Factor and Idiosyncratic VAR-Itô Volatility Matrix Models for Heavy-Tailed High-Frequency Financial Observations[*]

Minseok Shin[a], Donggyu Kim[b][†] Yazhen Wang[c], and Jianqing Fan[d][†]

[a]Korea Advanced Institute of Science and Technology (KAIST)

[b]University of California, Riverside

[c]University of Wisconsin-Madison

[d]Princeton University

July 8, 2024

Abstract

This paper introduces a novel Itô diffusion process for both factor and idiosyncratic volatility matrices whose eigenvalues follow the vector auto-regressive (VAR) model. We call it the factor and idiosyncratic VAR-Itô (FIVAR-Itô) model. The FIVAR-Itô model accounts for the dynamics of the factor and idiosyncratic volatilities and includes many parameters. In addition, many empirical studies have shown that high-frequency stock returns and volatilities often exhibit heavy tails. To handle these two problems simultaneously, we propose a penalized optimization procedure with a truncation scheme for parameter estimation. We apply the proposed parameter estimation procedure to predicting large volatility matrices and establish its asymptotic properties.

1 Introduction

Volatility analysis for high-frequency financial data is a vibrant research area in financial econometrics and statistics. With the wide availability of high-frequency financial data, several well-performing non-parametric estimation methods have been developed to estimate integrated volatilities (Aït-Sahalia et al., 2010; Barndorff-Nielsen et al., 2008, 2011a; Bibinger et al., 2014; Christensen et al., 2010; Fan and Kim, 2018; Fan and Wang, 2007; Jacod et al., 2009; Shin et al., 2023; Xiu, 2010; Zhang et al., 2005; Zhang, 2006, 2011). With these non-parametric (daily) realized volatility estimators, parametric models have been developed to account for volatility dynamics over time. Examples include the realized volatility-based modeling approaches (Andersen et al., 2003), the heterogeneous auto-regressive (HAR) models (Corsi, 2009), the realized GARCH models (Hansen et al., 2012), the high-frequency-based volatility (HEAVY) models (Shephard and Sheppard, 2010), and the unified GARCH-Itô models (Kim and Wang, 2016; Song et al., 2021). Their empirical studies showed that incorporating high-frequency information, such as realized volatility, helps capture the volatility dynamics for a finite number of assets. However, in financial practice, we often need to handle a large number of assets, which leads to an excessive number of parameters for typical sample sizes. To overcome this problem, the approximate factor model structure is often imposed on volatility matrices (Fan et al., 2013). For example, high-dimensional factor-based Itô processes are widely utilized with the sparsity assumption on the idiosyncratic volatility (Aït-Sahalia and Xiu, 2017; Fan et al., 2016; Kim et al., 2018; Kong, 2018). Recently, Kim and Fan (2019) developed the factor GARCH-Itô model, based on the high-dimensional factor-based Itô processes. The factor GARCH-Itô model assumes that the eigenvalue sequence of the latent factor volatility matrices admits some unified GARCH-Itô model structure (Kim and Wang, 2016) so that the dynamics of the volatility can be explained by the factors. See also Hetland et al. (2023) for the low-dimensional low-frequency setting and Kim et al. (2022, 2023) for the high-dimensional high-frequency setting. We note that when employing the approximate factor model structure, the

existing literature does not model the idiosyncratic volatility and assumes that the idiosyncratic volatility process is martingale.

However, several empirical studies indicate that idiosyncratic volatility also has a dynamic structure, and it comprises a large proportion of the total volatility (Barigozzi and Hallin, 2016; Connor et al., 2006; Herskovic et al., 2016). To provide evidence of the existence of the dynamics in the idiosyncratic process for the high-frequency financial returns, we estimated the 200 daily eigenvalues of the idiosyncratic volatility matrix based on the top 200 large trading volume stocks in the S&P 500 index. The estimation procedure will be described in Section 4.2 and Section 5.2. Figure 1 depicts the distribution of the first-order auto-correlations of the 200 time series of 200 daily estimated eigenvalues as well as the ACF plots for the time series of daily eigenvalue estimates of the 1st, 50th, 150th, and 200th eigenvalues. We note that other eigenvalues also have similar time series structures. Figure 1 shows that the lag-1 autocorrelations are quite strong, which supports a dynamic structure in the eigenvalue processes of the idiosyncratic volatility. In addition, these estimated eigenvalues exhibit fairly long memories, with significant autocorrelation of lags of about 1 to 4 weeks. Thus, simultaneously modeling the idiosyncratic volatility as well as the factor volatility is important to capture volatility dynamics. On the other hand, since the
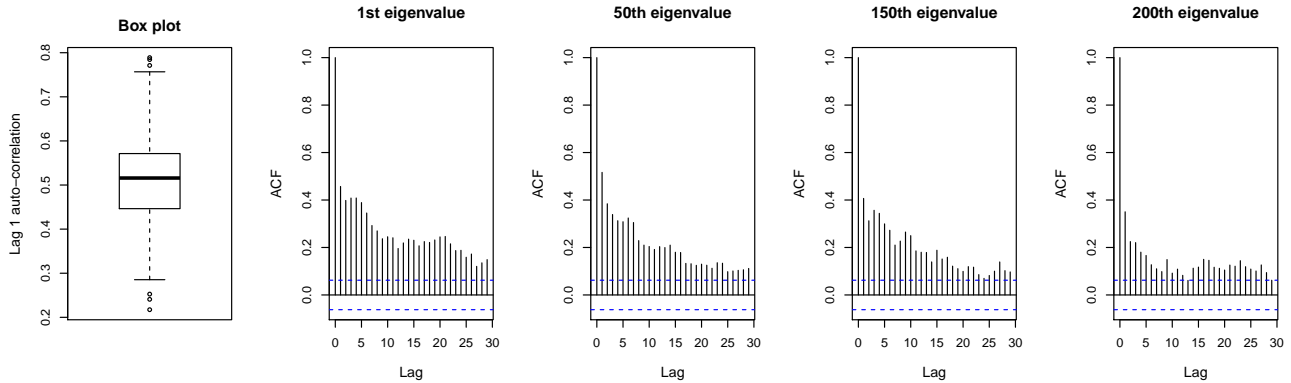
Figure 1: The box plot of the first-order auto-correlations for the time series of 200 daily estimated eigenvalues of the idiosyncratic volatility matrix and the ACF plots for the time series of the 1st, 50th, 150th, and 200th eigenvalues.

dimension of the idiosyncratic volatility is large, modeling the factor and idiosyncratic volatilities simultaneously results in the problem of over-parameterization. To address this issue, the sparsity of model parameters is often imposed, and high-dimensional estimation procedures, such as LASSO

(Tibshirani, 1996) and SCAD (Fan and Li, 2001), which are usually developed under a sub-Gaussian tail condition, are employed. However, this sub-Gaussian assumption is at odds with the empirical observations that the financial market exhibits heavy tails. For example, Figure 2 shows the boxplot of the 200 log kurtoses for the daily jump adjusted pre-averaging volatility estimators (Aït-Sahalia and Xiu, 2016; Christensen et al., 2010; Jacod et al., 2009) for 997 trading days in the period 2016–2019. The daily jump adjusted pre-averaging volatility estimators are estimated using 1-min log-returns of the most liquid 200 assets in the S&P 500 index. Detailed estimation procedure is presented in (A.1) in the Appendix. From Figure 2, we can see that the volatility processes have heavy-tailed distributions. See also Cont (2001); Fan and Kim (2018); Mao and Zhang (2018); Shin et al. (2023). Thus, the high-dimensional estimation procedure developed under the sub-Gaussian tail condition is inappropriate. These stylized features lead to the demands for developing a diffusion process for both factor and idiosyncratic volatilities with heavy-tailed observations.
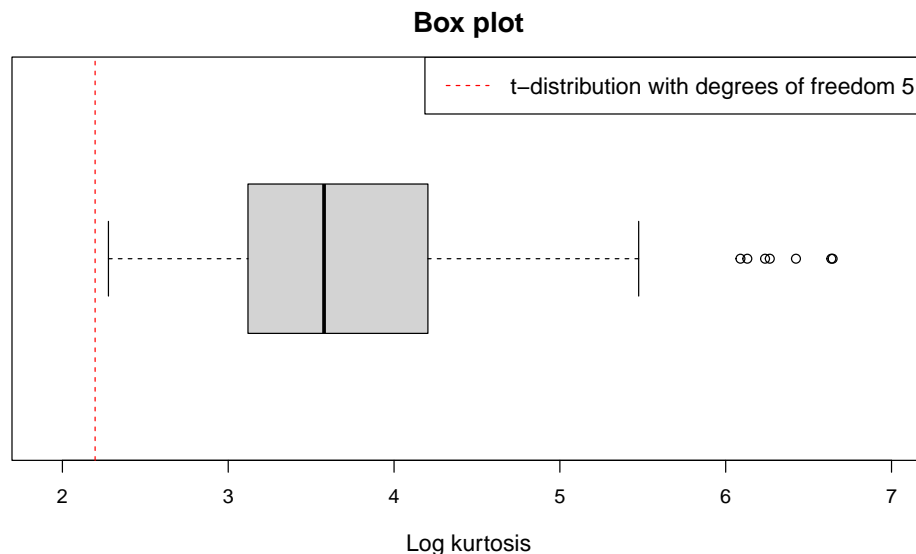
Figure 2: The boxplot of the 200 log kurtoses obtained from the daily jump adjusted pre-averaging volatility estimators for 997 trading days in the period 2016–2019. The daily jump adjusted pre-averaging volatility estimators are estimated using 1-min log-returns of the most liquid 200 assets in the S&P 500 index. The red dash represents the kurtosis of the t_5-distribution.

In this paper, we introduce a novel Itô diffusion process to account for the dynamics in the factor and idiosyncratic volatilities, based on the VAR model with heavy-tailed innovations. Specifically, it is assumed that the eigenvectors of the latent factor and idiosyncratic volatility matrices do not

vary over a time period. In contrast, we allow the eigenvalues to evolve with time and impose a parametric dynamic structure. In particular, the instantaneous eigenvalue processes of the latent factor and idiosyncratic instantaneous volatility matrices are continuous with respect to time and have a VAR structure at integer time points so that it is a form of some interpolation of the VAR structure. This structure makes the integrated eigenvalues have the VAR structure, and thus, the dynamics of the volatility can be explained by both the factor and idiosyncratic components. We call it the factor and idiosyncratic VAR-Itô (FIVAR-Itô) model. When it comes to estimating model parameters, the high dimensionality of the idiosyncratic volatility matrix causes over-parameterization. Furthermore, we allow the heavy-tailedness based on the bounded c_ϵ-th moment condition for $c_\epsilon > 4$. It is assumed that the model parameters are sparse so that an ℓ_1-penalty, such as LASSO, can be employed. The usual ℓ_1-penalty method does not work under the heavy-tailedness (Sun et al., 2020) and a Huber loss is employed to address this issue (Huber, 1964). We show that the proposed estimation procedure has robustness with the desirable convergence rate. We also propose a procedure for large volatility matrix prediction and investigate its asymptotic properties.

The rest of the paper is organized as follows. Section 2 introduces the FIVAR-Itô model, based on the high-dimensional factor-based Itô diffusion process, and investigates its properties. Section 3 proposes the robust parameter estimation method for a high-dimensional VAR model with heavy-tailedness and establishes its concentration properties. In Section 4, we apply the proposed estimator to large volatility matrix prediction. In Section 5, we conduct a simulation study to check the finite sample performance of the proposed estimator and apply the estimation method to high-frequency trading data. The conclusion is presented in Section 6, and the technical proofs are deferred to the Appendix.

Before closing this section, let us introduce some notations. For a given $p_1 \times p_2$ matrix $\mathbf{M} = (M_{ij})$, let

$$\|\mathbf{M}\|_1 = \max_{1 \leq j \leq p_2} \sum_{i=1}^{p_1} |M_{ij}|, \quad \|\mathbf{M}\|_\infty = \max_{1 \leq i \leq p_1} \sum_{j=1}^{p_2} |M_{ij}|, \quad \|\mathbf{M}\|_{\max} = \max_{i,j} |M_{ij}|.$$

Note that $\| \cdot \|_{\max}$ is not a matrix norm in general, but can be interpreted as a vector norm. The

matrix spectral norm $\|\mathbf{M}\|_2$ is the square root of the largest eigenvalue of $\mathbf{M}\mathbf{M}^\top$ and the Frobenius norm of $\mathbf{M}$ is denoted by $\|\mathbf{M}\|_F = \sqrt{\operatorname{tr}(\mathbf{M}^\top \mathbf{M})}$. When $\mathbf{M}$ is a square matrix, the spectral radius $\rho(\mathbf{M})$ is the largest value of the absolute eigenvalues of $\mathbf{M}$. For any vector $\mathbf{x} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$ and $q \geq 1$, the ℓ_q norm $\|\mathbf{x}\|_q = (\sum_{i=1}^p |x_i|^q)^{1/q}$. For any vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, we set $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$. For a function $f : \mathbb{R}^p \to \mathbb{R}$, its gradient vector is denoted by $\nabla f \in \mathbb{R}^p$ as long as it exists. We denote $\|Z\|_{L_q} = \{\mathrm{E}\,(|Z|^q)\}^{1/q}$ for a random variable $Z \in \mathbb{R}$ and $q \geq 1$. The half-vectorization, vech $(\mathbf{M})$, of the matrix $\mathbf{M}$ is the column vector obtained by vectorizing only the lower triangular part of $\mathbf{M}$. Also, tr $(\mathbf{M})$ is the trace of $\mathbf{M}$ and det $(\mathbf{M})$ is the determinant of $\mathbf{M}$. Diag $(\mathbf{M})$ denotes the square diagonal matrix with the elements of the main diagonal of $\mathbf{M}$. C's denote generic positive constants whose values are free of other parameters and may change from appearance to appearance.

