A Dynamic Structure for High Dimensional Covariance Matrices and its Application in Portfolio Allocation *

Shaojun Guo
Chinese Academy of Sciences, Beijing, People's Republic of China
& London School of Economics, London, United Kingdom

John Box
Department of Mathematics
University of York, United Kingdom

Wenyang Zhang
Department of Mathematics
University of York, United Kingdom

March 19, 2018

Abstract

Estimation of high dimensional covariance matrices is an interesting and important research topic. In this paper, we propose a dynamic structure and develop an estimation procedure for high dimensional covariance matrices. Asymptotic properties are derived to justify the estimation procedure and simulation studies are conducted to demonstrate its performance when the sample size is finite. By exploring a financial application, an empirical study shows that portfolio allocation based on dynamic high dimensional covariance matrices can significantly outperform the market from 1995 to 2014. Our proposed method also outperforms portfolio allocation based on the sample covariance matrix and the portfolio allocation proposed in Fan, Fan and Lv (2008).

KEY WORDS: Dynamic structure, factor models, high dimensional covariance matrices, iterative algorithm, kernel smoothing, portfolio allocation, single-index models.

SHORT TITLE: Dynamic Structure for HDCM.

*This research is supported by the Singapore National Research Foundation under its Cooperative Basic Research Grant and administered by the Singapore Ministry of Health's National Medical Research Council (Grant No. NMRC/CBRG/0014/2012) and the National Natural Science Foundation of China (Grant No. 11271242). John Box was supported by an EPSRC funded studentship through the University of York. Shaojun Guo was partly supported by Key Laboratory of RCSDS, Chinese Academy of Sciences and an EPSRC research grant in United Kingdom.

1 Introduction

Covariance matrix estimation is an important topic in statistics and econometrics with wide applications in many disciplines, such as economics, finance and psychology. A traditional approach to estimating covariance matrices is based on the sample covariance matrix. However, the sample covariance matrix would not be a good choice when the dimension is large, and especially when the inverse is required, which is often the case when constructing a portfolio allocation in finance. This is because the estimation errors would accumulate when using the inverse of the sample covariance matrix to estimate the inverse of the covariance matrix. When the size of the covariance matrix is large, the cumulative estimation error would become unacceptable even if the estimation error of each entry of the covariance matrix is tiny.

In recent years there has been various attempts to address high dimensional covariance matrix estimation. Usually, a sparsity condition is imposed to control the trade-off between variance and bias. See, Wu and Pourahmadi (2003), El Karoui (2008), Bickel and Levina (2008a, 2008b), Lam and Fan (2009), Fan, Liao, and Mincheva (2011), and the references therein. Fan, Fan and Lv (2008) considered a different approach by imposing a factor model and estimated the covariance matrix based on this structure.

Most of the literature addressing high dimensional covariance matrix estimation assumes that the covariance matrix is constant over time. However, in many applications, covariance matrices are dynamic. For example, today's optimal portfolio allocation may not be optimal tomorrow, or next month. Therefore, when applying the formula for Markowitz's optimal portfolio allocation (Markowitz 1959), the covariance matrix used should be dynamic and allowed to change over time.

In order to introduce a dynamic structure for covariance matrices, one cannot simply assume each entry of a covariance matrix is a function of time because this would not serve very well in prediction. Instead, we start with an approach stimulated by Fan, Fan and Lv (2008) which is based on the Fama-French three-factor model (Fama and French, 1992, 1993)

$$y_t = \alpha + X_t^{\mathrm{T}}\mathbf{a} + \epsilon_t, \tag{1.1}$$

where y_t is the excess return of an asset and X_t is the vector of the three factors at time t. To make (1.1) more flexible, we allow $\mathbf{a}$ to depend on the values of the three factors at time $t-1$. To avoid the so-called 'curse of dimensionality', we assume this dependence is through a linear combination of the values of the three factors at time $t-1$, which brings us to

$$y_t = \alpha(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + X_t^{\mathrm{T}}\mathbf{a}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + \epsilon_t. \tag{1.2}$$

This motivates a dynamic structure for the covariance matrix of a random vector Y_t through an adaptive varying coefficient model which we shall now introduce.

Suppose $(X_t^{\mathrm{T}},\ Y_t^{\mathrm{T}})$, $t = 1,\ \cdots,\ n$, is a time series, where Y_t is a p_n dimensional vector and 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. Also, we assume that X_t, $t = 1,\ \cdots,\ n$, is a stationary Markov process. We assume

$$Y_t = \mathbf{g}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + \boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})X_t + \boldsymbol{\epsilon}_t, \quad \|\boldsymbol{\beta}\| = 1, \quad \beta_1 > 0 \tag{1.3}$$

where $\boldsymbol{\beta} = (\beta_1,\ \cdots,\ \beta_q)^{\mathrm{T}}$, $\boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})$ is a factor loading matrix which is varying over $X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}$, and $\{\boldsymbol{\epsilon}_t,\ t = 1,\ \cdots,\ n\}$ are random errors which are independent of $\{X_t,\ t = 1,\ \cdots,\ n\}$. We assume

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

where

$$\sigma_{kt}^2 = \alpha_{k,0} + \sum_{i=1}^{m}\alpha_{k,i}\epsilon_{k,t-i}^2 + \sum_{j=1}^{s}\gamma_{k,j}\sigma_{k,t-j}^2, \quad t = 2,\ \cdots,\ n, \tag{1.4}$$

for each $k = 1,\ \cdots,\ p_n$ and for some integers m and s. Let $\mathcal{F}_t$ be the $\sigma-$algebra generated by $\{(X_l^{\mathrm{T}},\ \boldsymbol{\epsilon}_l^{\mathrm{T}}) :\ l \leq t\}$. The main focus of this paper is on the conditional covariance matrix

$$\mathrm{cov}(Y_t|\mathcal{F}_{t-1}) = \boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\Sigma}_x(X_{t-1})\boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}} + \boldsymbol{\Sigma}_{0,t} \tag{1.5}$$

where $\boldsymbol{\Sigma}_x(X_{t-1}) = \mathrm{cov}(X_t|X_{t-1})$. In (1.5), $\boldsymbol{\Phi}(\cdot)$, $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_x(\cdot)$, $\alpha_{k,i}$ and $\gamma_{k,j}$, $i = 0,\ \cdots,\ m$, $j = 1,\ \cdots,\ s$, are unknown and need to be estimated. Not only does (1.5) introduce a dynamic structure for $\mathrm{cov}(Y_t|\mathcal{F}_{t-1})$, but also reduces the number of unknown parameters from $p_n(p_n+1)/2$ to $p_nq + q^2$ unknown functions and $q + s + m + 1$ unknown parameters.

We remark that model (1.3) is interesting in its own right, since it combines single-index modelling (Carroll *et al.*, 1997, Härdle *et al.*, 1993, Yu and Ruppert, 2002, Xia and Härdle, 2006, Kong and Xia, 2014) and varying coefficient modelling (Fan and Zhang, 1999, 2000, Fan *et al.*, 2003, Sun *et al.*, 2007, Zhang *et al.*, 2009, Li and Zhang, 2011, Sun *et al.*, 2014). In this paper, as a by-product, an estimation procedure for (1.3) is proposed and an iterative algorithm is developed for implementation purposes.

This paper is organised as follows. We begin in Section 2 with a description of the proposed estimation procedure for $\mathrm{cov}(Y_t|\mathcal{F}_{t-1})$. A discussion on bandwidth selection is given in Section 3. In Section 4 we provide asymptotic properties of the estimation procedure. An iterative algorithm to implement the estimation procedure is suggested in Section 5. Using the proposed dynamic structure for covariance matrices and the developed estimation procedure, we outline a process for constructing a portfolio allocation based on the formula for Markowitz's optimal portfolio in

Section 6. The performance of the estimation procedure and portfolio allocation are also assessed by simulation studies in Section 7. In Section 8, we apply the portfolio allocation methodology to a data set consisting of 49 industry portfolios which are freely available from Kenneth French's website. We find that the proposed methodology works surprisingly well. All the detailed proofs are relegated to the appendix.

2 Estimation procedure

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

$$\widehat{\mathrm{cov}}(Y_t|\mathcal{F}_{t-1}) = \hat{\boldsymbol{\Phi}}(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\Sigma}}_x(X_{t-1})\hat{\boldsymbol{\Phi}}(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}})^{\mathrm{T}} + \hat{\boldsymbol{\Sigma}}_{0,t} \tag{2.1}$$

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

Throughout this paper, for any function $f(x)$, we use $\dot{f}(x)$ to denote its derivative. For any functional matrix $F = (f_{ij}(x))$, we define its derivative as $\dot{F} = (\dot{f}_{ij}(x))$. For any integers p and q, we use $\mathbf{0}_{p\times q}$ to denote a $p \times q$ matrix with each entry being 0, and $\mathbf{1}_p$ to denote a p-dimensional vector with each component being 1.

2.1 Estimation of $\boldsymbol{\beta}$

A Taylor expansion gives, for $X_i^{\mathrm{T}}\boldsymbol{\beta}$ in a neighbourhood of $X_j^{\mathrm{T}}\boldsymbol{\beta}$,

$$\boldsymbol{\Phi}(X_i^{\mathrm{T}}\boldsymbol{\beta}) \approx \boldsymbol{\Phi}(X_j^{\mathrm{T}}\boldsymbol{\beta}) + \dot{\boldsymbol{\Phi}}(X_j^{\mathrm{T}}\boldsymbol{\beta})(X_i - X_j)^{\mathrm{T}}\boldsymbol{\beta}$$

and

$$\mathbf{g}(X_i^{\mathrm{T}}\boldsymbol{\beta}) \approx \mathbf{g}(X_j^{\mathrm{T}}\boldsymbol{\beta}) + \dot{\mathbf{g}}(X_j^{\mathrm{T}}\boldsymbol{\beta})(X_i - X_j)^{\mathrm{T}}\boldsymbol{\beta}$$

for $j = 1, \cdots, n-1$. This, together with the idea of least squares estimation, brings us to the following local discrepancy function

$$\begin{aligned} & L(\mathbf{g}_1,\ \boldsymbol{\xi}_1,\ A_1,\ B_1,\ \cdots,\ \mathbf{g}_{n-1},\ \boldsymbol{\xi}_{n-1},\ A_{n-1},\ B_{n-1},\ \boldsymbol{\beta}) \\ = & \sum_{j=1}^{n-1}\sum_{i=2}^{n} \left\|Y_i - \mathbf{g}_j - A_jX_i - (\boldsymbol{\xi}_j + B_jX_i)(X_{i-1} - X_j)^{\mathrm{T}}\boldsymbol{\beta}\right\|^2 K_h((X_{i-1} - X_j)^{\mathrm{T}}\boldsymbol{\beta}), \end{aligned} \tag{2.2}$$

where: $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function; h is a bandwidth; and $\mathbf{g}_j$, $\boldsymbol{\xi}_j$, A_j and B_j are used to denote $\mathbf{g}(X_j^{\mathrm{T}}\boldsymbol{\beta})$, $\dot{\mathbf{g}}(X_j^{\mathrm{T}}\boldsymbol{\beta})$, $\boldsymbol{\Phi}(X_j^{\mathrm{T}}\boldsymbol{\beta})$ and $\dot{\boldsymbol{\Phi}}(X_j^{\mathrm{T}}\boldsymbol{\beta})$ respectively. By minimising

$$L(\mathbf{g}_1,\ \boldsymbol{\xi}_1,\ A_1,\ B_1,\ \cdots,\ \mathbf{g}_{n-1},\ \boldsymbol{\xi}_{n-1},\ A_{n-1},\ B_{n-1},\ \boldsymbol{\beta})$$

under the conditions

$$\|\boldsymbol{\beta}\| = 1, \quad \beta_1 > 0,$$

we use the corresponding value of $\boldsymbol{\beta}$ as the estimator and denote it by $\hat{\boldsymbol{\beta}}$.

2.2 Estimation of $\boldsymbol{\Phi}(\cdot)$ and $\mathbf{g}(\cdot)$

Once an estimate $\hat{\boldsymbol{\beta}}$ has been obtained, the estimators of $\boldsymbol{\Phi}(\cdot)$ and $\mathbf{g}(\cdot)$ can be constructed row by row through a standard univariate varying coefficient model for each component of Y_t. Let

$$\mathbf{g}(\cdot) = (g_1(\cdot),\ \cdots,\ g_{p_n}(\cdot))^{\mathrm{T}}, \quad \boldsymbol{\Phi}(\cdot) = (\mathbf{a}_1(\cdot),\ \cdots,\ \mathbf{a}_{p_n}(\cdot))^{\mathrm{T}}, \quad Y_t = (y_{1,t},\ \cdots,\ y_{p_n,t})^{\mathrm{T}}.$$

By (1.3), and for $k = 1,\ \cdots,\ p_n$, we have the following synthetic univariate varying coefficient model

$$y_{k,t} = g_k(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) + X_t^{\mathrm{T}}\mathbf{a}_k(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) + \epsilon_{kt},$$

for $t = 2,\ \cdots,\ n$. By local linear estimation for standard varying-coefficient models, and for any given u, we have

$$\hat{\mathbf{a}}_k(u) = (I_q,\ \mathbf{0}_{q\times(q+2)})\left(\mathcal{X}^{\mathrm{T}}W\mathcal{X}\right)^{-1}\mathcal{X}^{\mathrm{T}}W\mathbf{y}_k, \quad \hat{g}_k(u) = (\mathbf{0}_{1\times q},\ 1,\ \mathbf{0}_{1\times(q+1)})\left(\mathcal{X}^{\mathrm{T}}W\mathcal{X}\right)^{-1}\mathcal{X}^{\mathrm{T}}W\mathbf{y}_k,$$

where

$$\mathbf{y}_k = (y_{k,2},\ \cdots,\ y_{k,n})^{\mathrm{T}}, \quad \mathcal{X} = \begin{pmatrix} X_2^{\mathrm{T}} & 1 & (X_1^{\mathrm{T}}\hat{\boldsymbol{\beta}} - u) & (X_1^{\mathrm{T}}\hat{\boldsymbol{\beta}} - u)X_2^{\mathrm{T}} \\ \vdots & \vdots & \vdots & \vdots \\ X_n^{\mathrm{T}} & 1 & (X_{n-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - u) & (X_{n-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - u)X_n^{\mathrm{T}} \end{pmatrix},$$

$$W = \mathrm{diag}\left(K_{h_1}(X_1^{\mathrm{T}}\hat{\boldsymbol{\beta}} - u),\ \cdots,\ K_{h_1}(X_{n-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - u)\right),$$

and h_1 is a bandwidth.

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

In order to estimate $E(X_t|X_{t-1} = \mathbf{u})$ and $E(X_tX_t^{\mathrm{T}}|X_{t-1} = \mathbf{u})$, for any given $\mathbf{u}$, we use the local constant estimators

$$\widehat{E}(X_t|X_{t-1} = \mathbf{u}) = \frac{\sum_{t=2}^{n} X_t K_{h_2}(\|X_{t-1} - \mathbf{u}\|)}{\sum_{t=2}^{n} K_{h_2}(\|X_{t-1} - \mathbf{u}\|)}, \tag{2.3}$$

$$\widehat{E}(X_tX_t^{\mathrm{T}}|X_{t-1} = \mathbf{u}) = \frac{\sum_{t=2}^{n} X_tX_t^{\mathrm{T}} K_{h_2}(\|X_{t-1} - \mathbf{u}\|)}{\sum_{t=2}^{n} K_{h_2}(\|X_{t-1} - \mathbf{u}\|)}.$$

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

$$\begin{aligned}\hat{\mathbf{\Sigma}}_x(\mathbf{u}) &= \widehat{E}(X_t X_t^{\mathrm{T}}|X_{t-1}=\mathbf{u}) - \widehat{E}(X_t|X_{t-1}=\mathbf{u})\left\{\widehat{E}(X_t|X_{t-1}=\mathbf{u})\right\}^{\mathrm{T}} \\ &= \{\mathrm{tr}(\mathcal{W})\}^{-2}\mathbf{X}^{\mathrm{T}}\left\{\mathrm{tr}(\mathcal{W})\mathcal{W} - \mathcal{W}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathcal{W}\right\}\mathbf{X} \end{aligned} \tag{2.4}$$

where

$$\mathbf{X} = (X_2,\ \cdots,\ X_n)^{\mathrm{T}}, \quad \mathcal{W} = \mathrm{diag}(K_{h_2}(\|X_1-\mathbf{u}\|),\ \cdots,\ K_{h_2}(\|X_{n-1}-\mathbf{u}\|)),$$

and h_2 is a bandwidth.

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

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

$$r_{k,t} = \hat{\epsilon}_{k,t} = y_{k,t} - \hat{g}_k(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) - X_t^{\mathrm{T}}\hat{\mathbf{a}}_k(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}).$$

By (1.4), we have the following synthetic GARCH model

$$\sigma_{kt}^2 = \alpha_{k,0} + \sum_{i=1}^{m}\alpha_{k,i}r_{k,t-i}^2 + \sum_{j=1}^{s}\gamma_{k,j}\sigma_{k,t-j}^2, \quad t = 2,\ \cdots,\ n \tag{2.5}$$

which is equivalent to

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

where $\eta_{kt} = r_{k,t}^2 - \sigma_{kt}^2$, $\gamma_{k,i} = 0$ when $i > s$, and $\alpha_{k,i} = 0$ when $i > m$.

Once $\alpha_{k,i}$ and $\gamma_{k,j}$ have been estimated, by substituting them into (2.5) and setting $\sigma_{kl}^2 = r_{k,l}^2$ for $l \le \max(m,s)$, we can obtain an estimator $\hat{\sigma}_{kt}^2$ of σ_{kt}^2 and hence an estimator $\hat{\mathbf{\Sigma}}_{0,t}$ of $\mathbf{\Sigma}_{0,t}$.

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

$$Q_{k,n}(\boldsymbol{\theta}_k) = n^{-1}\sum_{t=2}^{n}\left\{\frac{r_{k,t}^2}{\sigma_{k,t}^2(\boldsymbol{\theta}_k)} + \log\sigma_{k,t}^2(\boldsymbol{\theta}_k)\right\} \tag{2.6}$$

where $\sigma_{k,t}^2(\boldsymbol{\theta}_k)$ are recursively defined by (2.5) with initial values being either

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

or

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

By minimising $Q_{k,n}(\boldsymbol{\theta}_k)$ with respect to $\boldsymbol{\theta}_k$ on a compact set $\mathbf{\Lambda}$ defined in (B3) in Appendix A, we use the minimiser $\hat{\boldsymbol{\theta}}_k$ to estimate $\boldsymbol{\theta}_k$.

