *Journal of Statistical Computation and Simulation*, **83**, 494-509.

Francq, C. and Zakoian, J. M. (2019). GARCH models: structure, statistical inference and financial applications. John Wiley & Sons.

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.

Harvey, A., Ruiz, E. and Shephard, N. (1994). Multivariate stochastic variance models. *Review of Economic Studies*, **61**, 247-264.

Haugen, R. A. and Baker, N. L. (1991). The efficient market inefficiency of capitalization-weighted stock portfolios. *Journal of Portfolio Management*, **17**, 35-40.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, **58**, 71–120.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, **54**, 1651-1684.

Kastner, G. (2019). Dealing with stochastic volatility in time series using the R package stochvol. *arXiv preprint arXiv:1906.12134, 2019 - arxiv.org*.

Ke, Y., Minsker, S., Ren, Z., Sun, Q. and Zhou, W.X. (2019). User-friendly covariance estimation for heavy-tailed distributions. *Statistical Science*, **34**, 454-471.

Lam, C. and Fan J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.*, **37**, 4254-4278.

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. (2011). Robust performances hypothesis testing with the variance. *Wilmott*, **2011**, 86-89.

Ledoit, O. and Wolf, M. (2008). Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance*, **15**, 850-859.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.*, **40**, 1024-1060.

Ledoit, O. and Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Annals of Statistics*. Forthcoming.

Markowitz, H.M. (1952). Portfolio selection *J. Finance*, **7**, 77-91.

Markowitz, H.M. (1959). Portfolio Selection: Efficient Diversification of Investments. John Wiley & Sons, New Jersey.

Mirsky, L. (1975). A trace inequality of John von Neumann. *Monatshefte fúr Mathematik*, **79**, 303–306.

Nielsen, F. and Aylursubramanian, R. (2008). Far from the madding crowd–Volatility efficient indices. *Research insights, MSCI Barra.*

Pakel, C., Shephard, N., Sheppard, K. and Engle, R. (2020). Fitting vast dimensional time-varying covariance models. *Journal of Business & Economic Statistics*, **38**, 1-17.

Rothman, A., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.*, **104**, 177-186.

Silvennoinen, A. and Teräsvirta, T. (2009). Modeling multivariate autoregressive conditional heteroskedasticity with the double smooth transition conditional correlation GARCH model. *Journal of Financial Econometrics*, **7**, 373-411.

Sun, Y., Zhang, W. and Tong, H. (2007). Estimation of the covariance matrix of random effects in longitudinal studies. *The Annals of Statistics*, **35**, 2795-2814.

Venkatraman, E.S. (1992). Consistency results in multiple change-point problems. *Thesis, Stanford University.*

Vogt, M. and Linton, O. B. (2017). Classification of nonparametric regression functions in heterogeneous panels. *Journal of the Royal Statistical Society Series B-Methodological*, **79**, 5–27.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, **90**, 831-884.

Xia Y. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis*, **97**, 1162-1184.

Xia, Y. and Li, W.K. (1999). On single-index coefficient regression models. *J. Amer. Statist. Assoc.*, **94**, 1275-1285.

Yu, J. and Meyer, R. (2006). Multivariate stochastic volatility models: Bayesian estimation and model comparison. *Econometric Reviews*, **25**, 361-384.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, **97**, 1042-1054.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, **11**, 2261-2286.

Zhao, W., Li, J. and Lian, H. (2018). Adaptive varying-coefficient linear quantile model: a profiled estimating equations approach. *Annals of the Institute of Statistical Mathematics*, **70**, 553–582.

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. (2011). Robust performances hypothesis testing with the variance. *Wilmott*, **2011**, 86-89.

Ledoit, O. and Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Annals of Statistics*. Forthcoming.