2   FIVAR-Itô model

Let $\mathbf{X}(t) = (X_1(t), \ldots, X_p(t))^\top$ be the vector of true log-prices of p assets at time t. To account for the cross-sectional dependence in financial asset prices, we employ the following factor-based jump-diffusion model:

$$d\mathbf{X}(t) = \boldsymbol{\mu}(t)dt + \mathbf{B}(t)d\mathbf{f}(t) + d\mathbf{u}(t) + \mathbf{J}(t)d\boldsymbol{\Lambda}(t), \tag{2.1}$$

where $\boldsymbol{\mu}(t) \in \mathbb{R}^p$ is a drift vector, $\mathbf{B}(t) \in \mathbb{R}^{p \times r}$ is an unknown factor loading matrix, $\mathbf{f}(t) \in \mathbb{R}^r$ is a latent factor process, and $\mathbf{u}(t) \in \mathbb{R}^p$ is an idiosyncratic process. For the jump part, $\mathbf{J}(t) = (J_1(t), \ldots, J_p(t))^\top$ is a jump size vector, and $\boldsymbol{\Lambda}(t) = (\Lambda_1(t), \ldots, \Lambda_p(t))^\top$ is a p-dimensional Poisson process with an intensity $\mathbf{I}(t) = (I_1(t), \ldots, I_p(t))^\top$. It is assumed that the factor and idiosyncratic processes $\mathbf{f}(t)$ and $\mathbf{u}(t)$ follow the continuous-time diffusion models:

$$d\mathbf{f}(t) = \boldsymbol{\vartheta}^\top(t)d\mathbf{W}(t) \quad \text{and} \quad d\mathbf{u}(t) = \boldsymbol{\Phi}^\top(t)d\mathbf{W}^*(t),$$

where $\boldsymbol{\vartheta}(t)$ and $\boldsymbol{\Phi}(t)$ are r by r and p by p instantaneous volatility matrices, respectively, and $\mathbf{W}(t)$ and $\mathbf{W}^*(t)$ are r-dimensional and p-dimensional independent Brownian motions, respectively.

Stochastic processes $\mathbf{X}(t)$, $\boldsymbol{\mu}(t)$, $\mathbf{B}(t)$, $\mathbf{f}(t)$, $\mathbf{u}(t)$, $\boldsymbol{\vartheta}(t)$, and $\boldsymbol{\Phi}(t)$ are defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t, t \in [0, \infty)\}, P)$ with filtration $\mathcal{F}_t$ satisfying the usual conditions, that is, the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t, t \in [0, \infty)\}, P)$ is complete and the filtration $\mathcal{F}_t$ is right-continuous. The instantaneous volatility matrix of the log-price $\mathbf{X}(t)$ is

$$\boldsymbol{\gamma}(t) = (\gamma_{ij}(t))_{1 \leq i, j \leq p} = \mathbf{B}(t)\boldsymbol{\vartheta}^\top(t)\boldsymbol{\vartheta}(t)\mathbf{B}^\top(t) + \boldsymbol{\Phi}^\top(t)\boldsymbol{\Phi}(t), \tag{2.2}$$

and the integrated volatility for the dth day is

$$\boldsymbol{\Gamma}_d = (\Gamma_{d,ij})_{i,j=1,\ldots,p} = \int_{d-1}^d \boldsymbol{\gamma}(t)dt = \boldsymbol{\Psi}_d + \boldsymbol{\Sigma}_d,$$

where $\boldsymbol{\Psi}_d = \int_{d-1}^d \mathbf{B}(t)\boldsymbol{\vartheta}^\top(t)\boldsymbol{\vartheta}(t)\mathbf{B}^\top(t)dt$ and $\boldsymbol{\Sigma}_d = \int_{d-1}^d \boldsymbol{\Phi}^\top(t)\boldsymbol{\Phi}(t)dt$ for $d \in \mathbb{N}$.

Let $\mathbf{q}_{t,1}^F, \ldots, \mathbf{q}_{t,r}^F$ and $\lambda_{t,1}(\boldsymbol{\theta}_1), \ldots, \lambda_{t,r}(\boldsymbol{\theta}_r)$ be the eigenvectors and eigenvalues of the instantaneous factor volatility matrix $\mathbf{B}(t)\boldsymbol{\vartheta}^\top(t)\boldsymbol{\vartheta}(t)\mathbf{B}^\top(t)/p$, respectively, and $\mathbf{q}_{t,1}^I, \ldots, \mathbf{q}_{t,p}^I$ and $\lambda_{t,r+1}(\boldsymbol{\theta}_{r+1})$, $\ldots, \lambda_{t,p+r}(\boldsymbol{\theta}_{p+r})$ be the eigenvectors and eigenvalues of the instantaneous idiosyncratic volatility matrix $\boldsymbol{\Phi}^\top(t)\boldsymbol{\Phi}(t)$, respectively. We note that in the high-dimensional factor model, the eigenvalues of the factor volatility matrix are usually assumed to diverge at the order of the dimension p. Therefore, to match the sizes of the factor and idiosyncratic parts, we divide the factor part by p. In this paper, to distinguish notations for the factor and idiosyncratic parts, we use subscript and superscript of F and I to their associated quantities, respectively. In the latent factor model, to identify the latent factor loading matrix and factors, it is often assumed that the latent factor loading matrix is orthonormal and the latent factors have a diagonal covariance matrix, which implies that the eigenvectors and eigenvalues are related to the factor loading matrix and factors, respectively (Aït-Sahalia and Xiu, 2017; Fan et al., 2013; Kim and Fan, 2019). In this paper, we also consider the eigenvalues as the latent factor and idiosyncratic associated variables. It is assumed that the eigenvectors, $\mathbf{q}_{t,i}^F$, $i = 1, \ldots, r$, are constant over time, that is, $\mathbf{q}_{t,i}^F = \mathbf{q}_i^F$ for $i = 1, \ldots, r$ and $t \in [0, \infty)$. Also, it is assumed that $\mathbf{q}_{t,i}^I = \mathbf{q}_i^I$ for $i = 1, \ldots, p$ and $t \in [0, \infty)$. We note that the constant assumption can be relaxed to the constant eigenvectors for each day. However, Kim and Fan (2019) shows that the estimation procedures with time-invariant eigenvec-

tors perform better. In light of this, it is assumed that the eigenvectors are constant over time, and hence the volatility dynamics are driven by those of the eigenvalues. Thus, to capture the volatility dynamics, we model the latent factor and idiosyncratic variables by the following factor and idiosyncratic VAR-Itô (FIVAR-Itô) model.

Definition 1. *We call a log-price vector $\mathbf{X}(t)$, $t \in [0, \infty)$, to follow a FIVAR(h)-Itô model if its associated values satisfy for $i = 1, \ldots, p + r$ and $t \in [0, \infty)$,*

$$
\begin{aligned}
\lambda_{t,i}(\boldsymbol{\theta}_i) &= (\lceil t \rceil - t)\lambda_{\lceil t-1 \rceil, i}(\boldsymbol{\theta}_i) + \sum_{j=1}^{p+r} \zeta_{1,i,j} \left\{ \int_{\lceil t-1 \rceil}^{t} \lambda_{s,j}(\boldsymbol{\theta}_j)\, ds \right\} + (\lceil t \rceil - t) \int_{\lceil t-1 \rceil}^{t} J_{\lambda,i}(s)\, d\widetilde{\Lambda}_{\lambda,i}(s) \\
&\quad + (t - \lceil t \rceil + 1)\left(a_i + \sum_{k=2}^{h} \sum_{j=1}^{p+r} \zeta_{k,i,j} \left\{ \int_{\lceil t \rceil - k}^{\lceil t \rceil - k + 1} \lambda_{s,j}(\boldsymbol{\theta}_j)\, ds \right\} \right) + (\lceil t \rceil - t) Z_{i,t}^2,
\end{aligned}
$$

where $\boldsymbol{\theta}_i = (\zeta_{k,i,j})_{1 \leq k \leq h, 1 \leq j \leq p+r}$ is the model parameter, $\lceil t \rceil$ is a ceiling function, which is the smallest integer greater than or equal to t, $J_{\lambda,i}(t)$ is a jump size process, $\widetilde{\Lambda}_{\lambda,i}(t)$ is a compensated Poisson process, and $Z_{i,t} = \int_{\lceil t-1 \rceil}^{t} z_{i,t}\, dW_{i,t}$, where $W_{i,t}$ is a standard Brownian motion and $z_{i,t}$ is a continuous process over each integer time interval.

In the proposed FIVAR-Itô model, the instantaneous eigenvalues at the integer time points (time unit is usually day) satisfy the VAR(h)-type structure as follows:

$$
\lambda_{d,i}(\boldsymbol{\theta}_i) = a_i + \sum_{k=1}^{h} \sum_{j=1}^{p+r} \zeta_{k,i,j} \left\{ \int_{d-k}^{d-k+1} \lambda_{t,j}(\boldsymbol{\theta}_j)\, dt \right\}. \tag{2.3}
$$

Thus, under the FIVAR-Itô model, the instantaneous eigenvalue process is an interpolation of the VAR structure. By introducing the jump process for the instantaneous eigenvalue process, we can account for the co-jumps with the return process (Bandi and Reno, 2016; Bibinger and Winkelmann, 2018; Jacod et al., 2017a; Jacod and Todorov, 2010). We note that to guarantee the positiveness of the eigenvalue process, we need some lower bound condition for the jump process, such as $J_{\lambda,i}(t)\widetilde{\Lambda}_{\lambda,i}(t) \geq -c$ a.s. for any t and some positive constant c that is related with the lower bound of the continuous part of the instantaneous eigenvalue process (e.g., a_i). The random fluctuation is modeled by $Z_{i,t}^2$, which is included in integrated eigenvalues. Thus, the volatility process is not deterministic. Furthermore, we model the heavy-tailedness by the heavy-tailed

8

random fluctuation $Z_{i,t}^2$.

The following proposition investigates the low-frequency time series structure of the integrated eigenvalues, which will be used for statistical inferences.

Proposition 1. *Let $\boldsymbol{\zeta}_k = (\zeta_{k,i,j})_{1 \leq i,j \leq p+r}$ for all $1 \leq k \leq h$, $\det(\boldsymbol{\zeta}_1) \neq 0$, and the spectral radius of $\boldsymbol{\zeta}_1$, $\rho(\boldsymbol{\zeta}_1) < 1$. Then, we have the following iterative relations for the FIVAR(h)-Itô model:*

$$\boldsymbol{\xi}_d = \boldsymbol{\nu} + \sum_{k=1}^h \mathbf{A}_k \boldsymbol{\xi}_{d-k} + \boldsymbol{\epsilon}_d \ \ a.s., \tag{2.4}$$

where $\boldsymbol{\xi}_d = (\xi_{d,1}, \ldots, \xi_{d,p+r})^\top = \int_{d-1}^d \boldsymbol{\lambda}_t(\boldsymbol{\theta}) \, dt$, $\boldsymbol{\lambda}_t(\boldsymbol{\theta}) = (\lambda_{t,1}(\boldsymbol{\theta}_1), \ldots, \boldsymbol{\lambda}_{t,p+r}(\boldsymbol{\theta}_{p+r}))^\top$, and the specific forms of $\mathbf{A}_k$'s, $\boldsymbol{\nu}$, and $\boldsymbol{\epsilon}_d$ are defined in (A.2) in the Appendix.

Proposition 1 shows that the daily integrated eigenvalues $\boldsymbol{\xi}_d$ follow the VAR model. We note that $\boldsymbol{\nu}$ is the the vector of the intercepts, $(a_1, \ldots, a_{p+r})^\top$ of instantaneous eigenvalues in (2.3) and the expected value, $(\mathrm{E}[z_{1,t}^2], \ldots, \mathrm{E}[z_{p+r,t}^2])^\top$, of the random fluctuations, and the coefficient, $\boldsymbol{\zeta}_1$, of the first AR order in (2.3) coming from handling the iterative relationship of the integrated eigenvalue. The coefficients $\mathbf{A}_k$'s are determined by the coefficients, $\boldsymbol{\zeta}_k$'s, of the instantaneous eigenvalue in (2.3). It is worth noting that we cannot specify all interpolation parameters in the instantaneous process because some quantities, such as $(a_1, \ldots, a_{p+r})^\top$ and the expected value of the random fluctuations, are not uniquely defined. Unlike the factor GARCH-Itô model (Kim and Fan, 2019), the FIVAR-Itô model considers not only the factor component but also the idiosyncratic component. In the empirical study, we find that the idiosyncratic eigenvalues have a time series structure, and incorporating the idiosyncratic dynamics helps capture the volatility dynamics. Details can be found in Section 5. We note that the proposed model is not the unique way to explain the observed auto-correlation structure in the empirical study (see Bollerslev et al. (2016); Cipollini et al. (2021); Hansen and Lunde (2014)). That is, the FIVAR-Itô model is one of the possible solutions, and we find its empirical benefits. However, incorporating the idiosyncratic component causes high-dimensionality. Furthermore, to account for the heavy-tailedness, we allow that the martingale noise $\boldsymbol{\epsilon}_d$ has heavy tails. That is, when it comes to statistical inferences for the proposed FIVAR-Itô model, we face two problems: the heavy-tailedness and over-parameterization.