3 Bandwidth selection

The choice of the bandwidth h, used in the estimation of $\boldsymbol{\beta}$, is not crucial. According to some numerical analysis not presented in this paper for brevity, the accuracy of the estimator $\hat{\boldsymbol{\beta}}$ is not very sensitive to h, as long as h is within in a reasonable range. In the computational algorithm for estimating $\boldsymbol{\beta}$, see Section 5, we recommend choosing a bandwidth h equal to around 20% of the following range

$$\max\{X_1^{\mathrm{T}}\tilde{\boldsymbol{\beta}}, \cdots, X_n^{\mathrm{T}}\tilde{\boldsymbol{\beta}}\} - \min\{X_1^{\mathrm{T}}\tilde{\boldsymbol{\beta}}, \cdots, X_n^{\mathrm{T}}\tilde{\boldsymbol{\beta}}\} \tag{3.1}$$

where $\tilde{\boldsymbol{\beta}}$ is a randomly chosen initial estimate of $\boldsymbol{\beta}$. We update h on subsequent iterations by replacing $\tilde{\boldsymbol{\beta}}$ in (3.1) with the most recent estimate of $\boldsymbol{\beta}$. This approach is employed in the simulation studies and real data analysis of this paper.

We now focus on the selection of the bandwidth h_1, used in the estimation of $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$. The proposed bandwidth selection is based on a k-nearest neighbours bandwidth with k being selected by cross-validation. We define the cross-validation statistic by

$$\mathrm{CV}(k) = \sum_{t=n-M}^{n} \left\| Y_t - \hat{\mathbf{g}}^{(t-1)}(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\Phi}}^{(t-1)}(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}})X_t \right\| \tag{3.2}$$

where $\hat{\mathbf{g}}^{(t-1)}(\cdot)$ and $\hat{\boldsymbol{\Phi}}^{(t-1)}(\cdot)$ are the respective estimators of $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$ using a k-nearest neighbours bandwidth based on $(X_l^{\mathrm{T}},\ Y_l^{\mathrm{T}})$, $l = 1, \cdots, t-1$, and where M is a look-back integer parameter such that $M < n-1$.

Hence, denoting the k that minimises $\mathrm{CV}(k)$ by $\hat{k}$, we use a $\hat{k}$-nearest neighbours bandwidth in the estimation of $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$. The bandwidth h_2 in the estimation of $\boldsymbol{\Sigma}_x(\cdot)$ or $E(X_t|X_{t-1} = \mathbf{u})$ can also be selected by cross-validation in a similar way.

4 Asymptotic properties

In this section, we are going to present the asymptotic properties of the proposed estimators. We first introduce the following notation which will be used throughout this paper. For any matrix $\mathbf{A} = (a_{ij})_{m\times N}$, we use $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ to denote respectively the smallest and largest eigenvalues of $\mathbf{A}$. The trace of $\mathbf{A}$ is denoted by $\mathrm{tr}(\mathbf{A})$, the Frobenius norm of $\mathbf{A}$ by $\|\mathbf{A}\|_F$, and the spectral norm (also called operator norm) and element-wise norm by

$$\|\mathbf{A}\| = \sqrt{\lambda_{\max}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)}, \quad \|\mathbf{A}\|_\infty = \max_{\substack{1\le i\le m\\ 1\le j\le N}} |a_{ij}|$$

respectively. We also define

$$\mathbf{U}_n = \frac{1}{np_n}\sum_{i=2}^{n}\sum_{k=1}^{p_n} f(X_i^{\mathrm{T}}\boldsymbol{\beta})\{X_{i-1} - E(X_i|X_{i-1}^{\mathrm{T}}\boldsymbol{\beta})\}\{\dot{g}_k(X_{i-1}^{\mathrm{T}}\boldsymbol{\beta}) + \dot{\mathbf{a}}_k(X_{i-1}^{\mathrm{T}}\boldsymbol{\beta})X_i\}\epsilon_{k,i}$$

and

$$\mathbf{V}_p = p_n^{-1}\sum_{k=1}^{p_n} E\left(f(X_1^{\mathrm{T}}\boldsymbol{\beta})\{X_1 - E(X_2|X_1^{\mathrm{T}}\boldsymbol{\beta})\}^{\otimes 2}\left\{\dot{g}_k(X_1^{\mathrm{T}}\boldsymbol{\beta}) + \dot{\mathbf{a}}_k(X_1^{\mathrm{T}}\boldsymbol{\beta})X_2\right\}^2\right).$$

Theorem 1. *Under assumptions (A1 - A5), (B1 - B4), (C1) and (C3) in Appendix A, there exists $C > 0$ and a small $\varepsilon > 0$ such that*

(I)

$$P\left\{\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} - \mathbf{V}_p^{-1}\mathbf{U}_n\right\| > C\left(h^3 + \frac{\log(n)}{nh}\right)\right\} \leq O\left(\frac{1}{n^{1+\varepsilon}}\right);$$

(II)

$$P\left\{\sup_{z\in\mathcal{Z}}\|\widehat{\mathbf{g}}(z) - \mathbf{g}(z)\|_\infty > C\left(h_1^2 + \sqrt{\frac{\log(n)}{nh_1}}\right)\right\} \leq O\left(\frac{1}{n^{1+\varepsilon}}\right);$$

(III)

$$P\left\{\sup_{z\in\mathcal{Z}}\left\|\widehat{\boldsymbol{\Phi}}(z) - \boldsymbol{\Phi}(z)\right\|_\infty > C\left(h_1^2 + \sqrt{\frac{\log(n)}{nh_1}}\right)\right\} \leq O\left(\frac{1}{n^{1+\varepsilon}}\right);$$

(IV)

$$P\left\{\sup_{1\leq k\leq p_n}\left\|\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\right\| > C\left(h_1^2 + \sqrt{\frac{\log(n)}{nh_1}}\right)\right\} \leq O\left(\frac{1}{n^{1+\varepsilon}}\right),$$

where $\mathcal{Z}$ is a compact subset of the range of $X_t^{\mathrm{T}}\boldsymbol{\beta}$.

Remark 1. Theorem 1 shows that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = o_P(n^{-1/2})$ when p_n diverges to ∞ as $n \to \infty$, provided that $\|\mathbf{U}_n\| = o_P(n^{-1/2})$. It indicates that the index $\boldsymbol{\beta}$ is estimated with a rate faster than the normal rate $n^{-1/2}$, which is the optimal rate if p_n is fixed. This is known as a 'blessing of high dimensionality'.

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

$$\left\|\hat{M} - M\right\|_{\boldsymbol{\Sigma}} = p_n^{-1/2}\left\|M^{-1/2}\left\{\hat{M} - M\right\}M^{-1/2}\right\|_F.$$

To facilitate our presentation, we focus on the convergence of $\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n) - \mathrm{cov}(Y_{n+1}|\mathcal{F}_n)$, after obtaining the data $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$.

Theorem 2. *Under assumptions (A1 - A5), (B1 - B4) and (C1 - C4) in Appendix A, there exist $C > 0$ and $\varepsilon > 0$ such that, with probability at least $1 - n^{-(1+\varepsilon)}$,*

$$\|\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n) - \mathrm{cov}(Y_{n+1}|\mathcal{F}_n)\|_{\mathbf{\Sigma}}^2 \leq p_n C \left\{ h_1^8 + \left(\frac{\log n}{n h_1} \right)^2 \right\} + C \left(h_1^4 + \frac{\log n}{n h_1} \right) + p_n^{-1} C \left(h_2^4 + \frac{\log n}{n h_2^q} \right).$$

Fan, Fan and Lv (2008) and Fan, Liao and Mincheva (2011) showed an estimator of a covariance matrix based on a certain structure would achieve a higher convergence rate than the sample covariance matrix. Theorem 2 tells us the same story. There are three terms to measure the accuracy of $\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n) - \mathrm{cov}(Y_{n+1}|\mathcal{F}_n)$. The first two terms tell us how the nonparametric smoothing steps in estimating $\mathbf{\Phi}(\cdot)$ affect the performance of $\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n)$, and the third term evaluates the influence of conditional covariance matrix $\mathbf{\Sigma}_x(X_n)$. It turns out that even though $q-$dimensional smoothing is required, its effect is small and often negligible if p_n is large.

5 Computational algorithm

To implement the proposed estimation procedure for $\mathrm{cov}(Y_t|\mathcal{F}_{t-1})$, the hardest part is to compute an estimate of $\boldsymbol{\beta}$, which is equivalent to finding the minimum of

$$L(\mathbf{g}_1,\ \boldsymbol{\xi}_1,\ A_1,\ B_1,\ \cdots,\ \mathbf{g}_{n-1},\ \boldsymbol{\xi}_{n-1},\ A_{n-1},\ B_{n-1},\ \boldsymbol{\beta})$$

under the conditions

$$\|\boldsymbol{\beta}\| = 1, \quad \beta_1 > 0.$$

We now introduce the proposed iterative algorithm which can be used to do this minimisation. Let

$$\begin{aligned} & \mathcal{Q}(\mathbf{g}_1,\ \boldsymbol{\xi}_1,\ A_1,\ B_1,\ \cdots,\ \mathbf{g}_{n-1},\ \boldsymbol{\xi}_{n-1},\ A_{n-1},\ B_{n-1},\ \boldsymbol{\beta},\ \mathbf{b}) \\ = & \sum_{j=1}^{n-1} \sum_{i=2}^{n} \left\| Y_i - \mathbf{g}_j - A_j X_i - (\boldsymbol{\xi}_j + B_j X_i)(X_{i-1} - X_j)^{\mathrm{T}} \boldsymbol{\beta} \right\|^2 K_h((X_{i-1} - X_j)^{\mathrm{T}} \mathbf{b}), \end{aligned}$$

which is $L(\mathbf{g}_1,\ \boldsymbol{\xi}_1,\ A_1,\ B_1,\ \cdots,\ \mathbf{g}_{n-1},\ \boldsymbol{\xi}_{n-1},\ A_{n-1},\ B_{n-1},\ \boldsymbol{\beta})$ with the $\boldsymbol{\beta}$ in the kernel function being replaced by $\mathbf{b}$. First of all, randomly choose an initial estimate for $\boldsymbol{\beta}$, denoted by $\tilde{\boldsymbol{\beta}}$, such that $\|\tilde{\boldsymbol{\beta}}\| = 1$ and the first component of $\tilde{\boldsymbol{\beta}}$ is positive. Then, iterate between the following two steps until convergence:

(Step 1) If this is the first iteration, let $\boldsymbol{\beta}_0 = \tilde{\boldsymbol{\beta}}$. Otherwise, set $\boldsymbol{\beta}_0$ equal to the $\hat{\boldsymbol{\beta}}$ obtained from Step 2 of the previous iteration. Minimise

$$L(\mathbf{g}_1,\ \boldsymbol{\xi}_1,\ A_1,\ B_1,\ \cdots,\ \mathbf{g}_{n-1},\ \boldsymbol{\xi}_{n-1},\ A_{n-1},\ B_{n-1},\ \boldsymbol{\beta}_0)$$

with respect to $\mathbf{g}_1,\ \boldsymbol{\xi}_1,\ A_1,\ B_1,\ \cdots,\ \mathbf{g}_{n-1},\ \boldsymbol{\xi}_{n-1},\ A_{n-1}$ and B_{n-1}, and denote the minimiser by $\hat{\mathbf{g}}_1, \hat{\boldsymbol{\xi}}_1, \hat{A}_1, \hat{B}_1, \cdots, \hat{\mathbf{g}}_{n-1}, \hat{\boldsymbol{\xi}}_{n-1}, \hat{A}_{n-1}$ and $\hat{B}_{n-1}$.

(Step 2) Minimise

$$\mathcal{Q}(\hat{\mathbf{g}}_1,\ \hat{\boldsymbol{\xi}}_1,\ \hat{A}_1,\ \hat{B}_1,\ \cdots,\ \hat{\mathbf{g}}_{n-1},\ \hat{\boldsymbol{\xi}}_{n-1},\ \hat{A}_{n-1},\ \hat{B}_{n-1},\ \boldsymbol{\beta},\ \boldsymbol{\beta}_0)$$

with respect to $\boldsymbol{\beta}$. Denote the minimiser by $\check{\boldsymbol{\beta}}$, and define $\hat{\boldsymbol{\beta}} = \check{\boldsymbol{\beta}}/\|\check{\boldsymbol{\beta}}\|$ when the first component of $\check{\boldsymbol{\beta}}$ is positive and $\hat{\boldsymbol{\beta}} = -\check{\boldsymbol{\beta}}/\|\check{\boldsymbol{\beta}}\|$ otherwise.

The $\hat{\boldsymbol{\beta}}$ resulting from the convergence is the final estimate of $\boldsymbol{\beta}$.

The proposed iterative algorithm is easy to implement as both minimisers in Step 1 and Step 2 have a closed form. Once an estimate of $\boldsymbol{\beta}$ is obtained, the remaining computation of $\mathrm{cov}(Y_t|\mathcal{F}_{t-1})$ becomes straightforward.

6 Portfolio allocation

In this section, we will briefly describe the construction of an estimated optimal portfolio allocation based on the proposed dynamic structure and the associated estimation procedure. Since the formula for optimal portfolio allocation contains $E(Y_t|\mathcal{F}_{t-1})$ we shall introduce its estimator $\widehat{E}(Y_t|\mathcal{F}_{t-1})$ first. By taking conditional expectation of (1.3), we have

$$E(Y_t|\mathcal{F}_{t-1}) = \mathbf{g}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + \boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})E(X_t|X_{t-1}).$$

Therefore, we use

$$\widehat{E}(Y_t|\mathcal{F}_{t-1}) = \hat{\mathbf{g}}(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{\Phi}}(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}})\hat{E}(X_t|X_{t-1}) \tag{6.1}$$

to estimate $E(Y_t|\mathcal{F}_{t-1})$ where $\hat{E}(X_t|X_{t-1})$ is defined in (2.3).

Our estimated optimal portfolio allocation builds on the mean-variance optimal portfolio by Markowitz (1952, 1959). The allocation vector $\mathbf{w}$ of p_n risky assets, to be held between times $t-1$ and t, is defined as the solution to

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

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

$$\hat{\mathbf{w}} = \frac{c_3 - c_2\delta}{c_1c_3 - c_2^2}\widehat{\mathrm{cov}}(Y_t|\mathcal{F}_{t-1})^{-1}\mathbf{1}_{p_n} + \frac{c_1\delta - c_2}{c_1c_3 - c_2^2}\widehat{\mathrm{cov}}(Y_t|\mathcal{F}_{t-1})^{-1}\widehat{E}(Y_t|\mathcal{F}_{t-1}), \tag{6.2}$$

where

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

7 Simulation studies

In this section, we are going to use a simulated example to show how well the proposed estimation procedure and portfolio allocation works. We shall use $a_{i,j}(\cdot)$ to denote the entry corresponding to the ith row and jth column of $\mathbf{\Phi}(\cdot)$.

We generate 1000 data sets from model (1.3) together with (1.4). We repeat this using the following combinations of n and p_n: $\{n = 1000, p_n = 50\}$, $\{n = 1000, p_n = 100\}$, $\{n = 2000, p_n = 50\}$ and $\{n = 2000, p_n = 100\}$. We set

$$q = 4, \quad m = 1, \quad s = 1, \quad \boldsymbol{\beta} = \frac{1}{3}(1,\ 2,\ 0,\ 2)^{\mathrm{T}}.$$

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

$$\alpha_{0,k} = 0.5, \quad \alpha_{1,k} = 0.1, \quad \beta_{1,k} = 0.1, \quad g_k(z) = \Xi_{0,k} + 3\exp(-z^2), \quad a_{k,1}(z) = \Xi_{1,k} + 0.8z,$$

$$a_{k,2}(z) = \Xi_{2,k}, \quad a_{k,3}(z) = \Xi_{3,k} + 1.5\sin(\pi z), \quad a_{k,4}(z) = \Xi_{4,k},$$

where $\Xi_{j,k}$ are some fixed parameters for $j = 0,\ \cdots,\ d$ and $k = 1,\ \cdots,\ p_n$. In order to define $\Xi_{j,k}$, we simulate them independently from a uniform distribution on $[-1,\ 1]$, and use these same values throughout all simulations. For $t = 1,\ \cdots,\ n+1$, we generate X_t independently from a uniform distribution on $[-1, 1]^q$, Z_t from p_n-variate standard normal distribution, and $\boldsymbol{\epsilon}_t$ through $\boldsymbol{\epsilon}_t = \mathbf{\Sigma}_{0,t}^{1/2} Z_t$. Once both X_t and $\boldsymbol{\epsilon}_t$ have been generated, Y_t can be generated through (1.3) for $t = 1,\ \cdots,\ n+1$.

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

$$R(\hat{\mathbf{w}}) = \hat{\mathbf{w}}^{\mathrm{T}} Y_{n+1} \tag{7.1}$$

of a portfolio allocation $\hat{\mathbf{w}}$ formed at time n based on data $(X_t^{\mathrm{T}},\ Y_t^{\mathrm{T}})$, $t = 1,\ \cdots,\ n$. In order to evaluate the performance of an estimator $\hat{M}$ of matrix M we use the following metric

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

We also use the Sharpe ratio

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

to evaluate the performance of $\hat{\mathbf{w}}$, where $\mathrm{SD}\left\{R(\hat{\mathbf{w}})\right\}$ is the standard deviation of $R(\hat{\mathbf{w}})$. We assume a zero risk-free rate for simplicity.

We first examine how well the estimation procedure works. We estimate $\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)$, and use $\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n)^{-1}$ to estimate $\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1}$. The kernel function in the estimation procedure is taken to be the Epanechnikov kernel $K(u) = 0.75(1-u^2)_+$, and the bandwidths are selected by the methodology described in Section 3. The results, presented in Tables 1 and 2, show both $\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n)$ and $\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n)^{-1}$ work very well.

Table 1: Mean and Standard Deviation of $\Delta\left(\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n),\ \mathrm{cov}(Y_{n+1}|\mathcal{F}_n)\right)$

	$n = 1000$ $p_n = 50$	$n = 1000$ $p_n = 100$	$n = 2000$ $p_n = 50$	$n = 2000$ $p_n = 100$
$E(D)$	0.183	0.189	0.136	0.141
$\mathrm{SD}(D)$	0.046	0.049	0.034	0.035

In this table, $D = \Delta\left(\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n),\ \mathrm{cov}(Y_{n+1}|\mathcal{F}_n)\right)$, and $\mathrm{SD}(D)$ is the standard deviation of D.

Table 2: Mean and Standard Deviation of $\Delta\left(\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n)^{-1},\ \mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1}\right)$

	$n = 1000$ $p_n = 50$	$n = 1000$ $p_n = 100$	$n = 2000$ $p_n = 50$	$n = 2000$ $p_n = 100$
$E(D_1)$	0.114	0.105	0.078	0.070
$\mathrm{SD}(D_1)$	0.017	0.013	0.012	0.009

In this table, $D_1 = \Delta\left(\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n)^{-1},\ \mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1}\right)$, and $\mathrm{SD}(D_1)$ is the standard deviation of D_1.