In the following section, we propose an estimation procedure which can handle the heavy-tailedness and high dimensionality.

Remark 1. *In this paper, we propose a prediction procedure for integrated volatility matrices based on the VAR model structure of daily integrated eigenvalues. On the other hand, the intraday dynamic structure of the eigenvalues is not used for the prediction procedure. Specifically, since we use only the VAR model structure for the interday prediction procedure, the intraday dynamic structure does not affect the prediction method as long as the integrated eigenvalues follow the proposed VAR model. That is, the intraday dynamics in Definition 1 are proposed to show the existence of the continuous eigenvalue process that satisfies the VAR model structure of the integrated eigenvalue process. This enables us to accommodate both a continuous-time Itô process and a discrete-time VAR model and to conduct a simulation study based on high-frequency simulated data. It is interesting and important to develop a unified model that can explain interday and intraday dynamics simultaneously. We leave this for a future study.*

Remark 2. *In this paper, we assume that the rank r is constant over time. However, it may be more realistic to allow the rank r to vary over time. To handle the time-varying rank r, we can consider a state heterogeneous structure of the volatility process as in Chun and Kim (2022). For example, we can assume that the number of common factors is the same under the same state. Then, we need to extend the one-dimensional case in Chun and Kim (2022) to the high-dimensional case. However, the extension to the high-dimensional case is not straightforward. Thus, we leave this for a future study.*

3 Estimation procedure for the heavy-tailed VAR model

In this section, we propose a robust parameter estimation method for the high-dimensional VAR model in (2.4), where $\boldsymbol{\epsilon}_d = (\epsilon_{d,1}, \ldots, \epsilon_{d,p+r})^\top$ is i.i.d. innovation at time d with $\mathrm{E}(\boldsymbol{\epsilon}_d) = \mathbf{0}_{p+r}$, which is independent of $\boldsymbol{\xi}_{d-l}$ for all $l \in \mathbb{N}$. Our idea is basically to robustly fit this model for each component. Let $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\nu} & \mathbf{A}_1 & \cdots & \mathbf{A}_h \end{pmatrix}$ and we denote $\boldsymbol{\beta}_i$ by the ith row of $\boldsymbol{\beta}$. To overcome the curse of dimensionality, the sparsity of $\boldsymbol{\beta}_i$ is assumed: the number of nonzero elements in each $\boldsymbol{\beta}_i$

is bounded by a small number $s_\beta \geq 1$. In contrast, for the factor-related parameter, the factor model usually assumes that the idiosyncratic variables do not affect the factor variable. To reflect this prior, $\mathbf{A}_{k,i,j} = 0$ for $k = 1, \ldots, h$, $i = 1, \ldots, r$, $j = r+1, \ldots, p+r$ is assumed. That is, the factor-related coefficients $\boldsymbol{\beta}_i$'s, $i = 1, \ldots, r$, have the specific sparse structure. We denote the true model parameter by $\boldsymbol{\beta}_0$ and its ith row by $\boldsymbol{\beta}_{i0}$. It is worth mentioning that the sparsity implies the Granger non-causality between the related variables. In practice, we do not know the number, r, of latent factors and AR lag h. In this section, it is assumed that r and h are given, and we will discuss how to choose them in Section 5.1.

To accommodate the sparsity structure, we often employ the penalized regression model, such as LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001). When analyzing data with the LASSO procedure, we need some sub-Gaussian tail conditions. However, as shown in Figure 2, the volatilities often exhibit heavy tails in financial applications. To tackle this heavy-tailedness, we often employ a robustification method (Catoni, 2012; Fan et al., 2017; Minsker, 2018; Sun et al., 2020). In this paper, we employ the Huber loss l_τ (Huber, 1964)

$$l_\tau(x) = x^2/2 I(|x| \leq \tau) + (\tau |x| - \tau^2/2) I(|x| > \tau),$$

where $\tau > 0$ is the robustification parameter, and the truncation (Winsorization) method

$$\psi_\varpi(x) = x I(|x| \leq \varpi) + \text{sign}(x) \varpi I(|x| > \varpi),$$

where $\varpi > 0$ is a truncation parameter. We denote $\psi_\varpi(\mathbf{x}) = (\psi_\varpi(x_1), \ldots, \psi_\varpi(x_{p_1}))^\top$ for any vector $\mathbf{x} = (x_1, \ldots, x_{p_1})^\top \in \mathbb{R}^{p_1}$.

By combining the truncation and ℓ_1-regularization methods, we can simultaneously deal with robustness and the curse of dimensionality. Specifically, we estimate the true sparse coefficient $\boldsymbol{\beta}_{i0}$ as follows:

$$\widehat{\boldsymbol{\beta}}_i = \arg \min_{\boldsymbol{\beta}_i \in \mathbb{R}^{h(p+r)+1}} \mathcal{L}_{\tau,\varpi}^{I,i}(\boldsymbol{\beta}_i) + \eta_I \|\boldsymbol{\beta}_i\|_1 \quad \text{for } i = r+1, \ldots, p+r, \tag{3.1}$$

where $\eta_I > 0$ is the regularization parameter, the empirical loss function is

$$\mathcal{L}_{\tau,\varpi}^{I,i}(\boldsymbol{\beta}_i) = (n-h)^{-1} \sum_{d=h+1}^{n} l_{\tau_I}\left(\widehat{\xi}_{d,i} - \langle \psi_{\varpi_I}(\widehat{\boldsymbol{\xi}}_{d-1}^I), \boldsymbol{\beta}_i \rangle\right), \tag{3.2}$$

n is the number of days in the sample, $\widehat{\boldsymbol{\xi}}_d^I = \left(1, \widehat{\boldsymbol{\xi}}_d^\top, \ldots, \widehat{\boldsymbol{\xi}}_{d-h+1}^\top\right)^\top$, and $\widehat{\boldsymbol{\xi}}_d = \left(\widehat{\xi}_{d,1}, \ldots, \widehat{\xi}_{d,p+r}\right)^\top$ is a non-parametric estimator for $\boldsymbol{\xi}_d$. Note that in (3.2), the Huber loss l_{τ_I} is used to handle the heavy-tailedness of $\boldsymbol{\epsilon}_d$ and the truncation function ψ_ϖ is used to guard against the tail of $\boldsymbol{\xi}_d$. In contrast, since the sparsity structure of the coefficients for the factor part is known, it is a low-dimensional problem. We do not need the ℓ_1 penalty term. However, we still need the truncation parts to handle the heavy-tailedness as follows:

$$\widehat{\boldsymbol{\beta}}_i = \arg \min_{\boldsymbol{\beta}_i \in \mathbb{R}^{hr+1}} \mathcal{L}_{\tau,\varpi}^{F,i}(\boldsymbol{\beta}_i) \quad \text{for } i = 1, \ldots, r, \tag{3.3}$$

where

$$\mathcal{L}_{\tau,\varpi}^{F,i}(\boldsymbol{\beta}_i) = (n-h)^{-1} \sum_{d=h+1}^{n} l_{\tau_F}\left(\widehat{\xi}_{d,i} - \langle \psi_{\varpi_F}(\widehat{\boldsymbol{\xi}}_{d-1}^F), \boldsymbol{\beta}_i \rangle\right), \tag{3.4}$$

$\widehat{\boldsymbol{\xi}}_d^F$ is an $(hr+1)$ by 1 vector obtained by stacking 1 and the first r elements of each $\widehat{\boldsymbol{\xi}}_{d-k}$, $k = 0, \ldots, h-1$. We note that, in financial practice, we cannot observe the true price or volatility process, so we employ the non-parametric estimator $\widehat{\boldsymbol{\xi}}_d$ of $\boldsymbol{\xi}_d$. We discuss the non-parametric estimators in Section 4.

We investigate the theoretical properties of $\widehat{\beta}$ under the following assumptions.

Assumption 1.

(a) The process $(\widetilde{\boldsymbol{\xi}}_d)_{d=1,2,\ldots}$ is strictly stationary and the spectral radius of $\widetilde{\mathbf{A}}$, $\rho(\widetilde{\mathbf{A}})$, is less than 1, where $\widetilde{\boldsymbol{\xi}}_d$ and $\widetilde{\mathbf{A}}$ are the vectorization of $\boldsymbol{\xi}_d$ and its corresponding coefficient matrix defined in (A.3) in the Appendix, respectively.

(b) The number of nonzero elements in each $\boldsymbol{\beta}_{i0}$ is bounded by a number s_β.

(c) $\epsilon_{d,i}$ and $\xi_{d,i}$ satisfy $\max_{1 \leq i \leq p+r} \mathrm{E}(|\epsilon_{d,i}|^{c_\epsilon}) < \infty$ and $\max_{1 \leq i \leq p+r} \mathrm{E}(|\xi_{d,i}|^{c_\epsilon}) < \infty$ for some constant $c_\epsilon > 4$.

12

(d) The process $(\widetilde{\boldsymbol{\xi}}_d)_{d=1,2,\dots}$ is α-mixing and the α-mixing coefficients satisfy $\alpha(k) = O\left(\varphi^k\right)$ for some $\varphi \in (0,1)$.

(e) The non-parametric estimator $\widehat{\boldsymbol{\xi}}_d$ satisfies

$$\max_{1 \leq d \leq n} \max_{1 \leq i \leq r} \left|\widehat{\xi}_{d,i} - \xi_{d,i}\right| \leq b^F_{m,n,p} \quad and \quad \max_{1 \leq d \leq n} \max_{r+1 \leq i \leq p+r} \left|\widehat{\xi}_{d,i} - \xi_{d,i}\right| \leq b^I_{m,n,p} \ a.s.,$$

where m represents the number of observations for estimating $\xi_{d,i}$, and $b^F_{m,n,p}$ and $b^I_{m,n,p}$ converge to zero as m, n, and p increase.

(f) There exists a constant $\kappa > 0$ such that the following inequality holds for some $D_F \geq (hr+1)\eta_F/\kappa$ and $1 \leq i \leq r$, where the bound of η_F is given in Theorem 1:

$$\inf\{\mathbf{w}^\top \nabla^2 \mathcal{L}^{F,i}_{\tau,\varpi}(\boldsymbol{\beta}_i)\mathbf{w} : \ \|\mathbf{w}\|_2 = 1, \ \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i0}\|_1 \leq D_F\} \geq \kappa.$$

(g) Define the ℓ_1-cone $\mathcal{W}_i = \{\mathbf{w} \in \mathbb{R}^{hp+1} : \ \|\mathbf{w}_{S^c_i}\|_1 \leq 3\|\mathbf{w}_{S_i}\|_1\}$, where $\mathbf{w}_{S^c_i}$ is the subvector obtained by stacking $\{\mathbf{w}_j : \ j \in S^c_i\}$, $\mathbf{w}_{S_i}$ is the subvector obtained by stacking $\{\mathbf{w}_j : \ j \in S_i\}$, and $S_i = \{j : j\text{-th element of } \boldsymbol{\beta}_{i0} \neq 0\}$. Then, there exists a constant $\kappa > 0$ such that the following inequality holds for some $D_I \geq 48s_\beta\eta_I/\kappa$ and $1 \leq i \leq p$, where the bound of η_I is given in Theorem 1:

$$\inf\{\mathbf{w}^\top \nabla^2 \mathcal{L}^{I,i}_{\tau,\varpi}(\boldsymbol{\beta}_i)\mathbf{w} : \ \mathbf{w} \in \mathcal{W}_i, \ \|\mathbf{w}\|_2 = 1, \ \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i0}\|_1 \leq D_I\} \geq \kappa.$$

Remark 3. *Assumption 1(a) is the strictly stationary and stable conditions for the VAR(1) representation of the model (2.4). Assumption 1(c) allows the heavy-tailedness in the VAR model. Since we consider the high-dimensional VAR model, we need the moment condition for $\xi_{d,i}$, such as $\max_{1 \leq i \leq p+r} \mathrm{E}(|\xi_{d,i}|^{c_\epsilon}) < \infty$. However, under Assumption 1(a)–(b), the condition $\max_{1 \leq i \leq p+r} \mathrm{E}(|\epsilon_{d,i}|^{c_\epsilon}) < \infty$ implies the condition $\max_{1 \leq i \leq p+r} \mathrm{E}(|\xi_{d,i}|^{c_\epsilon}) < \infty$ when s_β is bounded by some positive constant (see Lemma 1 in the Appendix). We note that we do not impose the bounded s_β throughout the paper; thus, we need the moment condition for $\xi_{d,i}$. Assumption 1(d) is required to handle the*

dependency in the VAR model. Under Assumption 1(a), Assumption 1(d) holds if the process
$(\widetilde{\boldsymbol{\xi}}_d)_{d=1,2,\dots}$ *is geometric ergodic (see Proposition 2 in Liebscher (2005) and Fact 5 in the online*
Appendix of Wong et al. (2020)). We note that the geometric ergodicity can be obtained under
the mild condition on $\boldsymbol{\epsilon}_d$ (see Example 3 in Wong et al. (2020)). Assumption 1(e) represents the
concentration property of the non-parametric estimator $\widehat{\boldsymbol{\xi}}_d$. In Section 4, we propose a method for
constructing $\widehat{\boldsymbol{\xi}}_d$ and show its associated inequality holds with high probability. Assumptions 1(f)–(g)
are the eigenvalue conditions for the Hessian matrices $\nabla^2 \mathcal{L}_{\tau,\varpi}^{F,i}(\boldsymbol{\beta}_i)$ and $\nabla^2 \mathcal{L}_{\tau,\varpi}^{I,i}(\boldsymbol{\beta}_i)$, respectively.
This is called the localized restricted eigenvalue (LRE) condition (Fan et al., 2018; Sun et al.,
2020), which implies strictly positive restricted eigenvalues over a local neighborhood.

The following theorem provides the convergence rate of $\widehat{\boldsymbol{\beta}}_i$ defined in (3.1) and (3.3).

Theorem 1. *Under the model (2.4), Assumption 1, $n \geq 3$, $\delta \geq 1$, $\sqrt{n\delta} + (\tau_F + \varpi_F)(\log n)^2 \delta = O(n)$, and $\eta_F \geq C[b_{m,n,p}^F + \tau_F^{-2} + \varpi_F^{-2} + \frac{\tau_F \varpi_F (\log n)^2 \delta + \sqrt{n\delta}}{n}]$, we have, for $i = 1, \dots, r$, with probability*
at least $1 - 4hre^{-\delta}$,

$$\left\| \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0} \right\|_2 \leq \frac{(hr+1)^{1/2} \eta_F}{\kappa}. \tag{3.5}$$

Furthermore, we assume that $s_\beta \sqrt{n\delta} + (\tau_I + \varpi_I)(\log n)^2 \delta = O(n)$ and $\eta_I \geq C[s_\beta (b_{m,n,p}^F + b_{m,n,p}^I) + s_\beta^3 \tau_I^{-2} + s_\beta \varpi_I^{-2} + \frac{\tau_I \varpi_I (\log n)^2 \delta + s_\beta \sqrt{n\delta}}{n}]$. Then, we have, for $i = r+1, \dots, p+r$, with probability at least
$1 - 4h(p+r)e^{-\delta}$,

$$\left\| \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0} \right\|_2 \leq \frac{12 s_\beta^{1/2} \eta_I}{\kappa}. \tag{3.6}$$

Remark 4. *Theorem 1 shows the convergence rates for the general setting of the low-dimensional*
and high-dimensional VAR models, where the covariates are not observable and observations are
heavy-tailed. Specifically, $b_{m,n,p}^F$ and $b_{m,n,p}^I$ in η_F and η_I are the costs to estimate the true co-
variates. When $\boldsymbol{\xi}_d$ is directly observable, $b_{m,n,p}^F$ and $b_{m,n,p}^I$ become zero. Take $\delta = 2\log p$, $\tau_F = \varpi_F = C(n/\log p)^{1/4}$, $\eta_F = C(\log n)^2 \sqrt{\log p/n}$, $\tau_I = Cs_\beta (n/\log p)^{1/4}$, $\varpi_I = C(n/\log p)^{1/4}$, and
$\eta_I = Cs_\beta (\log n)^2 \sqrt{\log p/n}$. Then, $\widehat{\boldsymbol{\beta}}_i$ for the factor and idiosyncratic parts have a near-optimal
convergence rate of $(\log n)^2 \sqrt{\log p/n}$ and $s_\beta^{3/2} (\log n)^2 \sqrt{\log p/n}$, respectively (Sun et al., 2020).
The additional $(\log n)^2$ term comes from handling the dependency in the process $(\boldsymbol{\xi}_d)_{d=1,2,\dots}$. When
comparing to the optimal rate for the high-dimensional case, established in Sun et al. (2020), we

have the additional $s_\beta (\log n)^2$. Usually, the sparsity level is small; thus, the proposed method does not lose significant efficiency, even for the dependent and heavy-tailed case.

Remark 5. *In addition to obtaining concentration inequalities for the model parameter estimation method, obtaining confidence intervals for the model parameters is also important. To do this, we need to adjust the bias of the proposed estimator and to obtain a debiased estimator. This bias is coming from Huber loss, truncation, high-dimensional observation error, and ℓ_1 regularization. It is a demanding task to simultaneously handle them. We leave this issue for a future study.*

4 Large volatility matrix prediction

4.1 A model set-up

In this section, using the estimation procedure in Section 3, we discuss how to predict the large volatility matrix, based on the FIVAR-Itô model. Given the observations of n days, the parameter of interest is the conditional expected volatility matrix $\mathrm{E}(\boldsymbol{\Gamma}_{n+1}|\mathcal{F}_n)$. Recall that the integrated volatility matrix $\boldsymbol{\Gamma}_d$ has the following low-rank plus sparse structure:

$$\boldsymbol{\Gamma}_d = \boldsymbol{\Psi}_d + \boldsymbol{\Sigma}_d = p \sum_{i=1}^{r} \xi_{d,i} \mathbf{q}_i^F \left(\mathbf{q}_i^F \right)^\top + \sum_{i=1}^{p} \xi_{d,i+r} \mathbf{q}_i^I \left(\mathbf{q}_i^I \right)^\top \quad \text{a.s.,}$$

where $p\xi_{d,i}$'s are the i-th largest eigenvalues of $\boldsymbol{\Psi}_d$ for $i = 1, \ldots, r$ and $\xi_{d,i+r}$'s are the i-th largest eigenvalues of $\boldsymbol{\Sigma}_d$ for $i = 1, \ldots, p$. It is assumed that the rank, r, of $\boldsymbol{\Psi}_d$ is bounded and the idiosyncratic volatility matrix $\boldsymbol{\Sigma}_d = (\Sigma_{d,ij})_{i,j=1,\ldots,p}$ satisfies the following sparse condition:

$$\max_{1 \leq d \leq n} \max_{1 \leq i \leq p} \sum_{j=1}^{p} |\Sigma_{d,ij}|^\Upsilon (\Sigma_{d,ii} \Sigma_{d,jj})^{(1-\Upsilon)/2} \leq M_I s_I \quad \text{a.s.,} \tag{4.1}$$

where M_I is a bounded positive random variable, $\Upsilon \in [0,1)$, and s_I is a deterministic function of p, which grows slowly in p. This low-rank plus sparse structure is widely employed when analyzing the large matrices (Aït-Sahalia and Xiu, 2017; Bai and Ng, 2002; Fan and Kim, 2018; Fan et al., 2013; Kim et al., 2018; Stock and Watson, 2002; Shin et al., 2023). We note that when we directly use the

total volatility without decomposition, we cannot explain the sparse structure of the eigenvectors of the idiosyncratic volatility matrices. This may introduce a lot of parameters and lead to higher complexity.

Unfortunately, the true log-price $\mathbf{X}(t)$ cannot be directly observed since the high-frequency data are contaminated by microstructure noise. To account for this, it is assumed that the observed log-price $Y_i(t_k)$ has the following additive noise structure:

$$Y_i(t_{d,k}) = X_i(t_{d,k}) + e_i(t_{d,k}) \quad \text{for } i = 1, \ldots, p, d = 1, \ldots, n, k = 0, \ldots, m, \qquad (4.2)$$

where $d - 1 = t_{d,0} < \cdots < t_{d,m} = d$, and the microstructure noise $e_i(t_{d,k})$ is a stationary random variable with mean zero. The empirical studies have shown that the microstructure noise is serial dependent and endogenous (Aït-Sahalia et al., 2011; Hansen and Lunde, 2006b; Jacod et al., 2017b; Li and Linton, 2022; Ubukata and Oya, 2009). Fortunately, as long as non-parametric integrated volatility matrix estimators satisfy (4.3) presented below, the dependent structure of the microstructure noise does not affect the main results of this paper. There are several estimation procedures that are robust to dependent structures of the microstructure noise (Barndorff-Nielsen et al., 2011b; Jacod et al., 2017b; Kim et al., 2016; Li and Linton, 2020). Similarly, the assumptions on the jumps do not affect the main results of this paper as long as (4.3) holds. There are also several estimation methods which can handle jumps when estimating integrated volatilities (Aït-Sahalia and Xiu, 2016; Shin et al., 2023). Thus, we only require condition (4.3). On the other hand, for simplicity, the observation time points are assumed to be synchronized and equally spaced: $t_{d,k} - t_{d,k-1} = m^{-1}$ for $d = 1, \ldots, n$ and $k = 1, \ldots, m$.

Remark 6. *In this paper, we mainly focus on the parametric structure of the volatility process, so it is assumed that the observation time points are synchronized and equally spaced for simplicity. The conditions for the observation time points can be relaxed to the non-synchronized and unequally spaced conditions by using generalized sampling time (Aït-Sahalia et al., 2010), refresh time (Barndorff-Nielsen et al., 2011a), and previous tick (Andersen et al., 2003; Barndorff-Nielsen et al., 2011a; Zhang, 2011) schemes. See also Bibinger et al. (2014); Fan and Kim (2019); Park et al. (2016).*

4.2 Large volatility matrix prediction

To predict the large volatility matrix, we first employ a non-parametric integrated volatility matrix estimator $\widehat{\boldsymbol{\Gamma}}_d$, which is robust to jumps and dependent structures of the microstructure noise (Aït-Sahalia and Xiu, 2016; Barndorff-Nielsen et al., 2011b; Bibinger and Winkelmann, 2015; Jacod et al., 2009; Kim et al., 2016; Koike, 2016; Li and Linton, 2020; Shin et al., 2023). Based on the non-parametric estimator $\widehat{\boldsymbol{\Gamma}}_d$, we estimate the eigenvectors and eigenvalues of factor and idiosyncratic volatility matrices as follows. For estimating the 'daily' integrated eigenvalues $\boldsymbol{\xi}_d$ on the factor volatility matrix $\boldsymbol{\Psi}_d$, based on the assumption of time-invariance of eigenvectors, we calculate r eigenvectors $\widehat{\mathbf{q}}_1^F, \ldots, \widehat{\mathbf{q}}_r^F$ of the average of the recent ℓ days' non-parametric realized volatility matrix estimators, $\frac{1}{\ell} \sum_{d=n-\ell+1}^{n} \widehat{\boldsymbol{\Gamma}}_d$, where ℓ is the window length for the eigenvector estimation, and obtain the estimators of time-dependent eigenvalues $\widehat{\xi}_{d,i} = \left(\widehat{\mathbf{q}}_i^F\right)^{\top} \widehat{\boldsymbol{\Gamma}}_d \widehat{\mathbf{q}}_i^F / p$ for $d = 1, \ldots, n$ and $i = 1, \ldots, r$. This provides a part of inputs for (3.2) and (3.4).

To provide the rest of the inputs, namely $\widehat{\xi}_{d,i}$ for $i > r$ for the idiosyncratic volatility matrix $\boldsymbol{\Sigma}_d$, we apply the principal orthogonal complement thresholding (POET) method (Fan et al., 2013) as follows. First, we decompose the input volatility matrix

$$\widehat{\boldsymbol{\Gamma}}_d = \sum_{k=1}^{p} \bar{\xi}_{d,k} \bar{\mathbf{q}}_{d,k} \bar{\mathbf{q}}_{d,k}^{\top},$$

where $\bar{\xi}_{d,k}$ is the k-th largest eigenvalue of $\widehat{\boldsymbol{\Gamma}}_d$ and $\bar{\mathbf{q}}_{d,k}$ is its corresponding eigenvector. We then obtain the input idiosyncratic volatility matrix estimator $\bar{\boldsymbol{\Sigma}}_d = (\bar{\Sigma}_{d,ij})_{1 \leq i,j \leq p} = \widehat{\boldsymbol{\Gamma}}_d - \sum_{k=1}^{r} \bar{\xi}_{d,k} \bar{\mathbf{q}}_{d,k} \bar{\mathbf{q}}_{d,k}^{\top}$ and apply the adaptive thresholding method to $\bar{\boldsymbol{\Sigma}}_d$ by computing

$$\widehat{\Sigma}_{d,ij} = \begin{cases} \bar{\Sigma}_{d,ij} \vee 0 & \text{if } i = j \\ g_{ij}(\bar{\Sigma}_{d,ij})\mathbf{1}(|\bar{\Sigma}_{d,ij}| \geq v_{ij}) & \text{if } i \neq j \end{cases} \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}}_d = (\widehat{\Sigma}_{d,ij})_{1 \leq i,j \leq p},$$

where the thresholding function $g_{ij}(\cdot)$ satisfies $|g_{ij}(x) - x| \leq v_{ij}$, and the adaptive thresholding level $v_{ij} = v_m \sqrt{(\bar{\Sigma}_{d,ii} \vee 0)(\bar{\Sigma}_{d,jj} \vee 0)}$. For example, we often utilize the soft thresholding function $g_{ij}(x) = x - \text{sign}(x)v_{ij}$ and the hard thresholding function $g_{ij}(x) = x$. The thresholding parameter v_m will be specified in Proposition 2. With the idiosyncratic volatility matrix estimator $\widehat{\boldsymbol{\Sigma}}_d$, we

calculate p eigenvectors, $\widehat{\mathbf{q}}_1^I, \ldots, \widehat{\mathbf{q}}_p^I$, of $\frac{1}{\ell} \sum_{d=n-\ell+1}^n \widehat{\boldsymbol{\Sigma}}_d$ and obtain $\widehat{\xi}_{d,i+r} = \left(\widehat{\mathbf{q}}_i^I\right)^\top \widehat{\boldsymbol{\Sigma}}_d \widehat{\mathbf{q}}_i^I$ for $d = 1, \ldots, n$ and $i = 1, \ldots, p$. Again, we here use the time-invariant assumption of the eigenvectors.