We now examine the performance of the proposed portfolio allocation, using a target return $\delta = 1\%$, by computing the return as described in (7.1). In order to see how much gain can be made by making use of the dynamic structure, we make a comparison with portfolio allocations based on Markowitz's formula but where the covariance matrix is estimated using the sample covariance matrix and also the estimator proposed by Fan, Fan and Lv (2008). The mean, standard deviation and Sharpe ratio of the returns are presented in Table 3. For each situation discussed, we see the Sharpe ratio of the proposed portfolio allocation is much bigger than the other two portfolio allocations. This suggests there is significant gain from making use of the dynamic structure of the covariance matrix.

Table 3: Means, Standard Deviations and Sharpe Ratios

	$n = 1000$ $p_n = 50$	$n = 1000$ $p_n = 100$	$n = 2000$ $p_n = 50$	$n = 2000$ $p_n = 100$
$E\{R(\hat{\mathbf{w}})\}$	0.99%	1.01%	1.03%	1.03%
$E\{R(\hat{\mathbf{w}}_1)\}$	0.96%	0.96%	1.02%	1.02%
$E\{R(\hat{\mathbf{w}}_2)\}$	0.96%	0.96%	1.02%	1.02%
$\mathrm{SD}\{R(\hat{\mathbf{w}})\}$	0.40%	0.28%	0.39%	0.27%
$\mathrm{SD}\{R(\hat{\mathbf{w}}_1)\}$	1.02%	1.03%	1.03%	1.02%
$\mathrm{SD}\{R(\hat{\mathbf{w}}_2)\}$	0.99%	0.97%	1.02%	1.00%
$\mathrm{SR}(\hat{\mathbf{w}})$	2.49	3.57	2.63	3.83
$\mathrm{SR}(\hat{\mathbf{w}}_1)$	0.94	0.93	0.99	1.00
$\mathrm{SR}(\hat{\mathbf{w}}_2)$	0.97	0.99	1.00	1.02

In this table we denote the proposed portfolio allocation by $\hat{\mathbf{w}}$, the portfolio allocation formed by Markowitz's formula using the sample covariance matrix by $\hat{\mathbf{w}}_1$, and the portfolio allocation formed by Markowitz's formula using the estimated covariance matrix from Fan, Fan and Lv (2008) by $\hat{\mathbf{w}}_2$.

8 Real data analysis

In this section, we are going to apply the dynamic structure for covariance matrices to a real data set. We use the term *Face* (Factor model with an Adaptive-varying-coefficient-model structure Covariance matrix Estimator) to denote the proposed portfolio allocation. This name was chosen because the estimator will 'face' the markets today based on what happened yesterday and adapt according to the dynamic structure. We compare Face with the allocation based on the sample covariance matrix (denoted by *Sam*), and the allocation proposed by Fan, Fan and Lv (2008) (denoted by *Fan*). In all three cases, we use the same target return $\delta = 1\%$. We also make a comparison with the market portfolio (denoted by *Market*) since this aids as an important benchmark indicating whether we are in a bull or bear market. In this section, the kernel function used in the construction of Face is still taken to be the Epanechnikov kernel, and the bandwidths are selected by the method described in Section 3.

All data used can be freely downloaded from Kenneth French's website `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html` and was accessed on 2nd April 2015. The response variable Y_t is chosen to be the vector of the daily returns of $p_n = 49$ industry portfolios (value weighted) minus the risk-free rate. The observable factors $x_{1,t}$, $x_{2,t}$ and $x_{3,t}$ are taken to be the market, size and value factors respectively from the Fama-French three-factor model. The labelling along with a brief description of $Y_t = (y_{1,t},\ \cdots, y_{49,t})^{\mathrm{T}}$ and $X_t = (x_{1,t},\ x_{2,t},\ x_{3,t})^{\mathrm{T}}$

can be found in Table 4 and Table 5 respectively.

There are various advantages of using the portfolio returns for $y_{k,t}$ as opposed to using individual stocks: we avoid having to merge different sources of data; we avoid survivorship bias (where we only picked companies that did not go bankrupt); and we attempt to avoid company specific risk. A further benefit is that the results we give are entirely reproducible since the data is free and presented in a spreadsheet format.

To have a better idea about what the data is like, we plot the observations from 3rd January 1995 to 31st December 2014 of the three factors and the risk-free rate in Figure 1, and the first four components of Y_t in Figure 2 corresponding to the industrial sectors Agriculture, Food Products, Candy & Soda, and Beer & Liquor. The plots show clearly that there are periods of large volatility around the 2008-2009 financial crisis. We will see Face performs reasonably well even during that period, whilst the others do not.

We compare the three portfolio allocations, (Face, Sam and Fan), along with the market portfolio, year by year from 1995 to 2014 using a simple trading strategy. For each year we trade on each trading day, which is approximately $T = 252$ trading days per year. At the beginning of each year we assume we have an initial balance of 100 pounds. Although this initial choice is arbitrary, it is a useful way of comparing the performance during the course of a year. We assume no transaction costs, allow for short selling, and assume that all possible portfolio allocations are attainable. Our trading strategy consists of forming a portfolio allocation $\hat{\mathbf{w}}$ the end of each trading day and holding it until the end of the next trading day. Between day $t-1$ and day t, we obtain the portfolio return

$$R_t(\hat{\mathbf{w}}) = \hat{\mathbf{w}}^{\mathrm{T}} Y_t$$

where $\hat{\mathbf{w}}$ is formed based on $(X_{t-j}^{\mathrm{T}},\ Y_{t-j}^{\mathrm{T}})$, $j = 1,\ \cdots,\ n$, for some look-back integer n. With the realised returns $R_t(\hat{\mathbf{w}})$, $t = 1,\ \cdots,\ T$, we can calculate the annualized Sharpe ratio

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

where

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

and $R_{f,t}$ is the risk-free rate on day t. Hence, for each year, and for each of the four trading strategies, we compute an annualized Sharpe ratio and the balance at the end of the final trading day of the year. We repeat this using $n = 100$, 300, and 500. From the the annualized Sharpe ratios presented in Figure 4 and the balances in Table 6, it is clear that Face performs significantly better than the other three.

We remark that although Face, Sam and Fan are all constructed based on Markowitz's formula, the difference between them lies in the way to estimate the covariance matrix of returns, which appears in Markowitz's formula. Both Sam and Fan do not take into account the dynamic feature of the covariance matrix in their estimation, but Face does. This is the fundamental reason why Face performs significantly better than Sam and Fan. One may argue that if Sam and Fan used fewer observations in their moving window to estimate the covariance matrix they would start to take the dynamic feature into account, potentially improving their performance. However when constructing Face, Sam and Fan, we tried a variety of n, ranging from 100 to 500, and found Face always performs better. This suggests that even if Sam and Fan only use the observations in a carefully chosen moving window, Face still outperforms them.

To have a tangible idea about whether the covariance matrix is dynamic or not, we plot the estimated intercept and coefficients of $x_{1,t}$, $x_{2,t}$ and $x_{3,t}$, interpreted as the impact of the factors, for each of the first four components of Y_t in Figure 3. One can see that these coefficients are dynamic rather than constant, which implies the covariance matrix is also dynamic.

It is interesting to have a closer look at the performances of the four strategies in the volatile time period 2007-2009 during which the financial crisis took place. Still assuming an initial balance of 100 pounds at the start of each year, and using $n = 500$, we plot the balances at the end of each trading day in Figure 5. During 2007, Face, Sam and Fan all perform reasonably well, with Face slightly better. The market does not make much profit, and is beaten by the other three. In 2008, Face continuously does well whilst the other three do not make profit at all. In 2009, although Face does not do very well during some time periods, it adapts to the market change quickly and almost breaks even. The reason that Face can adapt to market change quickly is because it takes into account the dynamic feature of the covariance matrix of returns. On the other hand, both Sam and Fan do very poorly, and in fact they almost lose all their money at the end of the year. In 2009, the market performs best, but still with very little profit.

Table 4: Description of the 49 industry portfolios

k	$y_{k,t}$	Industry name	k	$y_{k,t}$	Industry name
1	Agric	Agriculture	26	Guns	Defense
2	Food	Food Products	27	Gold	Precious Metals
3	Soda	Candy & Soda	28	Mines	Industrial Metal Mining
4	Beer	Beer & Liquor	29	Coal	Coal
5	Smoke	Tobacco Products	30	Oil	Petroleum and Natural Gas
6	Toys	Recreation	31	Util	Utilities
7	Fun	Entertainment	32	Telcm	Communication
8	Books	Printing and Publishing	33	PerSv	Personal Services
9	Hshld	Consumer Goods	34	BusSv	Business Services
10	Clths	Apparel	35	Hardw	Computers
11	Hlth	Healthcare	36	Softw	Computer Software
12	MedEq	Medical Equipment	37	Chips	Electronic Equipment
13	Drugs	Pharmaceutical Products	38	LabEq	Measuring and Control Equipment
14	Chems	Chemicals	39	Paper	Business Supplies
15	Rubbr	Rubber and Plastic Products	40	Boxes	Shipping Containers
16	Txtls	Textiles	41	Trans	Transportation
17	BldMt	Construction Materials	42	Whlsl	Wholesale
18	Cnstr	Construction	43	Rtail	Retail
19	Steel	Steel Works Etc	44	Meals	Restaurants, Hotels, Motels
20	FabPr	Fabricated Products	45	Banks	Banking
21	Mach	Machinery	46	Insur	Insurance
22	ElcEq	Electrical Equipment	47	RlEst	Real Estate
23	Autos	Automobiles and Trucks	48	Fin	Trading
24	Aero	Aircraft	49	Other	Almost Nothing
25	Ships	Shipbuilding, Railroad Equipment			

This table gives the labelling and a brief description of industrial sectors which form the 49 Industry Portfolios data set. Precise details of their construction are given on Kenneth French's website.

Table 5: Description of the Fama and French factors

j	Name of $x_{j,t}$	Description
1	Market factor	Return on the market minus the risk-free rate
2	Size factor	Excess returns of small caps over big caps
3	Value factor	Excess returns of value stocks over growth stocks

This table gives the labelling and a brief description of market, size and value factors from the Fama-French factors data set. Precise details of their construction are given on Kenneth French's website.

Figure 1: Returns plots of factors and the risk-free rate R_f

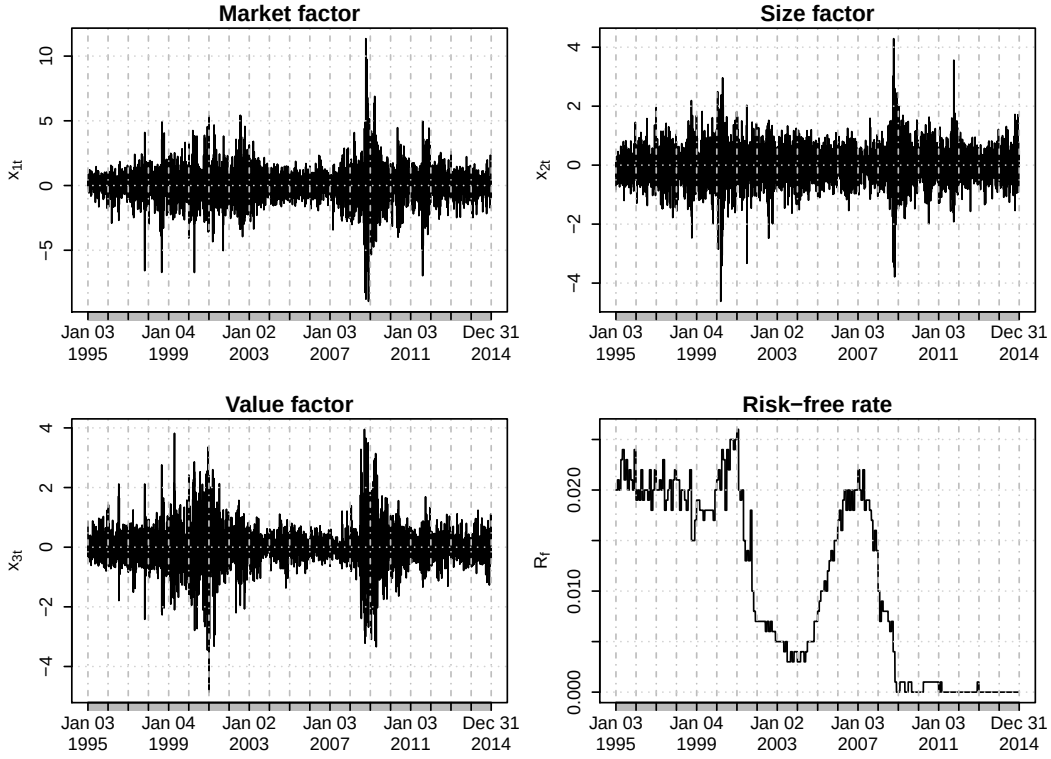


Figure 2: Returns plots of $y_{1,t}$, $y_{2,t}$, $y_{3,t}$, and $y_{4,t}$.

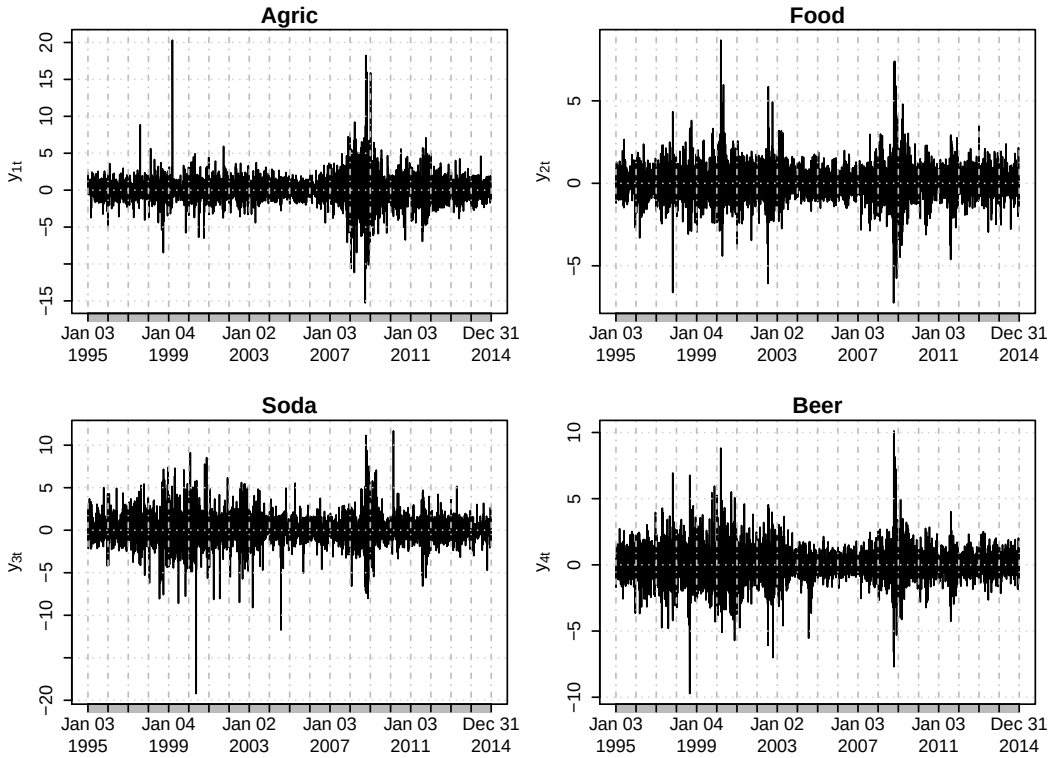
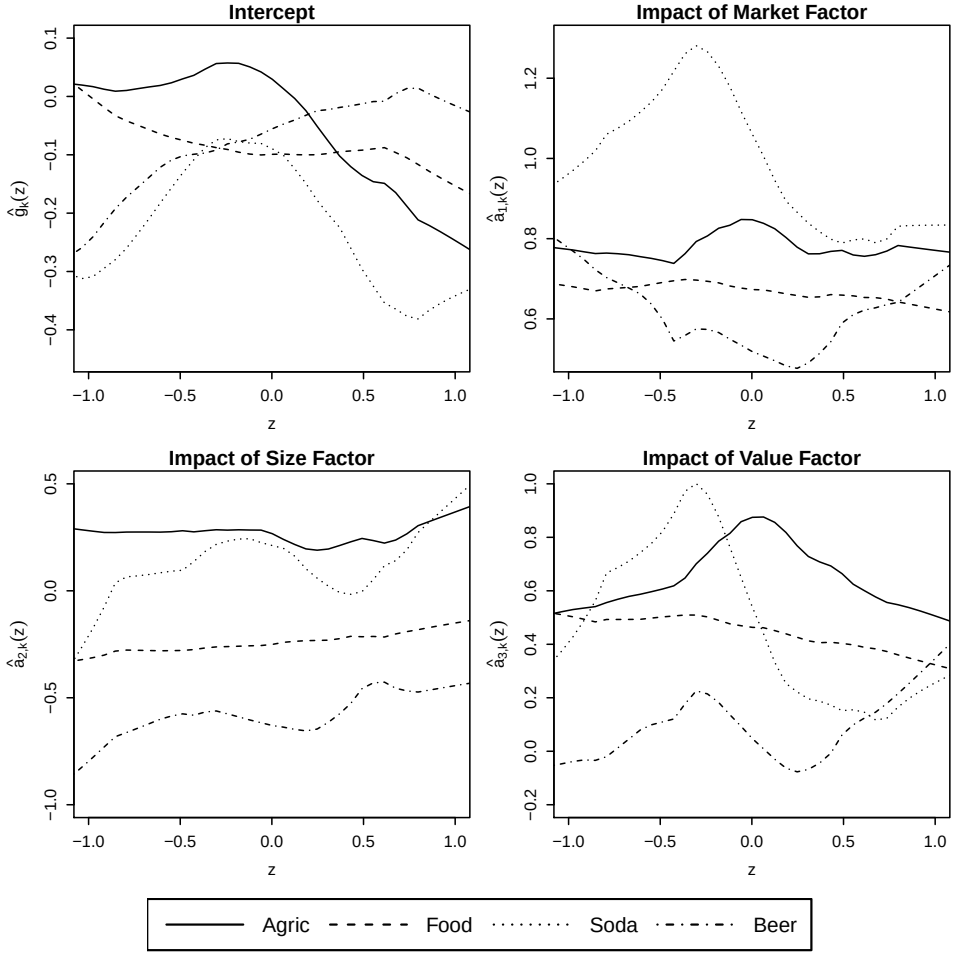
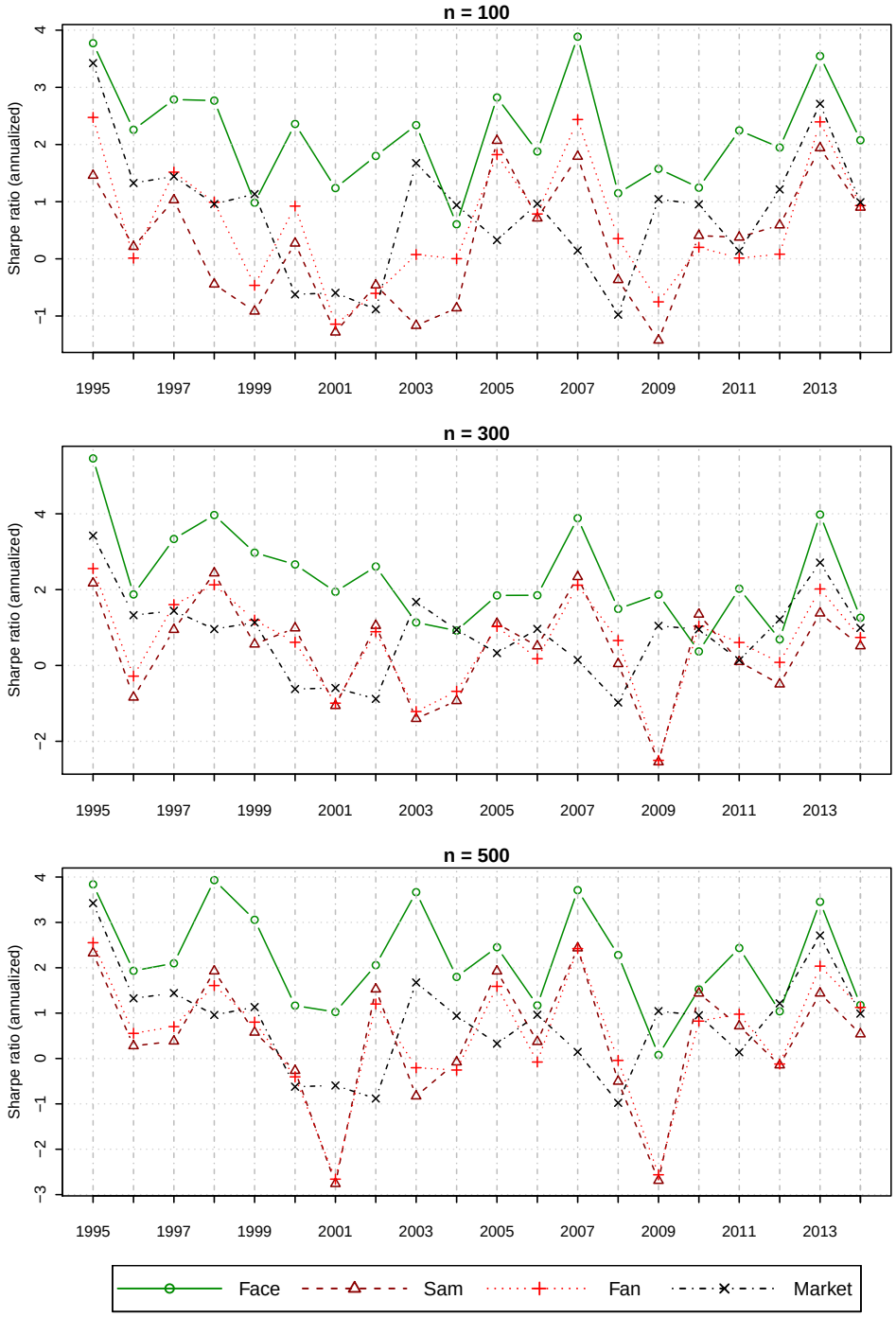