With these inputs, we can estimate the true model parameter $\boldsymbol{\beta}_0$ using the VAR model parameter estimation procedure in Section 3 and calculate the predicted eigenvalue estimator by $\widehat{\boldsymbol{\xi}}_{n+1} = \left(\widehat{\xi}_{n+1,1}, \ldots, \widehat{\xi}_{n+1,p+r}\right)^\top = \widehat{\boldsymbol{\nu}} + \sum_{k=1}^h \widehat{\mathbf{A}}_k \widehat{\boldsymbol{\xi}}_{n+1-k}$ using (2.4). Finally, we estimate the conditional expected volatility matrix by

$$\widetilde{\boldsymbol{\Gamma}}_{n+1} = \widehat{\boldsymbol{\Psi}}_{n+1} + \widehat{\boldsymbol{\Sigma}}_{n+1} = p \sum_{i=1}^r \widehat{\xi}_{n+1,i} \widehat{\mathbf{q}}_i^F \left(\widehat{\mathbf{q}}_i^F\right)^\top + \sum_{i=1}^p \widehat{\xi}_{n+1,i+r} \widehat{\mathbf{q}}_i^I \left(\widehat{\mathbf{q}}_i^I\right)^\top.$$

We describe the estimation procedure in Algorithm 1 in the Appendix.

Remark 7. *To estimate the eigenvectors, the constant eigenvector over time is assumed, and the window length, ℓ, for the eigenvector estimation can be from 1 to n.* *In the empirical study, we investigate the effect of the eigenvector estimation methods, and we find that the volatility matrix estimator with the previous 22-day observations (one month) shows the best performance. This shows that the averaging step helps mitigate volatile fluctuations in the volatility process, and by using the recent 22-day instead of the longer period, such as the whole period, we can explain the effect of the eigenvector dynamics. On the other hand, in high-frequency finance literature, the intraday time-varying patterns are often observed (Andersen et al., 2019, 2021; Kong et al., 2021). We also conducted hypothesis tests for the constant eigenvector, and the constant eigenvector hypothesis is often rejected (see Section A.2). Thus, it is more natural to assume intraday time-varying eigenvectors. However, under this condition, we need to calculate a lot of local eigenvalues and eigenvectors and accumulate the local estimators. Furthermore, the intraday dynamic structure of the volatility process becomes more complex since we need to consider two different dynamic structures. This complexity may cause large estimation errors and the possibility of over-parameterization. Thus, it is a demanding task to develop a parametric model that can explain the intraday and interday dynamics simultaneously and obtain robust prediction performance. We leave this for a future study.*

We investigate the theoretical properties of the POET estimator under the following assump-

tions. These conditions are often used when analyzing the asymptotic behaviors of the POET estimator (Fan and Kim, 2018; Shin et al., 2023).

Assumption 2.

(a) For some fixed constant C_1, we have $\frac{p}{r} \max_{1 \leq i \leq p} \sum_{j=1}^{r} q_{ij}^2 \leq C_1$ a.s., where $\mathbf{q}_j = (q_{1j}, \ldots, q_{pj})^\top$ is the jth eigenvector of $\mathbf{\Psi}_d$.

(b) For $d = 1, \ldots, n$, let $D_{d,\xi} = \min\{\xi_{d,i} - \xi_{d,i+1}, i = 1, \ldots, r-1\}$, $D_{d,\xi}$ and $\xi_{d,r} \geq C_2$ a.s., and $\xi_{d,1} \leq C_3$ a.s. for some generic positive constants C_2 and C_3.

(c) For $d = 1, \ldots, n$, $\xi_{d,r+1}$ is bounded by some positive constant and $\xi_{d,p+r}$ stays away from zero almost surely.

(d) $s_I / \sqrt{p} + \sqrt{\log p / m^{1/2}} = o(1)$.

The following proposition derives the concentration properties of the POET estimator.

Proposition 2. *Under the FIVAR(h)-Itô model, suppose that the concentration inequality,*

$$\Pr\left\{ \max_{1 \leq d \leq n} \max_{1 \leq i,j \leq p} \left| \widehat{\Gamma}_{d,ij} - \Gamma_{d,ij} \right| \geq C \sqrt{\frac{\log(pn \vee m)}{m^{1/2}}} \right\} \leq p^{-1}, \tag{4.3}$$

Assumption 2, and the sparsity condition (4.1) are met. Take $v_m = C_\varpi H_m$ for some large fixed constant C_ϖ, where $H_m = s_I/p + \sqrt{\log(pn \vee m)/m^{1/2}}$. Then, we have, for a sufficiently large m, with probability at least $1 - p^{-1}$,

$$\max_{1 \leq d \leq n} \max_{1 \leq i \leq r} \left| \widehat{\xi}_{d,i} - \xi_{d,i} \right| \leq C H_m, \tag{4.4}$$

$$\max_{1 \leq d \leq n} \max_{r+1 \leq i \leq p+r} \left| \widehat{\xi}_{d,i} - \xi_{d,i} \right| \leq C s_I H_m^{1-\Upsilon}, \tag{4.5}$$

$$\max_{1 \leq d \leq n} \|\widehat{\mathbf{\Sigma}}_d - \mathbf{\Sigma}_d\|_2 \leq C s_I H_m^{1-\Upsilon}, \quad and \tag{4.6}$$

$$\max_{1 \leq d \leq n} \|\widehat{\mathbf{\Sigma}}_d - \mathbf{\Sigma}_d\|_{\max} \leq C H_m. \tag{4.7}$$

Remark 8. *Under the locally boundedness condition of the instantaneous volatility process with the heavy-tailed observations, we can obtain the same asymptotic results (Fan and Kim, 2018; Shin*

et al., 2023). For example, continuous adapted processes are locally bounded, and more generally, left-continuous adapted processes are almost surely locally bounded on every finite time interval. The proposed FIVAR-Itô model is continuous; thus, the locally bounded condition is satisfied. Thus, the concentration tail condition (4.3) is not restrictive.

The concentration inequalities (4.4)–(4.5) show that Assumption 1(e) is satisfied with high probability. For example, we have $b^F_{m,n,p} = CH_m$ and $b^I_{m,n,p} = Cs_I H_m^{1-\Upsilon}$. Using Theorem 1, we can derive the following result.

Theorem 2. *Under the assumptions in Proposition 2 and Assumption 1 (except for Assumption 1(e)), let $n \geq 3$, $s_\beta \sqrt{n \log p} + s_\beta n^{1/4} (\log n)^2 (\log p)^{3/4} = O(n)$, $\tau_F = \varpi_F = C (n/\log p)^{1/4}$, $\tau_I = Cs_\beta (n/\log p)^{1/4}$, and $\varpi_I = C (n/\log p)^{1/4}$. Suppose that Assumption 1(f)–(g) hold with $\eta_F = C \left\{ H_m + (\log n)^2 \sqrt{\log p/n} \right\}$ and $\eta_I = C \left\{ s_\beta s_I H_m^{1-\Upsilon} + s_\beta (\log n)^2 \sqrt{\log p/n} \right\}$. Then, we have, for a sufficiently large m, with probability at least $1 - 2p^{-1}$,*

$$\max_{1 \leq i \leq r} \left\| \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0} \right\|_2 \leq C \left\{ H_m + (\log n)^2 \sqrt{\log p/n} \right\}, \quad and \tag{4.8}$$

$$\max_{r+1 \leq i \leq p+r} \left\| \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0} \right\|_2 \leq C \left\{ s_\beta^{3/2} s_I H_m^{1-\Upsilon} + s_\beta^{3/2} (\log n)^2 \sqrt{\log p/n} \right\}. \tag{4.9}$$

Remark 9. *Theorem 2 shows that the low-dimensional factor VAR has the convergence rate $s_I/p + \sqrt{\log (pn \vee m) / m^{1/2}} + (\log n)^2 \sqrt{\log p/n}$. The s_I/p term is the cost to identify the latent factor, and the $m^{-1/4}$ term comes from estimating the integrated volatility matrix. Finally, the $n^{-1/2}$ term is the usual convergence rate of estimating model parameters in low-frequency time series. In contrast, the high-dimensional factor VAR has the convergence rate $s_\beta^{3/2} s_I H_m^{1-\Upsilon} + s_\beta^{3/2} (\log n)^2 \sqrt{\log p/n}$. The first term, $s_\beta^{3/2} s_I H_m^{1-\Upsilon}$, is the cost to estimate the latent idiosyncratic volatility matrix with the noisy high-frequency data. The second term, $s_\beta^{3/2} (\log n)^2 \sqrt{\log p/n}$, is the convergence rate of the sparse high-dimensional regression. We note that the $(\log n)^2$ term is the cost of handling the dependency in the eigenvalue process.*

With the results in Theorem 2, we investigate the theoretical properties of the future volatility matrix estimator $\widetilde{\boldsymbol{\Gamma}}_{n+1}$. To study the future idiosyncratic volatility matrix estimator $\widehat{\boldsymbol{\Sigma}}_{n+1}$, we need the additional condition for the eigen-gap as follows.

Assumption 3. *For some $\chi \in (0,1)$ and $i = r+1, \ldots, p+r-1$, we have $C_4\chi^i \leq \xi_{d,i} - \xi_{d,i+1} \leq C_5\chi^i$ for some positive constants C_4 and C_5.*

Remark 10. *To have the bounded eigenvalues for the idiosyncratic volatility matrices such as Assumption 2(c), we cannot have that all eigen-gaps are some positive constants. Specifically, several eigen-gaps can be constant, but most of them may need to converge to zero. To check the behavior of the eigen-gaps, we draw the plot of the eigen-gaps of the idiosyncratic volatility matrix using high-frequency trading data (see Figure 7 in the Appendix). We find that the eigen-gaps have an exponentially decaying pattern. Thus, to account for this, we impose Assumption 3. We note that, even if the finite number of $\xi_{d,i}$'s do not satisfy this condition, we can obtain the same results in Theorem 3.*

The following theorem establishes the convergence rates of the future volatility matrix estimator.

Theorem 3. *Under the assumptions in Theorem 2 and Assumption 3, we have with probability at least $1 - 2p^{-1}$,*

$$
\max_{1 \leq i \leq r} \left| \widehat{\xi}_{n+1,i} - \mathrm{E}\left(\xi_{n+1,i} | \mathcal{F}_n \right) \right| \;\leq\; C\left[H_m + (\log n)^2 \sqrt{\log p/n} \right], \tag{4.10}
$$

$$
\max_{r+1 \leq i \leq p+r} \left| \widehat{\xi}_{n+1,i} - \mathrm{E}\left(\xi_{n+1,i} | \mathcal{F}_n \right) \right| \;\leq\; C\left[s_\beta^2 s_I H_m^{1-\Upsilon} + s_\beta^2 (\log n)^2 \sqrt{\log p/n} \right], \tag{4.11}
$$

$$
\left\| \widetilde{\boldsymbol{\Gamma}}_{n+1} - \mathrm{E}\left(\boldsymbol{\Gamma}_{n+1} | \mathcal{F}_n \right) \right\|_{\boldsymbol{\Gamma}^*} \;\leq\; C\Big[p^{1/2} H_m^2 + p^{1/2} \log p \, (\log n)^4 / n
$$
$$
+ s_\beta^2 s_I H_m^{1-\Upsilon} + s_\beta^2 (\log n)^2 \sqrt{\log p/n} \Big], \tag{4.12}
$$

where the relative Frobenius norm $\|\mathbf{M}\|_{\boldsymbol{\Gamma}^*}^2 = p^{-1}\|\boldsymbol{\Gamma}^{*-1/2}\mathbf{M}\boldsymbol{\Gamma}^{*-1/2}\|_F^2$ and $\boldsymbol{\Gamma}^* = \mathrm{E}\left(\boldsymbol{\Gamma}_{n+1} | \mathcal{F}_n \right)$.

Remark 11. *The relative Frobenius norm is used in Theorem 3 since the top eigenvalues of $\boldsymbol{\Gamma}^*$ are diverging (see Fan et al. (2008)). Theorem 3 indicates that the proposed estimator $\widetilde{\boldsymbol{\Gamma}}_{n+1}$ is consistent as long as $p = o(m \wedge n^2)$ in terms of the relative Frobenius norm. Its convergence rate is similar to that of Kim and Fan (2019) except for the additional terms, $p^{1/2} \log p \, (\log n)^4 / n$ and $s_\beta^2 (\log n)^2 \sqrt{\log p/n}$, which come from handling the VAR model structure in the factor and idiosyncratic volatility matrices, respectively.*

4.3 Discussion on the tuning parameter selection

To implement the proposed robust estimation method, we need to choose the tuning parameters. In this section, we discuss how to select the tuning parameters in (3.1)–(3.4). For the factor part, let $\sigma_F = \sqrt{\sum_{i=1}^{r} \sum_{d=1}^{n} \widehat{\xi}_{d,i}^2 / (nr)}$. We choose

$$\varpi_F = c_{F,1} \sigma_F \left(\frac{n}{\log p} \right)^{1/4} \quad \text{and} \quad \tau_F = c_{F,2} \sigma_F \left(\frac{n}{\log p} \right)^{1/4}, \tag{4.13}$$

where $c_{F,1}$ and $c_{F,2}$ are tuning parameters. For the idiosyncratic part, we first standardize the variables, $\widehat{\xi}_{d,i}$, $i = 1, \ldots, p + r$, to have mean zero and variance 1. Then, we choose

$$\varpi_I = c_{I,1} \left(\frac{n}{\log p} \right)^{1/4}, \tau_I = c_{I,2} \left(\frac{n}{\log p} \right)^{1/4}, \quad \text{and} \quad \eta_I = c_\eta \left(\frac{\log p}{n} \right)^{1/2}, \tag{4.14}$$

where $c_{I,1}$, $c_{I,2}$, and c_η are tuning parameters. We select $c_\eta \in [0.1, 10]$ by minimizing the corresponding Bayesian information criterion (BIC). In the simulation and empirical studies, we choose $c_{F,1} = 4$, $c_{F,2} = 1/4$, $c_{I,1} = 4$, and $c_{I,2} = 4$. These choices are based on the empirical study (Section 5.2). Specifically, we choose $c_{F,1}$, $c_{F,2}$, $c_{I,1}$, and $c_{I,2}$ which minimize the corresponding mean squared prediction error (MSPE). Details can be found in Section 5.2.

5 Numerical study

5.1 A simulation study

In this section, we conducted simulations to validate the finite sample performance of the proposed estimation methods. We generated the data for n days with frequency $1/m^{all}$ on each day and let $t_{d,j} = d - 1 + j/m^{all}$ for $d = 1, \ldots, n$ and $j = 0, \ldots, m^{all}$. We considered the jump-diffusion process with the FIVAR(h)-Itô model in Definition 1 and generated heavy-tailed and sub-Gaussian processes. The specific simulation setup is described in Appendix A.1. The noise-contaminated high-frequency data were generated from model (4.2), where the noise $e_i(t_{i,k})$ was obtained from the independent Gaussian distribution with mean zero and standard deviation $0.01\sqrt{\int_0^1 \gamma_{ii}(t)dt}$.

This choice is inspired by Wang and Zou (2010) who found that the relative noise level is typically around 1% for the stock index for high-frequency trading data. We first generated the data for 500 days, and we varied n from 100 to 500. For each n, we obtained the data from the last n days among the 500 days.

To estimate the integrated volatility matrix $\boldsymbol{\Gamma}_d = (\Gamma_{d,ij})_{i,j=1,\ldots,p}$, we utilized the jump adjusted pre-averaging volatility matrix (PAVM) estimator (Aït-Sahalia and Xiu, 2016; Christensen et al., 2010; Jacod et al., 2009) defined in (A.1) in the Appendix. Then, we estimated the conditional expected volatility matrix $\mathrm{E}\left(\boldsymbol{\Gamma}_{n+1}|\mathcal{F}_n\right)$, based on the estimation procedure in Section 4.2. Specifically, we first projected $\widehat{\boldsymbol{\Gamma}}_d$ onto the positive semi-definite cone in the spectral norm to make it positive semi-definite. Since the eigenvectors are constant over time, we estimated them using the n period observations. To determine the rank r, we employed the procedure in Aït-Sahalia and Xiu (2017) as follows:

$$\widehat{r} = \arg\min_{1 \leq j \leq r_{\max}} \sum_{d=1}^{n} \left[p^{-1}\bar{\xi}_{d,j} + j \times c_1 \left\{ \sqrt{\log p/m^{1/2}} + p^{-1}\log p \right\}^{c_2} \right] - 1, \qquad (5.1)$$

where $\bar{\xi}_{d,j}$ is the j-th largest eigenvalue of PAVM, $r_{\max} = 30$, $c_1 = 0.02 \times \bar{\xi}_{d,30}$, and $c_2 = 0.5$. For the POET estimation procedure, we employed the soft thresholding scheme and selected the thresholding level that minimizes the corresponding Frobenius norm. When estimating $\boldsymbol{\beta}_{i0}$'s, we used the tuning parameter selection method discussed in Section 4.3. To select the lag h, we utilized the Bayesian information criterion (BIC). We calculated the future volatility matrix estimator with $\widehat{\boldsymbol{\beta}}_i$ and call it the Huber-LASSO (H-LASSO) estimator.

For comparison, we employ the ordinary least squares (OLS) and LASSO estimators as follows. The OLS estimator only considers the dynamics of the factor volatility matrix and obtains $\widehat{\boldsymbol{\beta}}_i$, $i = 1, \ldots, \widehat{r}$, using the OLS method. The OLS estimator predicts the future idiosyncratic volatility matrix by the average of the previous 22-day's idiosyncratic volatility matrices to smooth random fluctuations. On the other hand, the LASSO estimator considers the dynamics in both factor and idiosyncratic volatility matrices. The LASSO estimator uses the same estimation procedure as the H-LASSO estimator, except for the truncation method. That is, the OLS estimator can explain only the dynamics from the factor component, while the LASSO estimator can account for the

dynamics from both the factor and idiosyncratic components. However, they cannot account for the heavy-tailedness. We also investigated the previous day's PAVM estimator from the POET procedure as the non-parametric benchmark. We call it the POET-PAVM. We calculated the average estimation errors under the Frobenius norm, the max norm, the relative Frobenius norm (see Theorem 3 for the definition), and the spectral norm by 500 iterations. Note that we conducted one-day-ahead forecasts for each of 500 iterations.

We first checked the performance of the methods for model parameter estimation. The parameter of interest is the true parameter matrix $\boldsymbol{\beta}_0$. Table 1 reports the Frobenius, max, and spectral norm errors of the LASSO and H-LASSO estimators, with $n = 100, 200, 500$ and $m = 250, 500, 2000$. We note that for both heavy-tailed and sub-Gaussian processes, the number of factors r and lag h are estimated without errors for all n and m. The reason is that the data generation process has a large eigen-gap between the factor and idiosyncratic volatility matrices and a strong time series structure. From Table 1, we find that the estimation errors of the proposed H-LASSO estimator are usually decreasing as the number of low-frequency or high-frequency observations increases. The exception is the max norm error for $n = 500$ and $m = 2000$, while the overall error performances, such as Frobenius and spectral norm errors, always decrease as n or m increases. An explanation is that bigger outliers for heavy tails are more frequently observed as the high-frequency observation increases. For example, when m is small, the relative frequency of outliers may be low due to the smoothing effect from the subsampling. Furthermore, the max norm measure is highly affected by the outlier. When comparing two estimation methods, the H-LASSO estimator performs better than the LASSO estimator for both heavy-tailed and sub-Gaussian processes. One possible explanation for this is that, even if the process is generated by the sub-Gaussian variables, the log-prices process can still have some heavy tails. The truncation method can reduce the variance of the estimator, which is larger than that of the increase in estimation bias, even for the sub-Gaussian case. From this result, we find the benefit of handling the heavy-tailedness. These results support the theoretical findings in Section 3.

One of the main objectives of this paper is to predict future volatility. Therefore, we checked the performance of predicting future volatility. Figures 3 and 4 plot the log Frobenius, max, relative

Table 1: The Frobenius, max, and spectral norm errors of the LASSO and H-LASSO estimators with $n = 100, 200, 500$ and $m = 250, 500, 2000$.

Tail	n	m	Frobenius		Max		Spectral	
			LASSO	H-LASSO	LASSO	H-LASSO	LASSO	H-LASSO
Heavy	100	250	0.805	0.647	0.460	0.286	0.586	0.396
		500	0.752	0.584	0.462	0.283	0.569	0.365
		2000	0.727	0.559	0.461	0.280	0.554	0.346
	200	250	0.708	0.595	0.378	0.236	0.477	0.335
		500	0.657	0.530	0.381	0.230	0.459	0.294
		2000	0.640	0.510	0.379	0.222	0.447	0.278
	500	250	0.604	0.561	0.250	0.187	0.352	0.300
		500	0.543	0.494	0.248	0.181	0.312	0.235
		2000	0.534	0.475	0.255	0.179	0.309	0.222
Sub-Gaussian	100	250	0.835	0.711	0.482	0.350	0.620	0.467
		500	0.782	0.646	0.491	0.344	0.608	0.443
		2000	0.758	0.612	0.482	0.325	0.589	0.415
	200	250	0.709	0.637	0.361	0.274	0.472	0.374
		500	0.647	0.569	0.353	0.265	0.448	0.343
		2000	0.621	0.544	0.347	0.256	0.427	0.326
	500	250	0.613	0.583	0.249	0.204	0.349	0.307
		500	0.546	0.513	0.243	0.194	0.314	0.262
		2000	0.528	0.492	0.249	0.197	0.307	0.250

Frobenius, and spectral norm errors of the future volatility matrix estimators with $n = 100, 200, 500$ and $m = 250, 500, 2000$ for the heavy-tailed and sub-Gaussian processes. From Figures 3 and 4, we find that the parametric estimation methods show better performance than the non-parametric POET-PAVM estimator. When comparing the OLS and LASSO estimators, the LASSO estimator performs better than the OLS estimator in terms of the relative Frobenius norm. One possible explanation for this is that the OLS estimator can partially explain the volatility dynamics via the factor component, but fails to explain the whole dynamics. On the other hand, the Frobenius and spectral norm errors are similar for the OLS and LASSO estimators. This may be because they are highly affected by the errors in estimating large eigenvalues, such as the eigenvalues of the factor volatility matrix. Furthermore, the max norm error is also similar for the OLS and LASSO estimators. This may be because the OLS estimator does not have as many outliers as the LASSO estimator since the OLS estimator uses the average of the previous 22-day's idiosyncratic volatility matrices. Finally, the H-LASSO estimator shows the best performance for the heavy-tailed and

sub-Gaussian processes. These results are consistent with our notion that the H-LASSO estimator is robust to the heavy-tailedness, and it can explain the dynamics from the factor and idiosyncratic components.

5.2 An empirical study

We applied the proposed FIVAR(h)-Itô model to real high-frequency trading data for 200 assets from January 2016 to December 2019 (997 trading days). The top 200 large trading volume stocks among the S&P 500 were selected from the Wharton Data Service (WRDS) system. The trading volumes were calculated using the data from years 2015 to 2020. To synchronize the high-frequency data, we used the previous tick scheme (Andersen et al., 2003; Barndorff-Nielsen et al., 2011a; Zhang, 2011) with equal distance intervals, which helps mitigate the errors from the irregularity of the observation times. When applying the refresh time scheme for the 200 assets, we find that the average number of daily synchronized samples is 593.49, which corresponds to 39.42-sec sampling frequency. Hence, we chose 1-min sampling frequency to enjoy the benefit of large samples, which corresponds to $m = 390$. We excluded days with half trading hours. The data of year 2020 was excluded to avoid the effects of extreme market conditions. Specifically, we performed the structural break test for the eigenvalue process in Section A.3 and found the non-stationarity when including 2020 data. We note that approximately 10 CPU cores with 2 GHz and 200 GB of RAM are required to conduct the whole empirical study within 3 days.

To apply the proposed estimation procedures, we first calculated 997 daily integrated volatility matrices using the jump adjusted pre-averaging volatility matrix (PAVM) estimator in (A.1). We chose K as 19 and $c_{i,u}$ as 7 times the sample standard deviation of the pre-averaged variables $m^{1/4}\bar{Y}_i(t_{d,k})$. We projected the daily PAVM estimators onto the positive semi-definite cone in the spectral norm to make them positive semi-definite. In the empirical study, to predict the future volatility matrix, we employed the rolloing window scheme with the in-sample period $n = 251$. For each in-sample period, we estimated the rank r based on the rank estimation procedure in (5.1) with $n = 251$.