Figure 3: Estimated coefficient functions for industry portfolios 1-4



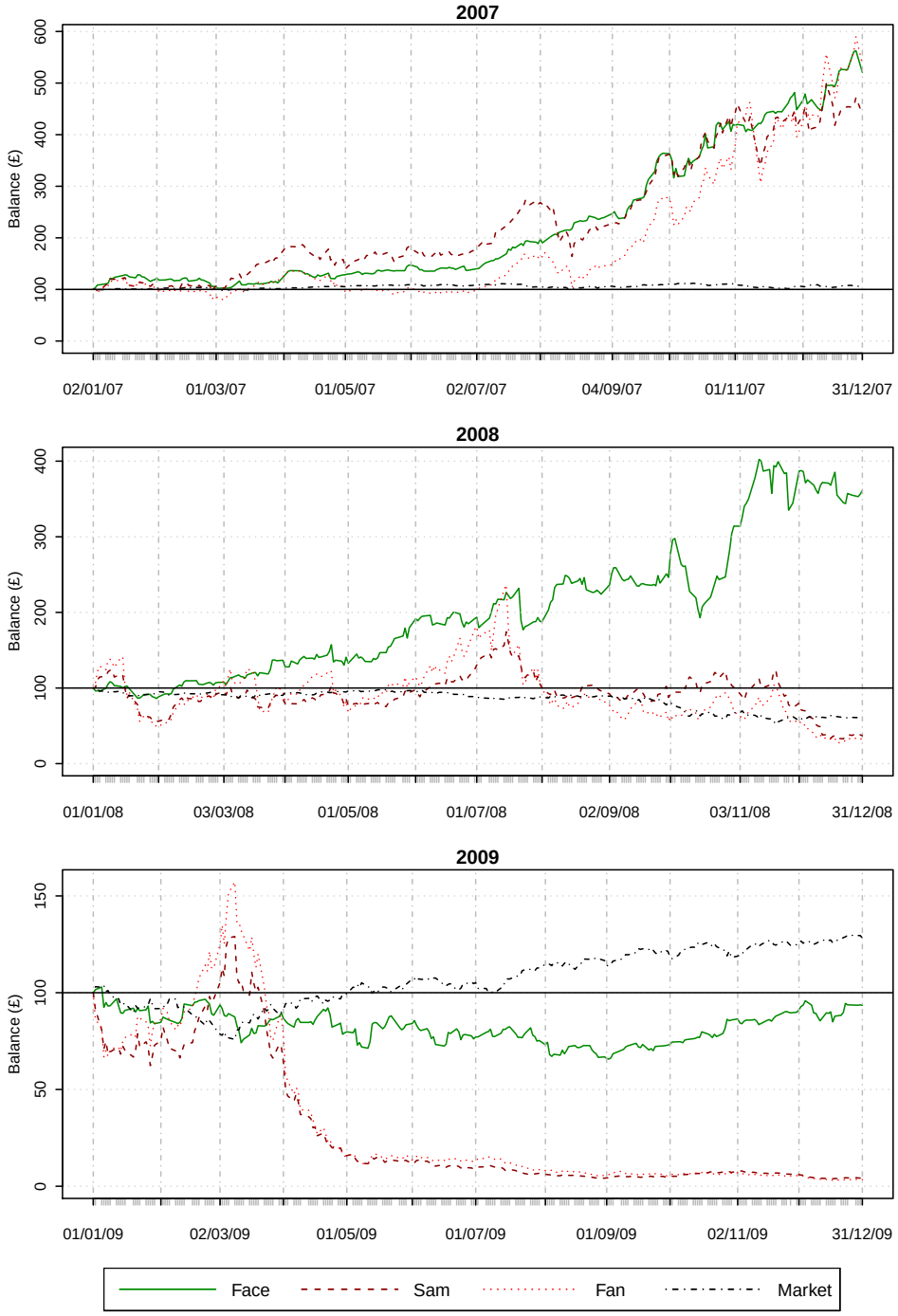
This figure shows the estimated intercept and coefficient functions for the market, size and value factors, for the first four industry portfolios (Agriculture, Food Products, Candy & Soda, and Beer & Liquor) on the first day of trading.

Figure 4: Annualized Sharpe Ratios



This figure shows the performance of the four trading strategies (Face, Sam, Fan and Market) in terms of the annualized Sharpe ratio, using different sample sizes $n = 100$, $n = 300$ *and* $n = 500$.

Figure 5: Trading strategies during the financial crisis



This figure shows the performance of the four trading strategies (Face, Sam, Fan and Market) using $n = 500$ *during 2007, 2008 and 2009 in terms of the end of day balances, assuming an initial balance of 100 pounds at the start of each year.*

Table 6: Comparison of Balances of Trading Strategies

Year	Market	$n = 100$			$n = 300$			$n = 500$		
		Face	Sam	Fan	Face	Sam	Fan	Face	Sam	Fan
1995	137	224	164	216	541	277	347	423	380	466
1996	121	159	101	96	184	56	72	212	95	115
1997	131	179	138	155	303	146	207	230	98	127
1998	124	178	79	134	317	330	299	442	340	273
1999	126	121	61	78	260	117	175	329	116	135
2000	88	176	102	133	253	155	120	160	54	42
2001	89	129	53	60	167	49	49	140	10	6
2002	79	164	73	69	222	150	142	196	212	176
2003	132	161	57	97	134	40	45	271	53	75
2004	112	112	67	95	132	55	56	180	75	63
2005	106	179	194	166	184	157	151	265	295	239
2006	115	149	119	121	184	114	95	150	103	76
2007	106	233	185	231	376	305	321	521	440	537
2008	63	143	73	104	203	79	114	361	37	32
2009	128	147	48	66	188	9	5	93	4	3
2010	117	129	109	100	107	169	148	152	220	140
2011	100	177	107	93	192	88	120	283	127	154
2012	116	158	117	96	122	60	83	144	71	68
2013	135	232	200	226	412	180	275	389	225	363
2014	112	158	133	134	152	114	131	162	114	178

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 Face, Sam and Fan are grouped according to $n = 100$ *(columns 3-5),* $n = 300$ *(columns 6-8) and* $n = 500$ *(columns 9-11).*

APPENDIX

Appendix A: Regularity conditions

We state the following assumptions.

Assumption A1. (i) $\{X_t\}_{t\geq 1}$ is stationary and ergodic; (ii) $\{\boldsymbol{\epsilon}_t\}_{t\geq 1}$ and $\{X_t\}_{t\geq 1}$ are independent; (iii) X_t's are bounded with support $\mathcal{X}$, that is, $\sup_{t\geq 1}\|X_t\|_\infty \leq L, a.s.$

Let $P(A)$ be the probability of a measurable set A and $E(X)$ be the expectation of a random variable X. The following strong mixing condition (A2) aims at conducting asymptotic properties of the index estimator and local linear estimators of nonparametric functions. Let $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_k^\infty$ be the $\sigma-$algebras generated by $\{X_t, t \leq 0\}$ and $\{X_t, t \geq T\}$, respectively and define the $\alpha-$mixing coefficient

$$\alpha(k) = \sup_{A\in\mathcal{F}_{-\infty}^0, B\in\mathcal{F}_k^\infty} |P(A)P(B) - P(AB)|.$$

Assumption A2. There exist positive constants c and $0 < \rho < 1$ such that for all $k = 1, 2, \cdots,$

$$\alpha(k) \leq c\rho^{-k}.$$

Assumption A3. (i) The kernel function $K(z)$ is a symmetric density function which is bounded with a bounded support and satisfies the Lipschitz condition; (ii) The density function $f_{\mathbf{b}}(z)$ of $X^{\mathrm{T}}\mathbf{b}$ is twice differentiable and bounded away from zero on $\{z = \mathbf{x}^{\mathrm{T}}\mathbf{b}; \mathbf{x} \in \mathcal{X}, \|\mathbf{b}-\boldsymbol{\beta}\|_2 \leq c_0\}$ with $0 < c_0 < 1$; (iii) The density function $f(\mathbf{x})$ of X_t is bounded away from zero and twice differentiable in $\mathcal{X}$ and the joint densities of X_1 and X_k for all $k \geq 2$ are bounded.

Assumption A4. $\mathbf{g}(z)$ and $\boldsymbol{\Phi}(z)$ have continuous third derivatives in $\mathcal{Z} = \{z : z = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{x} \in \mathcal{X}\}$.

Assumption A5. $\|\mathbf{V}_p - \mathbf{V}\| = o(1)$, as $p_n \to \infty$, for some $q \times q$ symmetric positive definite $\mathbf{V}$ such that $\lambda_{\min}(\mathbf{V})$ is bounded away from zero.

For the error process $\{\boldsymbol{\epsilon}_t, t \geq 1\}$, the following assumptions are stated. Denote the true value $\boldsymbol{\theta}_\ell = (\alpha_{\ell,0}, \cdots, \alpha_{\ell,m}, \gamma_{\ell,1}, \cdots, \gamma_{\ell,s})^{\mathrm{T}}$ for $\ell = 1, \cdots, p_n$.

Assumption B1. For each $\ell = 1 \cdots, p_n$, $\{(\epsilon_{\ell,t}, \sigma^2_{\ell,t}), t = 0, \pm 1, \pm 2, \cdots\}$ is a strictly stationary GARCH(m, s) process with $\sup_{1 \leq \ell \leq p_n} E\sigma^{2d}_{\ell,1} < \infty$ with $d > 4$.

Assumption B2. Let $\eta_{\ell,t} = \sigma^{-1}_{\ell,t}\epsilon_{\ell,t}$ for each t and ℓ. Then, for each $\ell = 1, \cdots, p_n$, the innovations $\eta_{\ell,t}$'s are *i.i.d.* and absolutely continuous with Lebesgue density being strictly positive in a neighbourhood of zero. Furthermore, $E\eta_{\ell,1} = 0$, $E\eta^2_{\ell,1} = 1$ and $\sup_{\ell \leq p_n} E(\eta^{2d}_{\ell,1}) < \infty$ with d defined in Assumption (B1).

Assumption B3. For each $\ell = 1, \cdots, p_n$, the true value $\boldsymbol{\theta}_{\ell,0}$ is an interior point of the compact set $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda} \subset (c, +\infty) \times (c, +\infty)^{m+s}$ for a constant $c > 0$.

Assumption B4. Let $\mathcal{A}_{\ell,\theta}(z) = \sum_{i=1}^{m} \alpha_{\ell,i} z^i$ and $\mathcal{B}_{\ell,\theta}(z) = 1 - \sum_{i=1}^{s} \gamma_{\ell,i} z^i$ for $\ell = 1, \cdots, p_n$. If $s > 0$, $\mathcal{A}_{\ell,\theta_{\ell,0}}(z)$ and $\mathcal{B}_{\ell,\theta_{\ell,0}}(z)$ have no common roots, $\mathcal{A}_{\ell,\theta_{\ell,0}}(1) \neq 0$, and $\alpha_{\ell,0m} + \gamma_{\ell,0s} \neq 0$.

For the bandwidths h, h_1, h_2 and the dimension p_n, we require the following assumptions.

Assumption C1. (i) The bandwidth h and h_1 satisfy $h = O(n^{-\tau})$ and $h_1 = O(n^{-\tau_1})$, respectively, with $1/6 < \tau, \tau_1 < 1/4$.

Assumption C2. The bandwidth h_2 satisfies $h_2 = O(n^{-\tau_2})$ with $1/(2q+4) < \tau_2 < 1/(2q+2)$.

Assumption C3. The dimension p_n satisfies $p_n \leq Cn^{d/2-2-2\varepsilon}$ for some constants $C > 0$ and $0 < 2\varepsilon < d/2 - 2$.

Our aim is to estimate $\mathrm{cov}(Y_t|\mathcal{F}_{t-1})$. Fan, Fan and Lv (2008) and Fan, Liao and Mincheva (2013) showed that by incorporating the factor structure into the covariance matrix, the resulting estimator has a better convergence rate than the usual sample covariance matrix under the norm

$\|\cdot\|_{\mathbf{\Sigma}}$. To prove the convergence rate of $\widehat{\text{cov}}(Y_t|\mathcal{F}_{t-1}) - \text{cov}(Y_t|\mathcal{F}_{t-1})$ under the norm $\|\cdot\|_{\mathbf{\Sigma}}$, we impose the following assumption:

Assumption C4. For each $\mathbf{x} \in \mathcal{X}$, $\|p_n^{-1}\{\mathbf{\Phi}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}}\mathbf{\Phi}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) - \mathbf{V}_2\| = o(1)$, as $p_n \to \infty$ for some $q \times q$ symmetric positive definite $\mathbf{V}_2$ such that $\lambda_{\min}(\mathbf{V}_2)$ is bounded away from zero.

The assumptions are regular. The strong mixing condition in the Assumption (A2) can be relaxed as $\alpha(k) \leq ck^{-\beta}$ with a large constant β. Assumption (B1) and (B2) guarantee the existence of the $2d-$th moment of $\epsilon_{\ell,1}$. For simplicity, we do not impose the conditions that ensure the finiteness of the $d-$th moment of $\sigma^2_{\ell,1}$. For more details, see Lindner (2009). Assumption (C4) requires that the factors should be pervasive, that is, impact every individual time series. It was also imposed in Fan, Fan and Lv (2008) and Fan, Liao and Mincheva (2011).

Appendix B: Proof of Theorem 1 (I)-(III)

For ease of presentation, we give some notation. Define

$$\delta_{\mathbf{b}} = \|\mathbf{b} - \boldsymbol{\beta}\|, \delta_{1n} = \left(\frac{\log(n)}{nh}\right)^{1/2}, \delta_{2n} = \left(\frac{\log(n)}{n}\right)^{1/2}, \delta_{3n} = \left(\frac{\log(n)}{nh_1}\right)^{1/2}$$

and $\tilde{\delta}_n = h^3 + h^2\delta_{1n} + \delta^2_{1n}$. Define $\mathbf{\Theta}$ to be a compact set $\{\mathbf{b} : \|\mathbf{b} - \boldsymbol{\beta}\| \leq c_0, \|\mathbf{b}\| = 1\}$ with a small $c_0 > 0$. For a random sequence a_n, $a_n = \bar{O}_{a.s.}(b_n)$ for some sequence b_n means that $P\{\|a_n\| > Cb_n\} = O(n^{-(1+\varepsilon)})$, where ε is defined in Assumption (C3).

To prove Theorem 1, the following lemma is useful.

Lemma B.1. Assume that Conditions (A1)-(A3) and (C3) in Appendix A hold and for some $d > 4$,

$$\sup_{1\leq\ell\leq p_n} E|\epsilon_{\ell,t}|^{2d} < \infty,$$

where d is defined in (C3). Then there exists a constant $C > 0$ such that

$$P\left\{\sup_{1\leq\ell\leq p_n}\sup_{(\mathbf{b},\mathbf{x})\in(\mathbf{\Theta},\mathcal{X})}\left|\frac{1}{n}\sum_{t=1}^{n}K_h(X_i^{\mathrm{T}}\mathbf{b} - \mathbf{x}^{\mathrm{T}}\mathbf{b})\epsilon_{\ell,t}\right| > C\delta_{1n}\right\} \leq O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

The proof of Lemma B.1 can be followed from the proof of Lemma 6.1 in Fan and Yao (2003). Of course, some constants involved in the proof need to be modified. For instance, we instead use $B_n = (nh)^{1/2}(\log(n))^{-2}$.