To estimate the idiosyncratic volatility matrix $\mathbf{\Sigma}_d$, we utilized the hard thresholding scheme

Figure 3: The log Frobenius, max, relative Frobenius, and spectral norm error plots of the POET-PAVM, OLS, LASSO, and H-LASSO estimators for the conditional expected integrated volatility matrix estimation with the heavy-tailed process, given $n = 100, 200, 500$ and $m = 250, 500, 2000$.
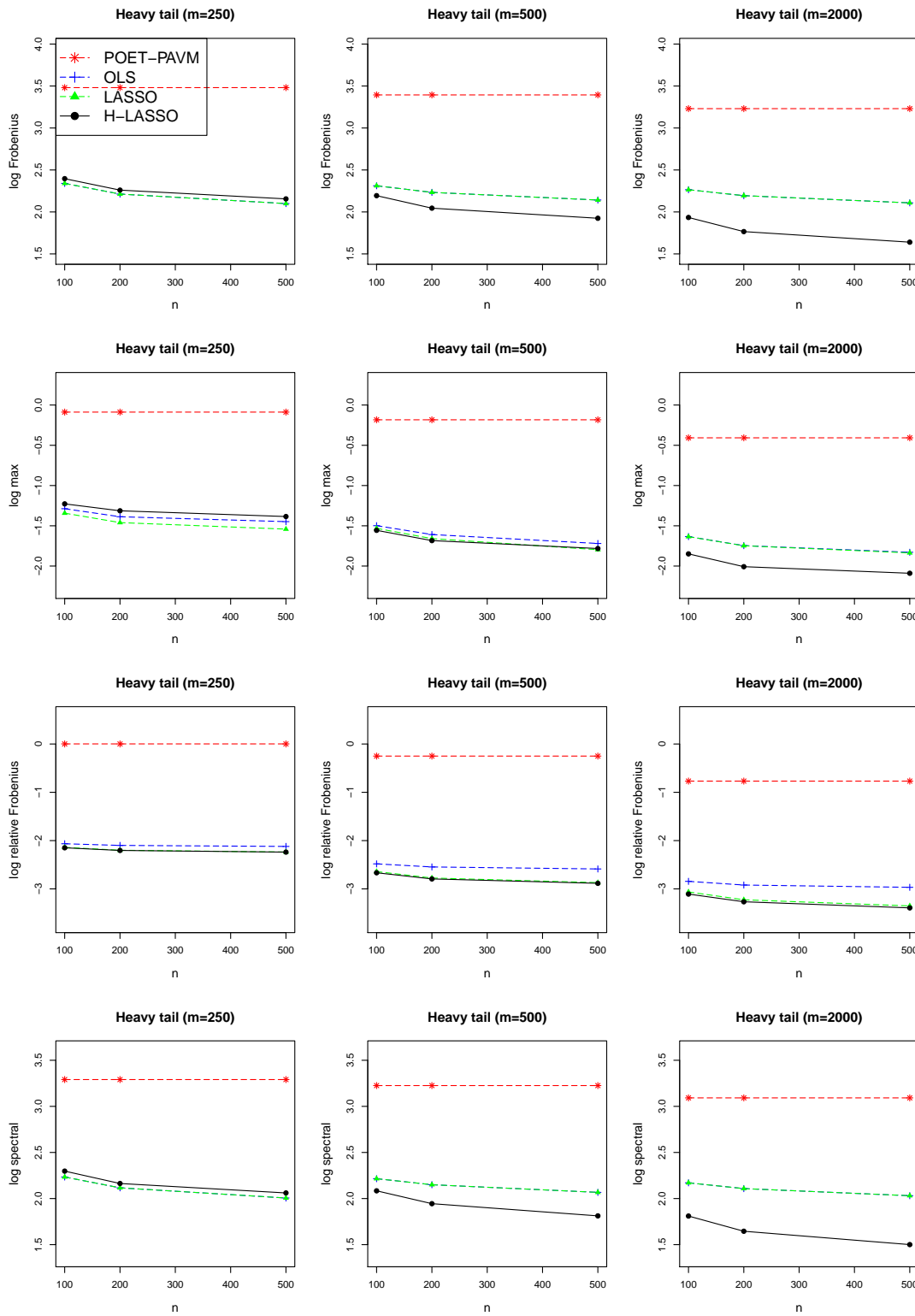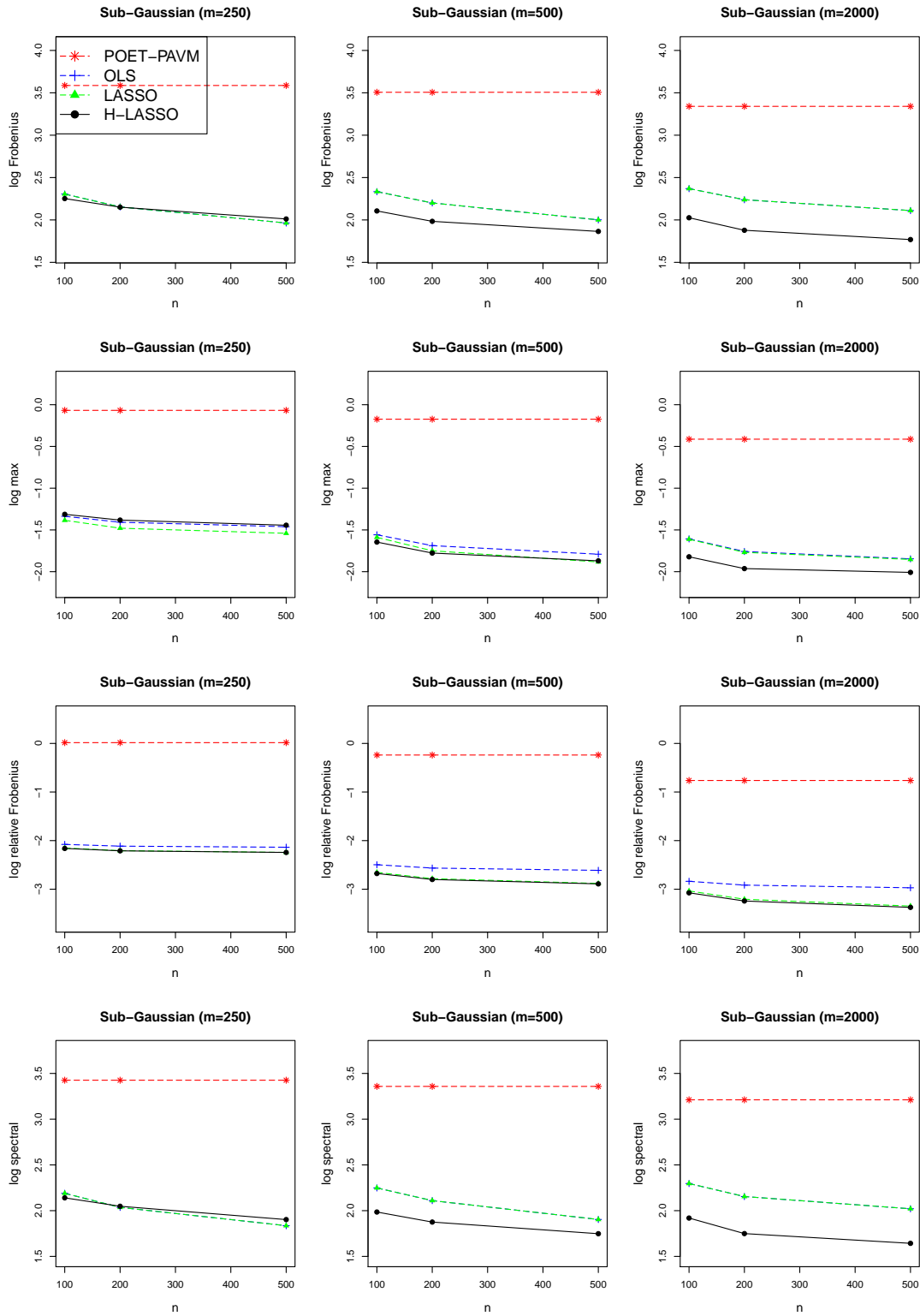
Figure 4: The log Frobenius, max, relative Frobenius, and spectral norm error plots of the POET-PAVM, OLS, LASSO, and H-LASSO estimators for the conditional expected integrated volatility matrix estimation with the sub-Gaussian process, given $n = 100, 200, 500$ and $m = 250, 500, 2000$.

based on the Global Industry Classification Standard (GICS) proposed by Fan et al. (2016). Specifically, the idiosyncratic components for the different sectors were set to zero, and we maintained those for the same sector. This corresponds to the hard-thresholding scheme with the sector information.

To choose the tuning parameters ℓ, $c_{F,1}$, $c_{F,2}$, $c_{I,1}$, and $c_{I,2}$, we defined

$$\Lambda^F(\ell, c_{F,1}, c_{F,2}) = \frac{1}{T} \sum_{d=1}^{T} \|\widehat{\boldsymbol{\Psi}}_d^{\text{H-LASSO}}(\ell, c_{F,1}, c_{F,2}) - \widehat{\boldsymbol{\Psi}}_{d+1}^{\text{POET}}\|_F^2,$$

$$\Lambda^I(\ell, c_{I,1}, c_{I,2}) = \frac{1}{T} \sum_{d=1}^{T} \|\widehat{\boldsymbol{\Sigma}}_d^{\text{H-LASSO}}(\ell, c_{I,1}, c_{I,2}) - \widehat{\boldsymbol{\Sigma}}_{d+1}^{\text{POET}}\|_F^2,$$

where $\widehat{\boldsymbol{\Psi}}_d^{\text{H-LASSO}}(\ell, c_{F,1}, c_{F,2})$ is the factor volatility matrix forecast from the H-LASSO estimator with the tuning parameters ℓ, $c_{F,1}$, and $c_{F,2}$ for the d-th day, and $\widehat{\boldsymbol{\Sigma}}_d^{\text{H-LASSO}}(\ell, c_{I,1}, c_{I,2})$ is the idiosyncratic volatility matrix forecast from the H-LASSO estimator with the tuning parameters ℓ, $c_{I,1}$, and $c_{I,2}$ for the d-th day. Also, $\widehat{\boldsymbol{\Psi}}_d^{\text{POET}}$ and $\widehat{\boldsymbol{\Sigma}}_d^{\text{POET}}$ are the factor and idiosyncratic volatility matrix estimators from the POET-PAVM estimator for the d-th day, respectively. Then, we selected ℓ, $c_{F,1}$, and $c_{F,2}$ by minimizing $\Lambda^F(\ell, c_{F,1}, c_{F,2})$ over $\ell \in \{1, 5, 22, 251\}$ and $c_{F,1}, c_{F,2} \in \{1/4, 1/2, 1, 2, 4\}$. Similarly, we chose ℓ, $c_{I,1}$, and $c_{I,2}$ by minimizing $\Lambda^I(\ell, c_{I,1}, c_{I,2})$ over $\ell \in \{1, 5, 22, 251\}$ and $c_{I,1}, c_{I,2} \in \{1/4, 1/2, 1, 2, 4\}$. For the choice of tuning parameters, we chose the in-sample period as $n = 251$ and out-of-sample period as day 252 to day 500 (year 2017). The selected parameters are $\ell = 22$, $c_{F,1} = 4$, $c_{F,2} = 1/4$, $c_{I,1} = 4$, and $c_{I,2} = 4$. We note that $\ell = 22$ was chosen for both $\Lambda^F(\ell, c_{F,1}, c_{F,2})$ and $\Lambda^I(\ell, c_{I,1}, c_{I,2})$. We also note that the stationarity of the volatility process is a reasonable assumption, which justifies the above tuning parameter choice procedure. To determine the lag h, we applied the Bayesian information criterion (BIC) to the VAR model. It leads to $h = 1$ for all in-sample period. Then, we estimated the conditional expected volatility matrix $\mathrm{E}\left(\boldsymbol{\Gamma}_{n+1} | \mathcal{F}_n\right)$ with the POET-PAVM, OLS, LASSO, and H-LASSO estimators.