Denote $Y = (Y_2, \cdots, Y_n)$, $W_h(z;\mathbf{b}) = \text{diag}\left\{K_h(X_1^{\mathrm{T}}\mathbf{b} - z), \cdots, K_h(X_{n-1}^{\mathrm{T}}\mathbf{b} - z)\right\}$ and

$$\widetilde{X}(z;\mathbf{b}) = \begin{pmatrix} \widetilde{X}_2^{\mathrm{T}}(z;\mathbf{b}) \\ \vdots \\ \widetilde{X}_n^{\mathrm{T}}(z;\mathbf{b}) \end{pmatrix} = \begin{pmatrix} 1 & X_2^{\mathrm{T}} & X_1^{\mathrm{T}}\mathbf{b} - z & (X_1^{\mathrm{T}}\mathbf{b} - z)X_2^{\mathrm{T}} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n^{\mathrm{T}} & X_{n-1}^{\mathrm{T}}\mathbf{b} - z & (X_{n-1}^{\mathrm{T}}\mathbf{b} - z)X_n^{\mathrm{T}} \end{pmatrix},$$

Let $H = \mathrm{diag}(\mathbf{1}_{1\times(q+1)}, h\mathbf{1}_{1\times(q+1)})$ and denote $\widehat{\Omega}_h(z;\mathbf{b}) = H^{-1}\{\widetilde{X}(z;\mathbf{b})\}^{\mathrm{T}}W_h(z;\mathbf{b})\widetilde{X}(z;\mathbf{b})H^{-1}$. Denote $\mu_2 = \int u^2K(u)du$, $\boldsymbol{\mu}_{\mathbf{b}}(z) = E(X|X^{\mathrm{T}}\mathbf{b} = z)$ and, for $\ell = 1,\cdots,p_n$, $\tilde{\boldsymbol{\epsilon}}_\ell = (\epsilon_{\ell,2},\cdots,\epsilon_{\ell,n})^{\mathrm{T}}$,

$$\begin{aligned}\widehat{\boldsymbol{\Gamma}}_\ell(z;\mathbf{b}) &= H^{-1}\left\{\widehat{\Omega}_h(z;\mathbf{b})\right\}^{-1}H^{-1}\widetilde{X}^{\mathrm{T}}(z;\mathbf{b})W_h(z;\mathbf{b})\mathbf{y}_\ell, \boldsymbol{\Gamma}_\ell(z) = (g_\ell(z),(\boldsymbol{\Phi}_\ell(z))^{\mathrm{T}},\dot{g}_\ell(z),(\dot{\boldsymbol{\Phi}}_\ell(z))^{\mathrm{T}})^{\mathrm{T}},\\ \boldsymbol{\Gamma}'_\ell(z) &= (\dot{g}_\ell(z),(\dot{\boldsymbol{\Phi}}_\ell(z))^{\mathrm{T}},\ddot{g}_\ell(z),(\ddot{\boldsymbol{\Phi}}_\ell(z))^{\mathrm{T}})^{\mathrm{T}}, \boldsymbol{\Gamma}''_\ell(z) = (\ddot{g}_\ell(z),(\ddot{\boldsymbol{\Phi}}_\ell(z))^{\mathrm{T}},\mathbf{0}_{1\times(q+1)})^{\mathrm{T}}.\end{aligned}$$

The following lemma gives the asymptotic representation of $\widehat{\boldsymbol{\Gamma}}_\ell(z)$.

Lemma B.2. Suppose that Assumption (A1)-(A4) in Appendix A hold. Then we have that

$$\begin{aligned}H\widehat{\boldsymbol{\Gamma}}_\ell(z;\mathbf{b}) &= H\boldsymbol{\Gamma}_\ell(z) + \left\{\widehat{\Omega}_h(z;\mathbf{b})\right\}^{-1}H^{-1}\widetilde{X}^{\mathrm{T}}(z;\mathbf{b})W_h(z;\mathbf{b})\tilde{\boldsymbol{\epsilon}}_\ell + H\boldsymbol{\Gamma}'_\ell(z)\big(\boldsymbol{\mu}_{\mathbf{b}}(z)\big)^{\mathrm{T}}(\boldsymbol{\beta}-\mathbf{b})\\ &\quad+\frac{1}{2}\mu_2h^2H\boldsymbol{\Gamma}''_\ell(z) + \bar{O}_{a.s.}\big(h\delta_{\mathbf{b}} + \delta_{1n}\delta_{\mathbf{b}} + \delta_{\mathbf{b}}^2 + \tilde{\delta}_n\big).\end{aligned}$$

Proof of Lemma B.2. For $i = 2,\cdots,n$, denote $z_i = X_{i-1}^{\mathrm{T}}\boldsymbol{\beta}$ and $z_{\mathbf{b},i} = X_{i-1}^{\mathrm{T}}\mathbf{b}$. Using a Taylor's expansion, we obtain that

$$y_{\ell,i} = g_\ell(z_i) + \boldsymbol{\Phi}_\ell(z_i)X_i + \epsilon_{\ell,i} = \widetilde{X}_i^{\mathrm{T}}(z;\mathbf{b})\boldsymbol{\Gamma}_\ell(z) + \epsilon_{\ell,i} + r_{\ell,\mathbf{b},i}^{(1)} + r_{\ell,\mathbf{b},i}^{(2)} + r_{\ell,\mathbf{b},i}^{(3)} + r_{\ell,\mathbf{b},i}^{(4)},$$

where $r_{\ell,\mathbf{b},i}^{(1)} = \widetilde{X}_{\mathbf{b},i}^{\mathrm{T}}(z)\boldsymbol{\Gamma}'_\ell(z)(z_i - z_{\mathbf{b},i}), r_{\ell,\mathbf{b},i}^{(2)} = 2^{-1}\widetilde{X}_{\mathbf{b},i}^{\mathrm{T}}(z)\boldsymbol{\Gamma}''_\ell(z)(z_{\mathbf{b},i} - z)^2$,
$r_{\ell,\mathbf{b},i}^{(3)} = 2^{-1}\widetilde{X}_{\mathbf{b},i}^{\mathrm{T}}(z)\boldsymbol{\Gamma}''_\ell(z)(z_i - z_{\mathbf{b},i})^2, r_{\ell,\mathbf{b},i}^{(4)} = O(|z_i - z|^3)$.

For $k = 1,\cdots,4$, denote $\mathbf{r}_{\ell,\mathbf{b}}^{(k)} = (\mathbf{r}_{\ell,\mathbf{b},2}^{(k)},\cdots,\mathbf{r}_{\ell,\mathbf{b},n}^{(k)})^{\mathrm{T}}$. Then

$$H\widehat{\boldsymbol{\Gamma}}_\ell(z;\mathbf{b}) - H\boldsymbol{\Gamma}_\ell(z) = \left\{\widehat{\Omega}_h(z;\mathbf{b})\right\}^{-1}H^{-1}\widetilde{X}^{\mathrm{T}}(z;\mathbf{b})W_h(z;\mathbf{b})\left(\tilde{\boldsymbol{\epsilon}}_\ell + \mathbf{r}_{\ell,\mathbf{b}}^{(1)} + \mathbf{r}_{\ell,\mathbf{b}}^{(2)} + \mathbf{r}_{\ell,\mathbf{b}}^{(3)} + \mathbf{r}_{\ell,\mathbf{b}}^{(4)}\right).$$

(I). Consider the term $\widehat{\Omega}_h(z;\mathbf{b})$. Following the proof of Theorem 5.3 in Fan and Yao (2003), we have that there exists a large $C > 0$ such that

$$P\left\{\sup_{(\mathbf{b},z)\in\boldsymbol{\Theta}\times\mathcal{Z}}\left\|\frac{1}{n}\left(\widehat{\Omega}_h(z;\mathbf{b}) - E\left\{\widehat{\Omega}_h(z;\mathbf{b})\right\}\right)\right\|_F > C\delta_{1n}\right\} \le O\left(\frac{1}{n^2}\right).$$

Let $\Omega(z;\mathbf{b}) = \lim_{n\to\infty}n^{-1}E\left\{\widehat{\Omega}_h(z;\mathbf{b})\right\}$. Note that $n^{-1}E\widehat{\Omega}_h(z;\mathbf{b}) = \Omega(z;\mathbf{b}) + O(h)$ and $\Omega(z;\mathbf{b})$ is positive definite. Therefore, $\widehat{\Omega}_h(z;\mathbf{b})$ is positive definite almost surely and

$$n^{-1}\widehat{\Omega}_h(z;\mathbf{b}) = \Omega(z;\mathbf{b}) + \bar{O}_{a.s.}(h + \delta_{1n}).$$

(II). Consider the term $H^{-1}\widetilde{X}^{\mathrm{T}}(z;\mathbf{b})W_h(z;\mathbf{b})\mathbf{r}_{\mathbf{b},\ell}^{(k)}$ $(k = 1,\cdots,4)$. By specific matrix calculations, we can show that

$$\begin{aligned}H^{-1}\widetilde{X}^{\mathrm{T}}(z;\mathbf{b})W_h(z;\mathbf{b})\mathbf{r}_{\mathbf{b},\ell}^{(1)} &= \Omega(z;\mathbf{b})H\boldsymbol{\Gamma}'_\ell(z)\big(\boldsymbol{\mu}_{\mathbf{b}}(z)\big)^{\mathrm{T}}(\boldsymbol{\beta}-\mathbf{b}) + \bar{O}_{a.s.}\Big(h\delta_{\mathbf{b}} + \delta_{1n}\delta_{\mathbf{b}}\Big),\\ H^{-1}\widetilde{X}^{\mathrm{T}}(z;\mathbf{b})W_h(z;\mathbf{b})\mathbf{r}_{\mathbf{b},\ell}^{(2)} &= \frac{1}{2}\mu^2h^2\Omega_(z;\mathbf{b})H\boldsymbol{\Gamma}''_\ell(z) + \bar{O}_{a.s.}\Big(h^3 + h^2\delta_{1n}\Big),\\ H^{-1}\widetilde{X}^{\mathrm{T}}(z;\mathbf{b})W_h(z;\mathbf{b})\mathbf{r}_{\mathbf{b},\ell}^{(3)} &= \bar{O}_{a.s.}\big(\delta_{\mathbf{b}}^2\big),\quad H^{-1}\widetilde{X}^{\mathrm{T}}(z;\mathbf{b})W_h(z;\mathbf{b})\mathbf{r}_{\mathbf{b},\ell}^{(4)} = \bar{O}_{a.s.}\big(\delta_{\mathbf{b}}^3 + h^3 + h^2\delta_{\mathbf{b}} + h\delta_{\mathbf{b}}^2\big).\end{aligned}$$

Combining (I) and (II), we obtain that

$$\begin{aligned} H\widehat{\mathbf{\Gamma}}_\ell(z;\mathbf{b}) &= H\mathbf{\Gamma}_\ell(z) + \big\{\widehat{\Omega}_h(z;\mathbf{b})\big\}^{-1} H^{-1}\widetilde{X}^{\mathrm{T}}(z;\mathbf{b})W_h(z;\mathbf{b})\tilde{\boldsymbol{\epsilon}}_\ell + H\mathbf{\Gamma}'_\ell(z)\big(\boldsymbol{\mu}_{\mathbf{b}}(z)\big)^{\mathrm{T}}(\boldsymbol{\beta}-\mathbf{b}) \\ &\quad + \frac{1}{2}\mu^2 h^2 H\mathbf{\Gamma}''_\ell(z) + \bar{O}_{a.s.}\big(h\delta_{\mathbf{b}} + \delta_{1n}\delta_{\mathbf{b}} + \delta_{\mathbf{b}}^2 + \tilde{\delta}_n\big). \end{aligned}$$

This completes the proof.

The following lemma, Lemma B.3, gives the asymptotic relationship between $\widehat{\boldsymbol{\beta}}_{m+1}$ and $\widehat{\boldsymbol{\beta}}_m$, where $\widehat{\boldsymbol{\beta}}_m$ is the mth step estimator based on our procedure in Section 2.

Without loss of generality, we consider $m = 1$. For each $i, j = 1, \cdots, n-1$, define

$$X_{ij} = X_i - X_j, w_{ij}(\mathbf{b}) = h^{-1}K\left\{X_{ij}^{\mathrm{T}}\mathbf{b}/h\right\}.$$

Given $\widehat{\boldsymbol{\beta}}_1$, for $j = 1, \cdots, n-1$, denote $\hat{z}_j = X_j^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_1$ and

$$\widehat{\mathbf{\Gamma}}_j = (\widehat{\mathbf{g}}_j, \widehat{\boldsymbol{\xi}}_j, \widehat{A}_j, \widehat{B}_j) = YW_h(\hat{z}_j;\widehat{\boldsymbol{\beta}}_1)\widetilde{X}(\hat{z}_j;\widehat{\boldsymbol{\beta}}_1)\left\{\widetilde{X}^{\mathrm{T}}(\hat{z}_j;\widehat{\boldsymbol{\beta}}_1)W_h(\hat{z}_j;\widehat{\boldsymbol{\beta}}_1)\widetilde{X}(\hat{z}_j;\widehat{\boldsymbol{\beta}}_1)\right\}^{-1}.$$

and

$$\begin{aligned} \widehat{\mathbf{V}}_n &= \frac{1}{n^2 p_n}\sum_{i,j=1}^{n-1} X_{ij}X_{ij}^{\mathrm{T}}\|\widehat{\boldsymbol{\xi}}_j + \widehat{B}_j X_{i+1}\|^2 w_{ij}(\widehat{\boldsymbol{\beta}}_1), \\ \widehat{\mathbf{U}}_n &= \frac{1}{n^2 p_n}\sum_{i,j=1}^{n-1} X_{ij}\Big(\widehat{\boldsymbol{\xi}}_j + \widehat{B}_j X_{i+1}\Big)^{\mathrm{T}}\Big\{Y_{i+1} - \widehat{\mathbf{\Gamma}}_j\widetilde{X}(\hat{z}_j;\widehat{\boldsymbol{\beta}}_1)\Big\}w_{ij}(\widehat{\boldsymbol{\beta}}_1), \\ \widehat{\boldsymbol{\beta}}_2 &= \widehat{\boldsymbol{\beta}}_1 + \widehat{\mathbf{V}}_n^{-1}\widehat{\mathbf{U}}_n. \end{aligned}$$

Lemma B.3. Suppose that Conditions (A1)-(A4), (B1)-(B4), (C1) and (C3) in Appendix A hold. Then, we have

$$\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta} = \frac{1}{2}\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\right) + \frac{1}{2}\mathbf{V}_p^{-1}\mathbf{U}_n + \mathbf{R}_n, \tag{A.1}$$

where $\mathbf{R}_n = \bar{O}_{a.s.}\left(h\delta_{2n} + h^{-1}\delta_{2n}^2 + \tilde{\delta}_n + h^{-1}\delta_{2n}\delta_{\widehat{\boldsymbol{\beta}}_1} + h\delta_{\widehat{\boldsymbol{\beta}}_1} + h^{-1}\delta_{\widehat{\boldsymbol{\beta}}_1}^2\right)$.

Proof of Lemma B.3. First, consider the term $\mathbf{U}_n$. For $i, j = 1, \cdots, n-1$, denote

$$\begin{aligned} \mathbf{e}_{ij,1} &= \mathbf{g}'(X_j^{\mathrm{T}}\boldsymbol{\beta}) + \mathbf{\Phi}'(X_j^{\mathrm{T}}\boldsymbol{\beta})X_{i+1}, \mathbf{e}_{ij,2} = \widehat{\boldsymbol{\xi}}_j + \widehat{\mathbf{B}}_j X_{i+1} - \mathbf{e}_{ij,1}, \\ \mathbf{e}_{i,3} &= \mathbf{g}(\hat{z}_i) + \mathbf{\Phi}(\hat{z}_i)X_{i+1} + (\mathbf{g}'(\hat{z}_i) + \mathbf{\Phi}'(\hat{z}_i)X_{i+1})(\boldsymbol{\mu}_{\boldsymbol{\beta}}(X_i^{\mathrm{T}}\boldsymbol{\beta}))^{\mathrm{T}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_1), \\ \mathbf{e}_{ij,4} &= \widehat{\mathbf{\Gamma}}_j\widetilde{X}_{i+1}(\hat{z}_j;\widehat{\boldsymbol{\beta}}_1) - \mathbf{e}_{i,3}. \end{aligned}$$

We decompose $\widehat{\mathbf{U}}_n$ as

$$\begin{aligned} \widehat{\mathbf{U}}_n &= \frac{1}{n^2 p_n}\sum_{i,j=1}^{n-1} X_{ij}\mathbf{e}_{ij,1}^{\mathrm{T}}\big(Y_{i+1} - \mathbf{e}_{i,3}\big)w_{ij}(\widehat{\boldsymbol{\beta}}_1) - \frac{1}{n^2 p_n}\sum_{i,j=1}^{n-1} X_{ij}\mathbf{e}_{ij,1}^{\mathrm{T}}\mathbf{e}_{j,4}w_{ij}(\widehat{\boldsymbol{\beta}}_1) \\ &\quad + \frac{1}{n^2 p_n}\sum_{i,j=1}^{n-1} X_{ij}\mathbf{e}_{ij,2}^{\mathrm{T}}\big(Y_{i+1} - \mathbf{e}_{i,3}\big)w_{ij}(\widehat{\boldsymbol{\beta}}_1) - \frac{1}{n^2 p_n}\sum_{i,j=1}^{n-1} X_{ij}\mathbf{e}_{ij,2}^{\mathrm{T}}\mathbf{e}_{ij,4}w_{ij}(\widehat{\boldsymbol{\beta}}_1) = \sum_{k=1}^{4}\mathbf{U}_{nk}. \end{aligned}$$

(a). Consider the main term $\mathbf{U}_{n1}$. Note that

$$Y_{i+1} - \mathbf{e}_{i,3} = \boldsymbol{\epsilon}_{i+1} + \left(\mathbf{g}'(X_i^{\mathrm{T}}\boldsymbol{\beta}) + \boldsymbol{\Phi}'(X_i^{\mathrm{T}}\boldsymbol{\beta})X_{i+1}\right)\left(X_i - \boldsymbol{\mu}_{\boldsymbol{\beta}}(X_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_1) + O(\delta^2_{\widehat{\boldsymbol{\beta}}_1}).$$

Analogous to Lemma A.2 of Xia, Tong and Li (2002), it follows that

$$\frac{1}{n^2 p_n}\sum_{i,j=1}^{n-1} X_{ij}\mathbf{e}_{ij,1}^{\mathrm{T}}\boldsymbol{\epsilon}_{i+1}w_{ij}(\widehat{\boldsymbol{\beta}}_1) = \mathbf{U}_n + \bar{O}_{a.s.}(\delta_{1n}\delta_{\widehat{\boldsymbol{\beta}}_1}).$$

Similarly, we obtain that

$$\frac{1}{n^2 p_n}\sum_{i,j=1}^{n-1} X_{ij}\mathbf{e}_{ij,1}^{\mathrm{T}}\left(\mathbf{g}'(X_i^{\mathrm{T}}\boldsymbol{\beta}) + \boldsymbol{\Phi}'(X_i^{\mathrm{T}}\boldsymbol{\beta})X_{i+1}\right)\left(X_i - \boldsymbol{\mu}_{\boldsymbol{\beta}}(X_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}w_{ij}(\widehat{\boldsymbol{\beta}}_1) = \mathbf{V}_p + \bar{O}_{a.s.}(\delta_{1n} + \delta_{\widehat{\boldsymbol{\beta}}_1}).$$

Hence, we approximate the term $\mathbf{U}_{n1}$ as

$$\mathbf{U}_{n1} = \mathbf{U}_n + \mathbf{V}_p(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_1) + \bar{O}_{a.s.}\big(\delta_{1n}\delta_{\widehat{\boldsymbol{\beta}}_1} + \delta^2_{\widehat{\boldsymbol{\beta}}_1}\big).$$

(b). With the help of asymptotic representation of $\widehat{\boldsymbol{\Gamma}}_j(z)$ and empirical approximation theories, we can show that

$$\mathbf{U}_{nk} = \bar{O}_{a.s.}\big(h\delta_{\widehat{\boldsymbol{\beta}}_1} + h^{-1}\delta_{2n}\delta_{\widehat{\boldsymbol{\beta}}_1} + h^{-1}\delta^2_{\widehat{\boldsymbol{\beta}}_1} + \tilde{\delta}_n\big), k = 2, 3, 4.$$

(c). In the similar fashion, we can also show that

$$\widehat{\mathbf{V}}_n = 2\mathbf{V}_p + \bar{O}_{a.s.}\left(\delta_{\widehat{\boldsymbol{\beta}}_1} + h + h^{-1}\delta_{2n}\right).$$

Therefore, $\widehat{\boldsymbol{\beta}}_2 - \widehat{\boldsymbol{\beta}}_1 = 2^{-1}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_1) + 2^{-1}\mathbf{V}_p^{-1}\mathbf{U}_n + \mathbf{R}_n$, which means that

$$\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta} = \frac{1}{2}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}) + \frac{1}{2}\mathbf{V}_p^{-1}\mathbf{U}_n + \mathbf{R}_n. \tag{A.2}$$

This completes the proof.

Proof of Theorem 1 (I). First, by Lemma B.3, for the $m-$th step ($m > 1$), we have

$$\widehat{\boldsymbol{\beta}}_{m+1} - \boldsymbol{\beta} = \frac{1}{2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) + \frac{1}{2}\mathbf{V}_p^{-1}\mathbf{U}_n + \mathbf{R}_{n,m}, \tag{A.3}$$

where $\|\mathbf{R}_{n,m}\| \leq M\left(\delta_{\widehat{\boldsymbol{\beta}}_m}(h + h^{-1}\delta_{2n} + h^{-1}\delta_{\widehat{\boldsymbol{\beta}}_m}) + \tilde{\delta}_n + h\delta_{2n} + h^{-1}\delta^2_{2n}\right)$ a.s. and $\|\mathbf{V}_p^{-1}\mathbf{U}_n\| \leq M\delta_{2n}$ a.s., with some large positive constant M. Here we take $M > 1$ and $h < 1$ for sufficiently large n. Note that as $n \to \infty$, the bandwidth h satisfies $h \to 0$, $h^{-1}\delta_{2n} \to 0$, $\tilde{\delta}_n h^{-1} \to 0$ and $h^{-2}\delta^2_{2n} \to 0$. We can assume that

$$h + h^{-1}\delta_{2n} \leq (8M)^{-1}, M(\tilde{\delta}_n + h\delta_{2n} + h^{-1}\delta^2_{2n}) + M\delta_{2n} \leq (32M)^{-1}h.$$

Then, if $\delta_{\boldsymbol{\beta}_m} \leq (8M)^{-1}h$, then $\delta_{\boldsymbol{\beta}_{m+1}} \leq (8M)^{-1}h$ and

$$\delta_{\boldsymbol{\beta}_{m+1}} \leq \frac{3}{4}\delta_{\boldsymbol{\beta}_m} + M(\tilde{\delta}_n + h\delta_{2n} + h^{-1}\delta_{2n}^2) + M\delta_{2n}.$$

Note that we can choose the initial estimator $\widehat{\boldsymbol{\beta}}_1$ which satisfies $\|\delta_{\boldsymbol{\beta}_1}\| \leq (8M)^{-1}h$ for sufficiently large n. Therefore,

$$\delta_{\boldsymbol{\beta}_{m+1}} \leq \left(\frac{3}{4}\right)^m \delta_{\boldsymbol{\beta}_1} + \left\{1 + \frac{3}{4} + \cdots + \left(\frac{3}{4}\right)^m\right\}\left\{M(\tilde{\delta}_n + h\delta_{2n} + h^{-1}\delta_{2n}^2 + \delta_{2n})\right\}.$$

Taking $m \to \infty$, it follows that the final estimator $\widehat{\boldsymbol{\beta}}$ satisfies $\delta_{\boldsymbol{\beta}} = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = \bar{O}_{a.s.}\left(\tilde{\delta}_n + \delta_{2n} + h^{-1}\delta_{2n}^2\right)$ and hence $\|\mathbf{R}_{n,\infty}\| = \bar{O}_{a.s.}\left(h^3 + \delta_{1n}^2\right)$. It also follows from the expression (A.3) that

$$P\left\{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} - \mathbf{V}_p^{-1}\mathbf{U}_n\| \geq C\left(h^3 + \delta_{1n}^2\right)\right\} \leq O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

This completes the proof of Theorem 1(I).

Proof of Theorem 1 (II) and (III). Lemma B.2 tells us that, for $\ell = 1, \cdots, p_n$,

$$\widehat{g}_\ell(z) - g_\ell(z) = e_1^{\mathrm{T}}\{\widehat{\Omega}_{h_1}(z;\widehat{\boldsymbol{\beta}})\}^{-1}H_1^{-1}\widetilde{X}^{\mathrm{T}}(z;\widehat{\boldsymbol{\beta}})W_{h_1}(z;\widehat{\boldsymbol{\beta}})\tilde{\boldsymbol{\epsilon}}_\ell + \frac{1}{2}\mu_2 h_1^2 \ddot{g}_\ell(z) + \mathbf{R}_n(z),$$

where $e_1 = (1, 0, \cdots, 0)^{\mathrm{T}}$, $H_1 = \mathrm{diag}(\mathbf{1}_{1\times(q+1)}, h_1\mathbf{1}_{1\times(q+1)})$ and

$$P\Big\{\sup_{z\in\mathcal{Z}}|\mathbf{R}_n(z)| > C(h^3 + \delta_{1n}^2 + n^{-1/2})\Big\} = O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

for some constant $C > 0$.

(a). Consider the term $\widehat{\Omega}_{h_1}(z;\mathbf{b})$. Following the proof of Theorem 5.3 in Fan and Yao (2003), we have that there exists a large $C > 0$ such that

$$P\left\{\sup_{(\mathbf{b},z)\in\boldsymbol{\Theta}\times\mathcal{Z}}\left\|\frac{1}{n}\left(\widehat{\Omega}_{h_1}(z;\mathbf{b}) - E\left\{\widehat{\Omega}_{h_1}(z;\mathbf{b})\right\}\right)\right\|_F > C\delta_{3n}\right\} \leq O\left(\frac{1}{n^2}\right).$$

Let $\Omega(z;\mathbf{b}) = \lim_{n\to\infty} n^{-1}E\left\{\widehat{\Omega}_{h_1}(z;\mathbf{b})\right\}$. Note that $n^{-1}E\widehat{\Omega}_{h_1}(z;\mathbf{b}) = \Omega(z;\mathbf{b}) + O(h_1)$ and $\Omega(z;\mathbf{b})$ is positive definite. Therefore, $\widehat{\Omega}_{h_1}(z;\mathbf{b})$ is positive definite almost surely and

$$P\left\{\sup_{(\mathbf{b},z)\in\boldsymbol{\Theta}\times\mathcal{Z}}\left\|\frac{1}{n}\widehat{\Omega}_{h_1}(z;\mathbf{b}) - \Omega(z;\mathbf{b})\right\|_2^2 > C(h_1 + \delta_{3n})\right\} \leq O\left(\frac{1}{n^2}\right).$$

(b). By Lemma B.1, we have

$$P\left(\sup_{1\leq\ell\leq p_n}\sup_{(\mathbf{b},z)\in\boldsymbol{\Theta}\times\mathcal{Z}}\|\frac{1}{n}H_1\widetilde{X}^{\mathrm{T}}(z;\mathbf{b})W_{h_1}(z;\mathbf{b})\tilde{\boldsymbol{\epsilon}}_\ell\|_2 > C\delta_{3n}\right) \leq O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

Therefore, combining (a) and (b), there exists a large $C > 0$ such that

$$P\left\{\sup_{z\in\mathcal{Z}}\|\widehat{\mathbf{g}}(z) - \mathbf{g}(z)\|_\infty > C(h_1^2 + \delta_{3n})\right\} \leq O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

This completes the proof of Theorem 1(II). Theorem 1(III) can be proven analogously.

Appendix C: Proof of Theorem 1 (IV)

Before we prove Theorem 1(IV), we first give the convergence rate of the difference between the estimated residual $\widehat{\boldsymbol{\epsilon}}_t$ and the true residual $\boldsymbol{\epsilon}_t$.

Lemma C.1. Suppose that Assumptions (A1)-(A5), (B1)-(B4) and (C1) and (C3) in Appendix A hold. Then there exists $C > 0$ and small $\varepsilon > 0$ such that

$$P\left\{\sup_{t\leq n}\|\widehat{\boldsymbol{\epsilon}}_t - \boldsymbol{\epsilon}_t\|_\infty > C\left(h_1^2 + \delta_{3n}\right)\right\} \leq O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

Proof of Lemma C.1. For each $t = 2, \cdots, n$,

$$\widehat{\boldsymbol{\epsilon}}_t - \boldsymbol{\epsilon}_t = \widehat{\mathbf{g}}(X_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) - \mathbf{g}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + \left(\widehat{\boldsymbol{\Phi}}(X_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})\right)X_t.$$

Note that $\widehat{\mathbf{g}}(X_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) - \mathbf{g}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{g}'(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}^*)X_{t-1}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \widehat{\mathbf{g}}(X_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) - \mathbf{g}(X_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$, and

$$\begin{aligned}(\widehat{\boldsymbol{\Phi}}(X_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}))X_t &= \boldsymbol{\Phi}'(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}^*)X_tX_{t-1}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\quad +(\widehat{\boldsymbol{\Phi}}(X_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}))X_t.\end{aligned}$$

Hence, there exists a large constant $C > 0$ such that

$$\|\widehat{\boldsymbol{\epsilon}}_t - \boldsymbol{\epsilon}_t\|_\infty \leq \sup_{z\in\mathcal{Z}}\|\widehat{\mathbf{g}}(z) - \mathbf{g}(z)\|_\infty + \sup_{z\in\mathcal{Z}}\|\widehat{\boldsymbol{\Phi}}(z) - \boldsymbol{\Phi}(z)\|_\infty + C\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|,$$

where $\sup_{z\in\mathcal{Z}}\left(\|\mathbf{g}'(z)\|_\infty + \|\boldsymbol{\Phi}'(z)\|_\infty\right) = O(1)$ is used in the last terms. For any $v > 0$, we have the following inequality

$$\begin{aligned}P\left\{\sup_{2\leq t\leq n}|\widehat{\boldsymbol{\epsilon}}_t - \boldsymbol{\epsilon}_t| > 3v\right\} &\leq P\left\{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| > v/C\right\} + P\left\{\sup_{z\in\mathcal{Z}}\|\widehat{\mathbf{g}}(z) - \mathbf{g}(z)\|_\infty > v\right\} \\ &\quad +P\left\{\sup_{z\in\mathcal{Z}}\left\|\widehat{\boldsymbol{\Phi}}(z) - \boldsymbol{\Phi}(z)\right\|_\infty > v\right\}.\end{aligned}$$

Take $v = C(h_1^2 + \delta_{3n})$ for a large constant $C > 0$. It follows from parts (II) and (III) of Theorem 1 that there exists a constant $C > 0$ such that

$$P\Big\{\sup_{2\leq t\leq n}\|\widehat{\boldsymbol{\epsilon}}_t - \boldsymbol{\epsilon}_t\|_\infty > C\left(h_1^2 + \delta_{3n}\right)\Big\} \leq O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

This completes the proof of Lemma C.1.

Now we are going to prove Theorem 1(IV). Define the quasi log-likelihood function

$$\widetilde{Q}_{\ell,n}(\boldsymbol{\theta}) = n^{-1}\sum_{t=1}^{n}\widetilde{v}_{\ell,t}(\boldsymbol{\theta}), \widetilde{v}_{\ell,t}(\boldsymbol{\theta}) = \frac{\epsilon_{\ell,t}^2}{\widetilde{\sigma}_{\ell,t}^2(\boldsymbol{\theta})} + \log\widetilde{\sigma}_{\ell,t}^2(\boldsymbol{\theta}),$$

where $\widetilde{\sigma}_{\ell,t}^2(\boldsymbol{\theta})$ is the solution of

$$\widetilde{\sigma}_{\ell,t}^2(\boldsymbol{\theta}) = \alpha_{\ell,0} + \sum_{i=1}^{m}\alpha_{\ell,i}\epsilon_{\ell,t-i}^2 + \sum_{i=1}^{s}\gamma_{\ell,i}\widetilde{\sigma}_{\ell,t-i}^2(\boldsymbol{\theta}).$$

For convenience, denote the true value of $\boldsymbol{\theta}_\ell$ by $\boldsymbol{\theta}_{\ell,0}$. First, we consider the consistency of $\widehat{\boldsymbol{\theta}}_\ell$. Recall that the observed quasi log likelihood function

$$Q_{\ell,n}(\boldsymbol{\theta}) = n^{-1}\sum_{t=1}^{n} v_{\ell,t}(\boldsymbol{\theta}), v_{\ell,t}(\boldsymbol{\theta}) = \frac{r_{\ell,t}^2}{\sigma_{\ell,t}^2(\boldsymbol{\theta})} + \log\sigma_{\ell,t}^2(\boldsymbol{\theta}),$$

where $\sigma_{\ell,t}^2(\boldsymbol{\theta})$ is defined in Section 2. Following the proof of Theorem 7.1 in Francq and Zokoian (2009), we shall establish the following results:

(a1) $\sup_{1\le\ell\le p_n}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Lambda}}|Q_{\ell,n}(\boldsymbol{\theta}) - \widetilde{Q}_{\ell,n}(\boldsymbol{\theta})| \to 0, a.s.$, as $n\to\infty$;

(a2) If there exists some t such that $\widetilde{\sigma}_{\ell,t}^2(\boldsymbol{\theta}) = \widetilde{\sigma}_{\ell,t}^2(\boldsymbol{\theta}_{\ell,0})$ a.s. in $P_{\boldsymbol{\theta}_{\ell,0}}$, then $\boldsymbol{\theta} = \boldsymbol{\theta}_{\ell,0}$;

(a3) $E_{\boldsymbol{\theta}_{\ell,0}}|\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})| < \infty$, and if $\boldsymbol{\theta}\neq\boldsymbol{\theta}_{\ell,0}$, $E_{\boldsymbol{\theta}_{\ell,0}}|\widetilde{v}_{\ell,t}(\boldsymbol{\theta})| > E_{\boldsymbol{\theta}_{\ell,0}}|\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})|$;

(a4) For any $\boldsymbol{\theta}\neq\boldsymbol{\theta}_{\ell,0}$, there exists a neighbourhood $U(\boldsymbol{\theta})$ such that

$$\lim\inf_{n\to\infty}\inf_{\boldsymbol{\theta}^*\in U(\boldsymbol{\theta})} Q_{\ell,2}(\boldsymbol{\theta}) > E_{\boldsymbol{\theta}_{\ell,0}}\widetilde{v}_{\ell,2}(\boldsymbol{\theta}_{\ell,0}), a.s.$$

By the proof of Theorem 7.1 in Francq and Zokoian (2009), we only need to prove (a1). Denote

$$\underline{\widetilde{\boldsymbol{\sigma}}}_{\ell,t}^2(\boldsymbol{\theta}) = \begin{pmatrix} \widetilde{\sigma}_{\ell,t}^2(\boldsymbol{\theta}) \\ \widetilde{\sigma}_{\ell,t-1}^2(\boldsymbol{\theta}) \\ \vdots \\ \widetilde{\sigma}_{\ell,t-m+1}^2(\boldsymbol{\theta}) \end{pmatrix}, \underline{\widetilde{\mathbf{c}}}_{\ell,t}(\boldsymbol{\theta}) = \begin{pmatrix} \alpha_0 + \sum_{j=1}^m \alpha_j\epsilon_{\ell,t-j}^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \mathbf{B}_\ell = \begin{pmatrix} \gamma_1 & \gamma_2 & \cdots & \gamma_s \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix}.$$

We have the relationship $\underline{\widetilde{\boldsymbol{\sigma}}}_{\ell,t}^2 = \underline{\widetilde{\mathbf{c}}}_{\ell,t} + \mathbf{B}_\ell\underline{\widetilde{\boldsymbol{\sigma}}}_{\ell,t-1}^2$. The condition (B2) and the compactness of $\boldsymbol{\Lambda}$ implies that $\rho = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Lambda}}\rho(\mathbf{B}_\ell) < 1$, where $\rho(\mathbf{B})$ means the spectral radius of $\mathbf{B}$. Furthermore, $\underline{\widetilde{\boldsymbol{\sigma}}}_{\ell,t}^2$ can be expressed as

$$\underline{\widetilde{\boldsymbol{\sigma}}}_{\ell,t}^2 = \sum_{k=0}^{t-1}\mathbf{B}_\ell^k\underline{\widetilde{\mathbf{c}}}_{\ell,t-k} + \mathbf{B}_\ell^t\underline{\widetilde{\boldsymbol{\sigma}}}_{\ell,0}^2.$$

Let $\underline{\boldsymbol{\sigma}}_{\ell,t}^2(\boldsymbol{\theta})$ be the vector obtained by replacing $\widetilde{\sigma}_{\ell,t-i}^2(\boldsymbol{\theta})$ by $\sigma_{\ell,t-i}^2(\boldsymbol{\theta})$ in $\underline{\widetilde{\boldsymbol{\sigma}}}_{\ell,t}^2(\boldsymbol{\theta})$, and let $\underline{\mathbf{c}}_{\ell,t}$ be the vector obtained by replacing $\epsilon_{\ell,t-i}^2$ by $r_{\ell,t-i}^2$ and $r_{\ell,1}^2,\cdots,r_{\ell,2-m}^2$ by the initial values. Then we have

$$\underline{\boldsymbol{\sigma}}_{\ell,t}^2 = \sum_{k=0}^{t-1}\mathbf{B}_\ell^k\underline{\mathbf{c}}_{\ell,t-k} + \mathbf{B}_\ell^t\underline{\boldsymbol{\sigma}}_{\ell,0}^2.$$

Denote $\tilde{d}_\ell = \sup_{t\le n}|r_{\ell,t} - \epsilon_{\ell,t}|$. Then, if $t\ge m+1$,

$$\|\underline{\widetilde{\mathbf{c}}}_{\ell,t} - \underline{\mathbf{c}}_{\ell,t}\| \le \left|\sum_{j=1}^m \alpha_j(r_{\ell,t-j}^2 - \epsilon_{\ell,t-j}^2)\right| \le \tilde{d}_\ell^2 + 2\tilde{d}_\ell\sum_{j=1}^m\alpha_j\left|\epsilon_{\ell,t-j}\right|.$$

As a result, for $t \geq m+1$, we obtain that

$$\begin{aligned}
\|\widetilde{\underline{\boldsymbol{\sigma}}}^2_{\ell,t} - \underline{\boldsymbol{\sigma}}^2_{\ell,t}\| &\leq \left\| \sum_{k=0}^{t-m+1} \mathbf{B}^k_\ell(\widetilde{\underline{\mathbf{c}}}_{\ell,t-k} - \underline{\mathbf{c}}_{\ell,t-k}) \right\| + \left\| \sum_{k=t-m+2}^{t-1} \mathbf{B}^k_\ell(\widetilde{\underline{\mathbf{c}}}_{\ell,t-k} - \underline{\mathbf{c}}_{\ell,t-k}) \right\| \\
&\quad + \left\| \mathbf{B}^t_\ell(\widetilde{\underline{\boldsymbol{\sigma}}}^2_{\ell,0} - \underline{\boldsymbol{\sigma}}^2_{\ell,0}) \right\| \\
&\leq C \cdot \left(\tilde{d}^2_\ell + \tilde{d}_\ell \sum_{k=0}^{t-1} \rho^k \sum_{j=1}^{m} \alpha_j \left| \epsilon_{\ell,t-k-j} \right| + \rho^t \|\widetilde{\underline{\boldsymbol{\sigma}}}^2_{\ell,0} - \underline{\boldsymbol{\sigma}}^2_{\ell,0})\| \right),
\end{aligned}$$

for some constant $C > 0$. We thus have

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in\boldsymbol{\Lambda}} |Q_{\ell,n}(\boldsymbol{\theta}) - \widetilde{Q}_{\ell,n}(\boldsymbol{\theta})| &\leq n^{-1} \sum_{t=2}^{n} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Lambda}} \left\{ \left| \frac{\tilde{\sigma}^2_{\ell,t} - \sigma^2_{\ell,t}}{\tilde{\sigma}^2_{\ell,t}\sigma^2_{\ell,t}} \right| \epsilon^2_{\ell,t} + \left| \log \frac{\sigma^2_{\ell,t}}{\tilde{\sigma}^2_{\ell,t}} \right| \right\} \\
&\leq \frac{1}{\alpha^2_L} \cdot C \cdot \left(\tilde{d}^2_\ell + \tilde{d}_\ell + n^{-1} \sum_{t=2}^{n} \rho^t \epsilon^2_{\ell,t} \right) + \frac{1}{\alpha_L} \cdot C \cdot n^{-1} \sum_{t=2}^{n} \rho^t,
\end{aligned}$$

where $\alpha_L = \inf_{\boldsymbol{\theta}\in\boldsymbol{\Lambda}} |\alpha_{\ell,0}|$. Note that $\tilde{d}_\ell \leq C \cdot (h^2_1 + \delta_{3n}), a.s.$ and $\sup_{\ell\leq p_n} E\epsilon^{2d}_{\ell,t} < \infty$ implies that $\rho^t \epsilon^2_{\ell,t} \to 0, a.s.$ Then $\sup_{1\leq\ell\leq p_n} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Lambda}} |Q_{\ell,n}(\boldsymbol{\theta}) - \widetilde{Q}_{\ell,n}(\boldsymbol{\theta})| \to 0, a.s.$, and part (a) follows.

Next, we consider the convergence rate of $\sup_{1\leq\ell\leq p_n} \|\widehat{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}_{\ell,0}\|$. The proof of this part is based on a standard Taylor expansion of $\widetilde{Q}_{\ell,n}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_{\ell,0}$. Since $\widehat{\boldsymbol{\theta}}_\ell$ converges to $\boldsymbol{\theta}_{\ell,0}$, which lies in the interior of the parameter space, we thus have

$$\begin{aligned}
0 &= n^{-1} \sum_{t=2}^{n} \frac{\partial v_{\ell,t}(\widehat{\boldsymbol{\theta}}_\ell)}{\partial \boldsymbol{\theta}} \\
&= n^{-1} \sum_{t=2}^{n} \frac{\partial v_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial \boldsymbol{\theta}} + \left(\frac{1}{n} \sum_{t=2}^{n} \frac{\partial^2 v_{\ell,t}(\boldsymbol{\theta}^*_\ell)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \right) \cdot (\widehat{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}_{\ell,0}),
\end{aligned}$$

where $\boldsymbol{\theta}^*_\ell$ is between $\widehat{\boldsymbol{\theta}}_\ell$ and $\boldsymbol{\theta}_{\ell,0}$. Suppose we have shown that there exist two positive constants C_1 and C_2 such that

$$P\left\{ \sup_{1\leq\ell\leq p_n} \left\| \frac{1}{n} \sum_{t=2}^{n} \frac{\partial v_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial \boldsymbol{\theta}} \right\| > C_1(h^2_1 + \delta_{3n}) \right\} = O\left(\frac{1}{n^{1+\varepsilon}} \right), \tag{A.4}$$

and

$$P\left\{ \inf_{1\leq\ell\leq p_n} \inf_{\boldsymbol{\theta}\in V(\boldsymbol{\theta}_0)} \lambda_{\min} \left(\sum_{t=2}^{n} \frac{\partial^2 v_{\ell,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \right) \leq nC_2 \right\} = O\left(\frac{1}{n^{1+\varepsilon}} \right). \tag{A.5}$$

Denote

$$\mathcal{A}_n = \left\{ \inf_{1\leq\ell\leq p_n} \inf_{\boldsymbol{\theta}\in V(\boldsymbol{\theta}_{\ell,0})} \lambda_{\min} \left(n^{-1} \sum_{t=2}^{n} \frac{\partial^2 v_{\ell,t}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \right) > C_2 \right\},$$

where C_2 is defined in (A.5). Then, for each $x > 0$,

$$P\left\{ \sup_{1\leq\ell\leq p_n} \left\| \widehat{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}_{\ell,0} \right\| > x \right\} \leq P\left\{ \sup_{1\leq\ell\leq p_n} \left\| \sum_{t=2}^{n} \frac{\partial v_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial \boldsymbol{\theta}} \right\| > nC_2 x \right\} + P(\mathcal{A}^C_n). \tag{A.6}$$

Take $x = C_1(h_1^2 + \delta_{3n})/C_2$ and the proof of Theorem 1(IV) follows immediately from (A.4) and (A.5).

Now we prove (A.4) and (A.5). To establish (A.4) and (A.5), it suffices to prove the following five parts:

(b1) There exists a constant $C > 0$ such that

$$P\left\{\sup_{1\le\ell\le p_n}\left\|\sum_{t=2}^{n}\frac{\partial\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}}\right\| > Cn\delta_{3n}\right\} = o(1),$$

(b2) There exists a constant $C > 0$ such that

$$P\left\{\sup_{1\le\ell\le p_n}\left|\sum_{t=2}^{n}\frac{\partial\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}} - \sum_{t=2}^{n}\frac{\partial\tilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}}\right| > Cn(h_1^2 + \delta_{3n})\right\} = O\left(\frac{1}{n^{1+\varepsilon}}\right),$$

(b3) There exists a constant $C > 0$ such that

$$P\left\{\inf_{1\le\ell\le p_n}\lambda_{\min}\left(\sum_{t=2}^{n}\frac{\partial^2\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right) \le nC\right\} = O\left(\frac{1}{n^{1+\varepsilon}}\right),$$

(b4) For any $C > 0$, we have

$$P\left\{\sup_{1\le\ell\le p_n}\sup_{\boldsymbol{\theta}\in V(\boldsymbol{\theta}_0)}\left\|\sum_{t=2}^{n}\frac{\partial^2 v_{\ell,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} - \sum_{t=2}^{n}\frac{\partial^2\tilde{v}_{\ell,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right\| > nC\right\} = O\left(\frac{1}{n^{1+\varepsilon}}\right),$$

(b5) For each $i, j, k = 1, \cdots, m + s + 1$, there exists a constant $C > 0$ and very small constant $c > 0$ such that

$$P\left\{\sup_{1\le\ell\le p_n}\sup_{\boldsymbol{\theta}\in V(\boldsymbol{\theta}_0)}\left|n^{-1}\sum_{t=2}^{n}\frac{\partial^2\widetilde{v}_{\ell,t}(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j\partial\theta_k}\right| \le Cn^c\right\} = O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

It is not hard to see that (A.4) can be proved from (b1) and (b2) and (A.5) follows from (b3)-(b5). We now prove them separately.

(b1). It is easy to show that

$$\frac{\partial\widetilde{v}_{\ell,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \left(1 - \frac{\epsilon_{\ell,t}^2}{\widetilde{\sigma}_{\ell,t}^2(\boldsymbol{\theta})}\right)\left(\frac{1}{\widetilde{\sigma}_{\ell,t}^2(\boldsymbol{\theta})}\frac{\partial\widetilde{\sigma}_{\ell,t}^2(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)$$

and

$$E\left\|\frac{\partial\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}}\right\|^d < \infty.$$

Note that $\{\epsilon_{\ell,t}, t \le n\}$ are strictly stationary and $\alpha-$mixing with geometric rate. (Also see Lindner (2009).) It follows from Theorem 2 (ii) of Liu, Xiao and Wu (2013) that, there exist positive constants C_1, C_2 and C_3 such that for all $x > 0$,

$$P\left\{\left\|\sum_{t=2}^{n}\frac{\partial\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\theta}\right\| > x\right\} \le \frac{C_1 n}{x^d} + C_2\exp\left(-\frac{C_3 x^2}{n^{1/2}}\right).$$

Hence, by taking $x = C\delta_{2n}$ for a large constant $C > 0$, we obtain that

$$\begin{aligned} P\left\{\sup_{1\le\ell\le p_n}\left\|\sum_{t=2}^{n}\frac{\partial\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}}\right\| > C\delta_{2n}\right\} &\le \frac{C_1 n^{1-d/2}p_n}{C^d(\log(n))^{d/2}} + C_2 p_n\exp\left(-C_3C^2\log(n)\right) \\ &\le O\left(\frac{1}{n^{1+\varepsilon}}\right). \end{aligned}$$

(b2). Similar to (a1) in this proof, we have that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Lambda}}\left\|\frac{\partial\tilde{\sigma}^2_{\ell,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - \frac{\partial\sigma^2_{\ell,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\| \le C(\tilde{d}^2_\ell + \tilde{d}_\ell\sum_{k=0}^{t-1}\rho^k\sum_{j=1}^{m}|\epsilon_{t-k-j}| + \rho^t).$$

We also obtain that

$$\widetilde{\sigma}^2_{\ell,t}\left|\frac{1}{\sigma^2_{\ell,t}} - \frac{1}{\widetilde{\sigma}^2_{\ell,t}}\right| \le C(\tilde{d}^2_\ell + \tilde{d}_\ell + \rho^t), \frac{\widetilde{\sigma}^2_{\ell,t}}{\sigma^2_{\ell,t}} \le 1 + C(\tilde{d}^2_\ell + \tilde{d}_\ell + \rho^t).$$

As a result, for $i = 1, \cdots, m+s+1$, the i-th component of the difference $\left|\frac{\partial v_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}_i} - \frac{\partial\tilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}_i}\right|$ is bounded above by

$$\begin{aligned} &\left|\frac{\partial\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}_i} - \frac{\partial v_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}_i}\right| \\ &\le \left|\left(\frac{\epsilon^2_{\ell,t}}{\sigma^2_{\ell,t}} - \frac{\epsilon^2_{\ell,t}}{\widetilde{\sigma}^2_{\ell,t}}\right)\left(\frac{1}{\widetilde{\sigma}^2_{\ell,t}}\frac{\partial\tilde{\sigma}^2_{\ell,t}}{\partial\theta_i}\right) + \left(1 - \frac{\epsilon^2_{\ell,t}}{\sigma^2_{\ell,t}}\right)\left(\frac{1}{\widetilde{\sigma}^2_{\ell,t}} - \frac{1}{\sigma^2_{\ell,t}}\right)\frac{\partial\sigma^2_{\ell,t}}{\partial\theta_i}\right. \\ &\left. + \left(1 - \frac{\epsilon^2_{\ell,t}}{\sigma^2_{\ell,t}}\right)\frac{1}{\sigma^2_{\ell,t}}\left(\frac{\partial\sigma^2_{\ell,t}}{\partial\theta_i} - \frac{\partial\widetilde{\sigma}^2_{\ell,t}}{\partial\theta_i}\right)\right|(\boldsymbol{\theta}_{\ell,0}) + \frac{\tilde{d}^2_\ell + \tilde{d}_\ell|\epsilon_{\ell,t}|}{\sigma^2_{\ell,t}}\left|\frac{1}{\sigma^2_{\ell,t}}\frac{\partial\sigma^2_{\ell,t}}{\partial\theta_i}(\boldsymbol{\theta}_{\ell,0})\right| \\ &\le C(\tilde{d}^2_\ell + \tilde{d}_\ell + \rho^t)(1+\eta^2_{\ell,t})\left|1 + \frac{1}{\widetilde{\sigma}^2_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}\frac{\partial\widetilde{\sigma}^2_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\theta_i}\right|. \end{aligned}$$

Then it follows that, for $i = 1, \cdots, m+s+1$,

$$\begin{aligned} \left|\sum_{t=2}^{n}\frac{\partial v_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\theta_i} - \sum_{t=2}^{n}\frac{\partial\tilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\theta_i}\right| &\le C(\tilde{d}^2_\ell + \tilde{d}_\ell)\sum_{t=2}^{n}(1+\eta^2_{\ell,t})\left|1 + \frac{1}{\widetilde{\sigma}^2_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}\frac{\partial\widetilde{\sigma}^2_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\theta_i}\right| \\ &\quad + C\sum_{t=2}^{n}\rho^t(1+\eta^2_{\ell,t})\left|1 + \frac{1}{\widetilde{\sigma}^2_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}\frac{\partial\widetilde{\sigma}^2_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\theta_i}\right|. \end{aligned}$$

By Markov and bulkholder inequalities for martingales, we claim that there exists a constant $C > 0$ such that

$$P\left\{\sup_{1\le\ell\le p_n}\sum_{t=2}^{n}\rho^t(1+\eta^2_{\ell,t})\left|1 + \frac{1}{\widetilde{\sigma}^2_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}\frac{\partial\widetilde{\sigma}^2_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\theta_i}\right| > Cn^{1/2}\right\} = O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

and

$$P\left\{\sup_{1\le\ell\le p_n}\sum_{t=2}^{n}(1+\eta_{\ell,t}^2)\left|1+\frac{1}{\widetilde{\sigma}_{\ell,t}^2(\boldsymbol{\theta}_{\ell,0})}\frac{\partial\widetilde{\sigma}_{\ell,t}^2(\boldsymbol{\theta}_{\ell,0})}{\partial\theta_i}\right|>Cn\right\}=O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

Note that $\sup_{\ell\le p_n}|\tilde{d}_\ell|=\bar{O}_{a.s}(h_1^2+\delta_{3n})$. Hence, it follows that there exists a constant $C>0$ such that

$$P\left\{\sup_{1\le\ell\le p_n}n^{-1}\left\|\sum_{t=2}^{n}\frac{\partial v_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}}-\sum_{t=2}^{n}\frac{\partial\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}}\right\|>C(h_1^2+\delta_{3n})\right\}=O\left(\frac{1}{n^{1+\varepsilon}}\right),$$

and part (b2) follows.

(b3). $n^{-1}\sum_{t=2}^{n}\frac{\partial^2\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}$ can be expressed as

$$n^{-1}\sum_{t=2}^{n}\frac{\partial^2\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}=n^{-1}\sum_{t=2}^{n}\left\{\frac{\partial^2\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}-E\left(\frac{\partial^2\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right)\right\}+E\left\{\frac{\partial^2\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right\}.$$

Note that $\inf_{1\le\ell\le p_n}E\left\{\frac{\partial^2\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right\}$ is positive definite. It suffices to show that, for any constant $c>0$,

$$P\left\{\sup_{1\le\ell\le p_n}n^{-1}\left|\sum_{t=2}^{n}\left\{\frac{\partial^2\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}-E\left(\frac{\partial^2\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right)\right\}\right|>c\right\}=O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

Similar to (b1), we claim that there exist three positive constants C_1, C_2 and C_3 such that

$$\begin{aligned}&P\left\{\sup_{1\le\ell\le p_n}n^{-1}\left|\sum_{t=2}^{n}\left\{\frac{\partial^2\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}-E\left(\frac{\partial^2\widetilde{v}_{\ell,t}(\boldsymbol{\theta}_{\ell,0})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right)\right\}\right|>c\right\}\\&\le C_1\frac{p_nn}{(nc)^d}+C_2p_n\exp(-C_3n^2c^2)=O\left(\frac{1}{n^{1+\varepsilon}}\right).\end{aligned}$$

Part (b3) follows.

(b4) and **(b5)**. Together with the proof of (c) in Theorem 7.2 of Francq and Zakoian(2011), the proofs of these two parts can be proved in a similar fashion to (b2) and (b3).

Appendix D: Proof of Theorem 2

Define $\widehat{\mathbf{E}}_n=\widehat{\boldsymbol{\Phi}}(X_n^{\mathrm{T}}\widehat{\boldsymbol{\beta}})-\boldsymbol{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})$, $\widehat{\mathbf{F}}_n=\widehat{\boldsymbol{\Sigma}}_x(X_n)-\boldsymbol{\Sigma}_x(X_n)$. The difference $\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n)-\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)$ can be decomposed into four parts:

$$\begin{aligned}\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n)-\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)&=\widehat{\mathbf{E}}_n\widehat{\boldsymbol{\Sigma}}_x(X_n)\widehat{\mathbf{E}}_n^{\mathrm{T}}+\boldsymbol{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\widehat{\mathbf{F}}_n\{\boldsymbol{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}}+\left(\widehat{\boldsymbol{\Sigma}}_{0,n}-\boldsymbol{\Sigma}_{0,n}\right)\\&\quad+\left(\boldsymbol{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\widehat{\boldsymbol{\Sigma}}_x(X_n)\widehat{\mathbf{E}}_n^{\mathrm{T}}+\widehat{\mathbf{E}}_n\widehat{\boldsymbol{\Sigma}}_x(X_n)\left\{\boldsymbol{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\right\}^{\mathrm{T}}\right).\end{aligned}$$

We thus bound $\|\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n) - \mathrm{cov}(Y_{n+1}|\mathcal{F}_n)\|_{\mathbf{\Sigma}}^2$ by

$$4\left\|\widehat{\mathbf{E}}_n\widehat{\mathbf{\Sigma}}_x(X_n)\widehat{\mathbf{E}}_n^{\mathrm{T}}\right\|_{\mathbf{\Sigma}}^2 + 4\left\|\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\widehat{\mathbf{F}}_n\{\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}}\right\|_{\mathbf{\Sigma}}^2 + 4\left\|\widehat{\mathbf{\Sigma}}_{0,n} - \mathbf{\Sigma}_{0,n}\right\|_{\mathbf{\Sigma}}^2$$
$$+4\left\|\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\widehat{\mathbf{\Sigma}}_x(X_n)\widehat{\mathbf{E}}_n^{\mathrm{T}} + \widehat{\mathbf{E}}_n\widehat{\mathbf{\Sigma}}_x(X_n)\left\{\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\right\}^{\mathrm{T}}\right\|_{\mathbf{\Sigma}}^2.$$

To bound these terms, we first introduce the following two lemmas.

Lemma D.1. Suppose that Assumptions (A1)-(A5), (B1)-(B4) and (C1)-(C4) in Appendix A hold. Then there exists a large $C > 0$ such that

(i)
$$P\left\{\left\|\widehat{\mathbf{E}}_n\right\|_F^2 > Cp_n\left(h_1^4 + \frac{\log(n)}{nh_1}\right)\right\} \leq O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

(ii)
$$P\left\{\left\|\widehat{\mathbf{F}}_n\right\|_F^2 > C\left(h_2^4 + \frac{\log(n)}{nh_2^q}\right)\right\} \leq O\left(\frac{1}{n^2}\right).$$

Proof of Lemma D.1. (i) Observe that

$$\widehat{\mathbf{\Phi}}(X_n^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) - \mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{\Phi}'(X_n^{\mathrm{T}}\boldsymbol{\beta}^*)X_n^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \left(\widehat{\mathbf{\Phi}}(X_n^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) - \mathbf{\Phi}(X_n^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\right),$$

where $\boldsymbol{\beta}^*$ is between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. As a result,

$$\|\widehat{\mathbf{E}}_n\|_F^2 \leq 2\|\sup_{z\in\mathcal{Z}}\mathbf{\Phi}'(z)\|_2^2 \cdot \|X_n\|^2 \cdot \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 + 2\cdot\sup_{z\in\mathcal{Z}}\left\|\widehat{\mathbf{\Phi}}(z) - \mathbf{\Phi}(z)\right\|_F^2.$$

Note that $\|\sup_{z\in\mathcal{Z}}\mathbf{\Phi}'(z)\|_F^2 \cdot \|X_n\|_2 = O(p_n)$. Therefore, part (i) follows from Theorem 1(I) and (III).

(ii) Let $\widetilde{K}_{h_2,t}(\mathbf{u}) = \widetilde{K}_{h_2}(X_{t-1} - \mathbf{u})$ and $\varphi(X_t)$ be a bounded function uniformly over $X_t \in \mathcal{X}$. By following the proof of Theorem 5.3 in Fan and Yao (2003), we can see that there exists a large $C > 0$ such that

$$P\left\{\sup_{\mathbf{u}\in\mathcal{X}}\frac{1}{n}\left|\sum_{t=2}^{n}\varphi(X_t)\widetilde{K}_{h_2,t}(\mathbf{u}) - E\left\{\varphi(X_t)\widetilde{K}_{h_2,t}(\mathbf{u})\right\}\right| > C\sqrt{\frac{\log(n)}{nh_2^q}}\right\} \leq O\left(\frac{1}{n^2}\right).$$

By setting $\varphi(X_t) = 1, X_j, X_jX_k, (j,k = 1,\cdots,q)$, part (ii) follows.

Lemma D.2. Suppose that Assumptions (A1)-(A5), (B1)-(B4) and (C1) and (C3) in Appendix A hold. Then there exists $C > 0$ and small $\varepsilon > 0$ such that

$$P\left\{\sup_{1\leq\ell\leq p_n}\left|\frac{\widehat{\sigma}_{\ell,n+1}^2 - \sigma_{\ell,n+1}^2}{\sigma_{\ell,n+1}^2}\right| > C\left(h_1^2 + \delta_{3n}\right)\right\} \leq O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

Proof of Lemma D.2. Let $\mathbf{B}(i,j)$ be the (i,j)th element of the matrix $\mathbf{B}$ and $\mathbf{A}(i)$ be the ith entry of a vector $\mathbf{A}$. The conditional covariance $\widehat{\sigma}_{\ell,n+1}^2$ can be expressed as

$$\widehat{\sigma}_{\ell,n+1}^2 = \sum_{k=0}^{n}\widehat{\mathbf{B}}_\ell^k(1,1)\underline{\widehat{\mathbf{c}}}_{\ell,n+1-k}(1) + \sum_{i=1}^{s}\widehat{\mathbf{B}}_\ell^{n+1}(1,i)\underline{\widehat{\boldsymbol{\sigma}}}_{\ell,0}^2(i),$$

where $\widehat{\mathbf{B}}_\ell$ is the matrix obtained by replacing $\widehat{\gamma}_{\ell,j}$ by $\gamma_{\ell,j}$ in $\mathbf{B}_\ell$ and $\widehat{\underline{\mathbf{c}}}_{\ell,t}$ and $\widehat{\underline{\boldsymbol{\sigma}}}^2_{\ell,0}$ are defined accordingly. Note that the true conditional variance

$$\sigma^2_{\ell,n+1} = \sum_{k=0}^{n} \mathbf{B}^k_\ell(1,1)\underline{\mathbf{c}}_{\ell,n+1-k}(1) + \sum_{i=1}^{s} \mathbf{B}^{n+1}_\ell(1,i)\underline{\boldsymbol{\sigma}}^2_{\ell,0}(i).$$

We thus have that

$$\begin{aligned}\widehat{\sigma}^2_{\ell,n+1} - \sigma^2_{\ell,n+1} &= \sum_{k=0}^{n} \widehat{\mathbf{B}}^k_\ell(1,1)\left(\widehat{\underline{\mathbf{c}}}_{\ell,n+1-k}(1) - \underline{\mathbf{c}}_{\ell,n+1-k}(1)\right) + \sum_{k=1}^{n}\left(\widehat{\mathbf{B}}^k_\ell - \mathbf{B}^k_\ell\right)(1,1)\underline{\mathbf{c}}_{\ell,n+1-k}(1) \\ &+ \sum_{i=1}^{s}\left(\widehat{\mathbf{B}}^{n+1}_\ell(1,i)\widehat{\underline{\boldsymbol{\sigma}}}^2_{\ell,0}(i) - \mathbf{B}^{n+1}_\ell(1,i)\underline{\boldsymbol{\sigma}}^2_{\ell,0}(i)\right) = U_{\ell,1} + U_{\ell,2} + U_{\ell,3}.\end{aligned}$$

(a) Consider the term $U_{\ell,1}$ and observe that $\|\widehat{\underline{\mathbf{c}}}_{\ell,t} - \underline{\mathbf{c}}_{\ell,t}\| \leq |\widehat{\alpha}_{\ell,0} - \alpha_{\ell,0}| + \tilde{d}^2_\ell + 2\tilde{d}_\ell \sum_{j=1}^{m} |\epsilon_{\ell,t-j}|$. Then, there exists a constant $C > 0$ such that

$$|U_{\ell,1}| \leq C\big(|\widehat{\alpha}_{\ell,0} - \alpha_{\ell,0}| + \tilde{d}^2_\ell + \tilde{d}_\ell \sum_{j=1}^{m}\sum_{k=1}^{n} \rho^k |\epsilon_{\ell,t-k-j}|\big).$$

Since $\sum_{j=1}^{m}\sum_{k=1}^{n} \rho^k |\epsilon_{\ell,t-k-j}| / \sigma^2_{\ell,n+1}$ is bounded and $\sigma^2_{\ell,n+1} \geq \alpha_{\ell,0} > 0$, this means that

$$\left|\frac{U_{\ell,1}}{\sigma^2_{\ell,n+1}}\right| \leq C\big(|\widehat{\alpha}_{\ell,0} - \alpha_{\ell,0}| + \tilde{d}_\ell\big).$$

and consequently, there exists a large constant $C > 0$ such that

$$P\left\{\sup_{1\leq\ell\leq p_n}\left|\frac{U_{\ell,1}}{\sigma^2_{\ell,n+1}}\right| > C(h_1^2 + \delta_{3n})\right\} = O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

(b) Consider the term $U_{\ell,2}$. Denote $\hat{\delta}_\ell = \sup_{1\leq i\leq s} |\hat{\gamma}_{\ell,i} - \gamma_{\ell,i}|/\gamma_{\ell,i}$. By the definition of $\widehat{\mathbf{B}}_\ell$ and $\mathbf{B}$, it is seen that

$$\left|\frac{\widehat{\mathbf{B}}^k_\ell(1,1) - \mathbf{B}^k_\ell(1,1)}{\mathbf{B}^k_\ell(1,1)}\right| \leq \max\{|(1-\hat{\delta}_\ell)^k - 1|, |(1+\hat{\delta}_\ell)^k - 1|\} \leq 2\hat{\delta}_\ell k(1+\hat{\delta}_\ell)^{k-1},$$

for small $\hat{\delta}_\ell$. Note that $\sigma^2_{\ell,n+1} \geq \alpha_{\ell,0} + \mathbf{B}^k_\ell(1,1)\underline{\mathbf{c}}_{\ell,n+1-k}(1)$ and the relation $x/(1+x) \leq x^\delta$ for all $x \geq 0$ and $\delta \in (0,1)$. We have that

$$\begin{aligned}\left|\frac{U_{\ell,2}}{\sigma^2_{\ell,n+1}}\right| &\leq \sum_{k=1}^{n}\left|\frac{\left(\widehat{\mathbf{B}}^k_\ell - \mathbf{B}^k_\ell\right)(1,1)}{\mathbf{B}^k(1,1)}\right| \frac{\mathbf{B}^k(1,1)\underline{\mathbf{c}}_{\ell,n+1-k}(1)}{\alpha_{\ell,0} + \mathbf{B}^k_\ell(1,1)\underline{\mathbf{c}}_{\ell,n+1-k}(1)} \\ &\leq 2\hat{\delta}_\ell \sum_{k=1}^{n} k(1+\hat{\delta}_\ell)^k \rho^{k\delta} \underline{\mathbf{c}}^\delta_{\ell,n+1-k}(1).\end{aligned}$$

Hence, by choosing a suitable but small δ, it follows from Theorem 1(IV) that there exists a large positive constant C such that

$$P\left\{\sup_{1\le\ell\le p_n}\left|\frac{U_{\ell,2}}{\sigma^2_{\ell,n+1}}\right|>C(h_1^2+\delta_{3n})\right\}\le P\left\{\sup_{1\le\ell\le p_n}\hat{\delta}_\ell>C(h_1^2+\delta_{3n})\right\}=O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

(c) It is easy to see that $\|U_{\ell,3}\|/\sigma^2_{\ell,n+1}$ is bounded. Lemma D.2 follows.

Proof of Theorem 2.

(a). Now we bound $\left\|\widehat{\mathbf{E}}_n\widehat{\mathbf{\Sigma}}_x(X_n)\widehat{\mathbf{E}}_n^{\mathrm{T}}\right\|^2_{\mathbf{\Sigma}}$. Observe that

$$p_n\left\|\widehat{\mathbf{E}}_n\widehat{\mathbf{\Sigma}}_x(X_n)\widehat{\mathbf{E}}_n^{\mathrm{T}}\right\|^2_{\mathbf{\Sigma}}=\lambda^2_{\max}\left(\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1}\right)\lambda^2_{\max}\left(\widehat{\mathbf{\Sigma}}_x(X_n)\right)\left\|\widehat{\mathbf{E}}_n\right\|^4_F.$$

Hence, it follows from Lemma D.1 that there exists $C>0$ such that

$$P\left\{\left\|\widehat{\mathbf{E}}_n\widehat{\mathbf{\Sigma}}_x(X_n)\widehat{\mathbf{E}}_n^{\mathrm{T}}\right\|^2_{\mathbf{\Sigma}}>Cp_n\left(h_1^8+\frac{\log^2(n)}{(nh_1)^2}\right)\right\}=O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

(b). We bound $\left\|\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\widehat{\mathbf{F}}_n\{\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}}\right\|^2_{\mathbf{\Sigma}}$. Note that $\|\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}\left(\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)\right)^{-1}\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\|=O(1)$. Hence, we have that $\left\|\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\widehat{\mathbf{F}}_n\{\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}}\right\|^2_{\mathbf{\Sigma}}\le O(p_n^{-1})\|\widehat{\mathbf{F}}_n\|^2_F$, and consequently, by Lemma D.1, there exists $C>0$ such that

$$P\left\{\left\|\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\widehat{\mathbf{F}}_n\{\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}}\right\|^2_{\mathbf{\Sigma}}>Cp_n^{-1}\left(h_2^4+\frac{\log(n)}{nh_2^q}\right)\right\}=O\left(\frac{1}{n^2}\right).$$

(c). We bound $\left\|\widehat{\mathbf{\Sigma}}_{0,n}-\mathbf{\Sigma}_{0,n}\right\|^2_{\mathbf{\Sigma}}$. Note that

$$\begin{aligned}\left\|\widehat{\mathbf{\Sigma}}_{0,n}-\mathbf{\Sigma}_{0,n}\right\|^2_{\mathbf{\Sigma}}&\le\left\|\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1/2}\left(\widehat{\mathbf{\Sigma}}_{0,n}-\mathbf{\Sigma}_{0,n}\right)\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1/2}\right\|^2_2\\&\le\sup_{1\le\ell\le p_n}\left|\frac{\widehat{\sigma}^2_{\ell,n+1}-\sigma^2_{\ell,n+1}}{\sigma^2_{\ell,n+1}}\right|^2.\end{aligned}$$

Hence we obtain from Lemma D.2 that there exists $C>0$ such that

$$P\left\{\left\|\widehat{\mathbf{\Sigma}}_{0,n}-\mathbf{\Sigma}_{0,n}\right\|^2_{\mathbf{\Sigma}}>C\left(h_1^4+\frac{\log(n)}{nh_1}\right)\right\}=O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

(d). Now we bound $\left\|\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\widehat{\mathbf{\Sigma}}_x(X_n)\widehat{\mathbf{E}}_n^{\mathrm{T}}+\widehat{\mathbf{E}}_n\widehat{\mathbf{\Sigma}}_x(X_n)\{\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}}\right\|^2_{\mathbf{\Sigma}}$. Note that for two $q\times q$ matrix $\mathbf{A}$ and $\mathbf{B}$, $\|\mathbf{A}+\mathbf{B}\|^2_F\le 2(\|\mathbf{A}\|^2_F+\|\mathbf{B}\|^2_F)$, $\|\mathbf{AB}\|_F\le\|\mathbf{A}\|_F\|\mathbf{B}\|_F$ and $|\mathrm{tr}(\mathbf{AB})|\le\|\mathbf{A}\|_F\|\mathbf{B}\|_F$. We have that

$$\begin{aligned}&p_n\left\|\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\widehat{\mathbf{\Sigma}}_x(X_n)\widehat{\mathbf{E}}_n^{\mathrm{T}}+\widehat{\mathbf{E}}_n\widehat{\mathbf{\Sigma}}_x(X_n)\{\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}}\right\|^2_{\mathbf{\Sigma}}\\&\le 2\|\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1/2}\mathbf{\Phi}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\widehat{\mathbf{\Sigma}}_{\mathbf{X}}(\mathbf{x})\widehat{\mathbf{E}}_n^{\mathrm{T}}\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1/2}\|^2_F\\&=2\mathrm{tr}\left(\widehat{\mathbf{\Sigma}}_x(X_n)\widehat{\mathbf{E}}_n^{\mathrm{T}}\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1}\widehat{\mathbf{E}}_n\widehat{\mathbf{\Sigma}}_x(X_n)\{\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}}\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1}\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\right)\\&\le 2q^2\|\widehat{\mathbf{\Sigma}}_x(X_n)\|^2_F\lambda_{\max}\left(\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1}\right)\lambda_{\max}\left(\{\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}}\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1}\mathbf{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\right)\cdot\|\widehat{\mathbf{E}}_n\|^2_F.\end{aligned}$$

Hence, by Lemma D.1 , together with $\lambda_{\max}\big(\{\boldsymbol{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}}\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1}\boldsymbol{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\big) = O(1)$, it follows that there exists $C > 0$ such that

$$P\left\{\left\|\boldsymbol{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\widehat{\boldsymbol{\Sigma}}_x(X_n)\widehat{\mathbf{E}}_n^{\mathrm{T}} + \widehat{\mathbf{E}}_n\widehat{\boldsymbol{\Sigma}}_x(X_n)\{\boldsymbol{\Phi}(X_n^{\mathrm{T}}\boldsymbol{\beta})\}^{\mathrm{T}}\right\|_{\boldsymbol{\Sigma}}^2 > C\left(h_1^4 + \frac{\log n}{nh_1}\right)\right\} \leq O\left(\frac{1}{n^{1+\varepsilon}}\right).$$

Combining (a)-(d), Theorem 2 follows. This completes the proof of Theorem 2.

References

Bosq, D. (1996). Nonparametric statistics for stochastic processes (Vol. 110). New York: Springer.

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.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M.P. (1997). Generalized partially linear single-Index models. *Journal of American Statistical Association*, **92**, 477-489.

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

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. and Yao, Q. (2003). Nonlinear Time Series: Nonparametric and Parametric Methods. Springer.

Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of Royal Statistical Society B*, **65**, 57-80.

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

Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*, **27**, 715-731.

Francq, C. and Zakoian, J. M. (2011). GARCH models: structure, statistical inference and financial applications. John Wiley & Sons.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.*, **21**, 157-178.

Kong, E. and Xia, Y. (2014). An adaptive composite quantile approach to dimension reduction. *Annals of Statistics*, **42**, 1657-1688.

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

Li, J. and Zhang, W. (2011). A semiparametric threshold model for censored longitudinal data analysis. *Journal of the American Statistical Association*, **106**, 685-696.

Lindner, A. M. (2009). Stationarity, mixing, distributional properties and moments of GARCH(p, q) processes. In *Handbook of financial time series* (pp. 43-69). Springer Berlin Heidelberg.

Liu, W., Xiao, H. and Wu, W. B. (2013). Probability and moment inequalities under dependence. *Statist. Sinca.*, **23(3)**, 1257-1272.

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.

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

Sun, Y., Yan, H., Zhang, W. and Lu, Z. (2014). A semiparametric spatial dynamic model. *The Annals of Statistics*, **42**, 700-727.

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.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, **94**, 1a17.

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.

Xia, Y., Tong, H., and Li, W. K. (2002). Single-index volatility models and estimation. *Statist. Sinica.*, **12(3)**, 785-799.

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

Zhang, W., Fan, J. and Sun, Y. (2009). A semiparametric model for cluster data. *The Annals of Statistics*, **37**, 2377-2408.