For a comparison, we employed the DCC-NL estimator (De Nard et al., 2021; Engle et al., 2019; Ledoit and Wolf, 2015, 2022), which employs the nonlinear shrinkage estimator and the dynamic conditional correlation (DCC) model (Engle, 2002). Specifically, let $\Delta \mathbf{Y}_d = (\Delta Y_{1,d}, \ldots, \Delta Y_{p,d})^\top$ and $\Delta Y_{i,d}$ be the daily return for the i-th asset and d-th day. To obtain the DCC-NL estimator,

we first employed the following GARCH$(1, 1)$ model:

$$U_{i,d}^2 = a_i + b_{1,i}\Delta Y_{i,d-1}^2 + b_{2,i}U_{i,d-1}^2,$$

where the conditional variance $U_{i,d}^2 = \text{var}\left(\Delta Y_{i,d}|\mathcal{F}_{d-1}\right)$. Based on the GARCH model, we calculated the conditional variance for the next day, $U_{i,n+1}^2$, and obtained the devolatilized returns

$$\Delta \mathbf{Y}_d^s = \left(\Delta Y_{1,d}/U_{1,d}, \ldots, \Delta Y_{p,d}/U_{p,d}\right)^\top.$$

With this devolatilized return series $\{\Delta \mathbf{Y}_d^s\}$, we obtained $\text{cov}\left(\Delta \mathbf{Y}_d^s\right)$ based on the nonlinear shrinkage. Then, we applied the DCC model with $\text{cov}\left(\Delta \mathbf{Y}_{i,d}^s\right)$ being used for correlation targeting, and calculated the conditional correlation matrix for the next day, $\mathbf{R}_{n+1} = \text{corr}\left(\Delta \mathbf{Y}_{n+1}|\mathcal{F}_n\right)$. Finally, we estimated the conditional covariance matrix for the next day, $\text{cov}\left(\Delta \mathbf{Y}_{n+1}|\mathcal{F}_n\right)$, as $\mathbf{U}_{n+1}\mathbf{R}_{n+1}\mathbf{U}_{n+1}$, where $\mathbf{U}_{n+1} = \text{Diag}\left(U_{1,n+1}, \ldots, U_{p,n+1}\right)$. Detailed estimation procedure can be found in Engle et al. (2019). We also employed the HAR-DRD model (Oh and Patton, 2016) based on the POET-PAVM estimator. Specifically, we first decomposed the POET-PAVM estimator for the d-th day, $\widehat{\mathbf{\Gamma}}_d^{POET} = \left(\widehat{\Gamma}_{d,ij}^{POET}\right)_{1 \leq i,j \leq p}$, into

$$\widehat{\mathbf{\Gamma}}_d^{POET} = \sqrt{\mathbf{D}_d}\mathbf{R}_d\sqrt{\mathbf{D}_d},$$

where $\mathbf{D}_d = (D_{d,ij})_{1 \leq i,j \leq p} = \text{Diag}\left(\widehat{\Gamma}_{d,11}^{POET}, \ldots, \widehat{\Gamma}_{d,pp}^{POET}\right)$ is the diagonal matrix of realized volatilities and $\mathbf{R}_d$ is the realized correlation matrix. Then, we applied the following HAR model to each realized volatility:

$$D_{d+1,ii} = a_i + b_i^{(day)}D_{d,ii} + b_i^{(week)}\frac{1}{5}\sum_{j=0}^{4}D_{d-j,ii} + b_i^{(month)}\frac{1}{22}\sum_{j=0}^{21}D_{d-j,ii} + e_{d+1,i}, \quad i = 1, \ldots, p.$$

We then obtained the conditional realized volatilities for the next day. To ensure that the volatility forecasts are positive, we set their lower bound as 10^{-7}. For the realized correlations, we applied the following HAR-type model:

$$\text{vech}\left(\mathbf{R}_{d+1}\right) = \text{vech}\left(\bar{\mathbf{R}}_d\right)(1 - a - b - c) + a \cdot \text{vech}\left(\mathbf{R}_d\right)$$

$$+b \cdot \frac{1}{5} \sum_{j=0}^{4} \text{vech} \left(\mathbf{R}_{d-j} \right) + c \cdot \frac{1}{22} \sum_{j=0}^{21} \text{vech} \left(\mathbf{R}_{d-j} \right) + \mathbf{e}_{d+1},$$

where $\bar{\mathbf{R}}_d = \frac{1}{d} \sum_{j=1}^{d} \mathbf{R}_j$ and $(a, b, c) \in \mathbb{R}^3$. Then, we forecast the next day's volatility matrix based on the conditional realized volatilities and correlations. We call it the HAR-DRD estimator. To predict the future volatility matrix, we also employed the rolling window scheme with the in-sample period of 251 days for the DCC-NL and HAR-DRD models. We note that for all estimators, including the HAR-DRD estimator, the logarithm of the realized volatility is not used to check the effect of modeling idiosyncratic volatilities in linear modeling approaches. In fact, there are some cases in which the logarithm improves the performance of the volatility estimators. However, it is difficult to model the log-volatility in the high-dimensional high-frequency set-up. For example, Kim (2022) introduced the exponential GARCH-Itô volatility model for the one-dimensional case, but the extension to the high-dimensional case is not straightforward. We leave this issue for a future study. We note that all estimators except the DCC-NL estimator use the POET-PAVM estimator as an input. Since the DCC-NL estimator uses the daily total returns, only the DCC-NL estimator forecasts the total volatility. Thus, we adjusted the jump component in the following applications. To do this, we obtained the jump volatility matrix estimator by subtracting the jump adjusted PAVM estimator from the PAVM estimator. The PAVM estimator can be obtained by setting the truncation parameters as infinity in (A.1).

To investigate the performance of the future volatility matrix estimators, we employed the high-frequency data from 2017 to 2019. We chose the in-sample period as $n = 251$ (one year), and we used three different out-of-sample periods: 2018 and 2019 (period 1), 2018 only (period 2), and 2019 only (period 3). For the period 1, we calculated the average number of non-zero elements in $\widehat{\boldsymbol{\beta}}_i$ excluding the intercept term over $i = 4, \ldots, 203$. The results are 2.790 and 2.857 for the H-LASSO and LASSO estimators, respectively.

To check the performance of the proposed estimation procedures, we first investigated the following mean squared prediction error (MSPE) and QLIKE (Bollerslev et al., 2018; Laurent

et al., 2013):

$$\text{MSPE}(\widetilde{\mathbf{\Gamma}}) = \frac{1}{T} \sum_{d=1}^{T} \|\widetilde{\mathbf{\Gamma}}_d - \widehat{\mathbf{\Gamma}}_d^{POET}\|_F^2,$$

$$\text{QLIKE}(\widetilde{\mathbf{\Gamma}}) = \frac{1}{T} \sum_{d=1}^{T} \log \left(\det \left(\widetilde{\mathbf{\Gamma}}_d \right) \right) + \text{tr} \left(\widetilde{\mathbf{\Gamma}}_d^{-1} \widehat{\mathbf{\Gamma}}_d^{POET} \right), \tag{5.2}$$

where T is the number of days in the out-of-sample period, $\widetilde{\mathbf{\Gamma}}_d$ is one of the future volatility matrix forecasts from the POET-PAVM, OLS, LASSO, H-LASSO, DCC-NL, and HAR-DRD estimators for the d-th day of the out-of-sample period, and $\widehat{\mathbf{\Gamma}}_d^{POET}$ is the POET-PAVM estimator for the d-th day, which is a proxy of the ground truth. For the DCC-NL estimator, we subtracted the next day's jump volatility matrix estimator, which helps improve the performance of the DCC-NL estimator. We note that MSPE is a form of the mean squared error that is one of the robust loss functions for volatility comparisons (Hansen and Lunde, 2006a; Patton, 2011; Patton and Sheppard, 2009). Also, QLIKE is robust to the presence of noise in the volatility proxy (Hansen and Lunde, 2006a; Laurent et al., 2013; Patton, 2011; Patton and Sheppard, 2009). Table 2 reports the MSPE and QLIKE results of the POET-PAVM, OLS, LASSO, H-LASSO, DCC-NL, and HAR-DRD estimators for three out-of-sample periods. We find that the H-LASSO estimator shows good performance in terms of both MSPE and QLIKE. We note that the LASSO and H-LASSO estimators have similar performance in terms of the QLIKE. This may be because the QLIKE is highly affected by the small eigenvalue estimation, and the small eigenvalues are less affected by the heavy-tailedness. On the other hand, the HAR-DRD estimator shows the best performance in terms of MSPE, but it did not perform well for the QLIKE loss. These results show the proposed H-LASSO estimator can help explain the dynamics of the idiosyncratic volatility matrix under the sparsity condition and the heavy-tailedness of the financial data.

To investigate the out-of-sample portfolio allocation performance, we applied the proposed estimators to the following minimum variance portfolio allocation problem:

$$\min_{\omega} \omega^\top \widetilde{\mathbf{\Gamma}}_d \omega, \quad \text{subject to } \omega^\top \mathbf{J} = 1 \text{ and } \|\omega\|_1 \leq c_0,$$

Table 2: The MSPE and QLIKE of the POET-PAVM, OLS, LASSO, H-LASSO, DCC-NL, and HAR-DRD estimators (period 1, from 2018 to 2019; period 2, 2018; period 3, 2019).

		POET-PAVM	OLS	LASSO	H-LASSO	DCC-NL	HAR-DRD
Period 1	MSPE $\times 10^4$	3.401	3.056	3.058	2.597	3.002	2.583
	QLIKE $\times 10^{-3}$	-1.050	-1.660	-1.670	-1.669	-1.641	-1.530
Period 2	MSPE $\times 10^4$	4.522	4.512	4.516	3.682	4.243	3.748
	QLIKE $\times 10^{-3}$	-1.017	-1.638	-1.643	-1.643	-1.615	-1.390
Period 3	MSPE $\times 10^4$	2.275	1.594	1.594	1.507	1.756	1.413
	QLIKE $\times 10^{-3}$	-1.082	-1.681	-1.696	-1.696	-1.666	-1.670

Figure 5: The out-of-sample risks of the minimum variance portfolios constructed by the POET-PAVM, OLS, LASSO, H-LASSO, DCC-NL, and HAR-DRD estimators.

where $\mathbf{J} = (1, \ldots, 1)^\top \in \mathbb{R}^p$, c_0 is the gross exposure constraint that changed from 1 to 3, and $\widetilde{\mathbf{\Gamma}}_d$ is one of the future volatility matrix estimators from POET-PAVM, OLS, LASSO, H-LASSO, DCC-NL, and HAR-DRD. Specifically, for all estimators except the DCC-NL estimator, we added the previous day's jump volatility matrix estimator to obtain $\widetilde{\mathbf{\Gamma}}_d$. To obtain the out-of-sample risks, we constructed the portfolios at the beginning of each trading day, based on the stock weights calculated using each future volatility matrix estimator. The portfolios were maintained for one day, and we calculated the realized volatility using the 10-min portfolio log-returns to mitigate the microstructural noise effect. We measured the out-of-sample risk using the average of their square root for each out-of-sample period. Figure 5 depicts the out-of-sample risks of the port-
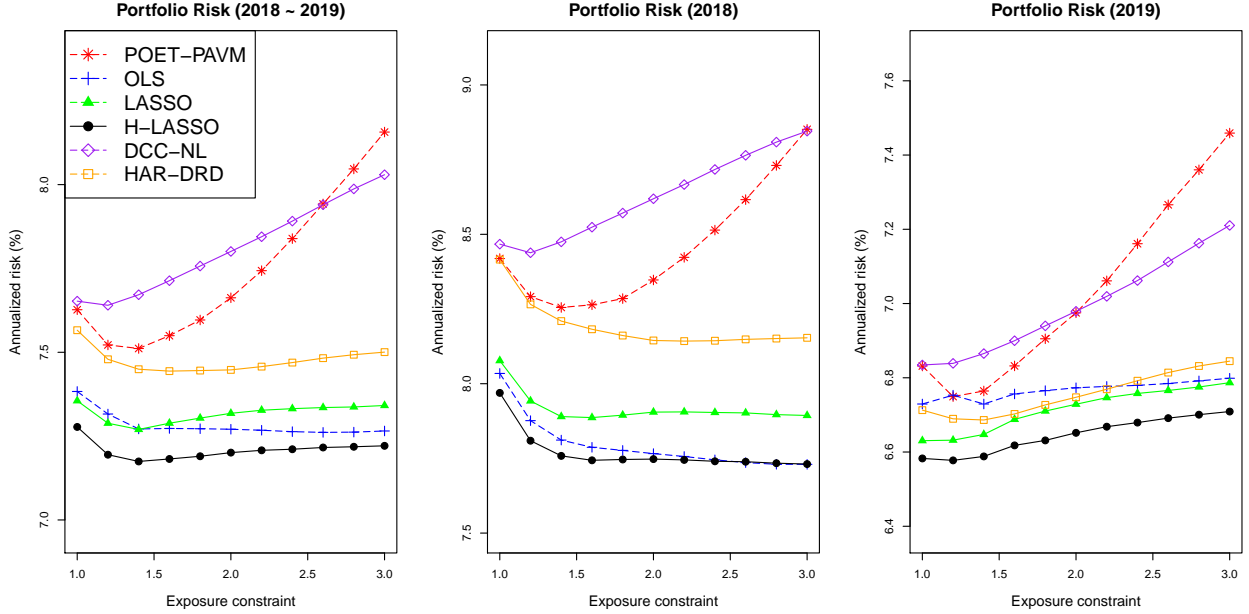
folios constructed using the POET-PAVM, OLS, LASSO, H-LASSO, DCC-NL, and HAR-DRD estimators. From Figure 5, we find that the POET-PAVM and DCC-NL estimators become unstable as the gross exposure constraint increases. This may be because the POET-PAVM estimator cannot explain the dynamics of the volatility process, and the DCC-NL estimator only uses the low-frequency information. On the other hand, the H-LASSO estimator has a stable result and has the smallest risk overall. These results indicate that considering both dynamic structures in idiosyncratic volatility and heavy-tailedness in financial data helps account for the dynamics of large volatility matrix processes.

6 Conclusion

In this paper, we develop a novel factor and idiosyncratic VAR-Itô (FIVAR-Itô) model to account for the dynamic structure of the large volatility matrix, which has the low-rank plus sparse structure. Under the FIVAR-Itô model, we show that the daily eigenvalues of the factor and idiosyncratic volatility matrices have the VAR model structure. To further allow the heavy-tailedness in financial data, we use the bounded moment condition for the VAR model. Then, we propose a robust estimation procedure for the VAR model parameters, which employs the truncation method and ℓ_1-penalty to deal with the heavy-tailedness and explore the sparsity. We show that it can handle the heavy-tailedness, observation error, and high dimensionality with a desirable convergence rate. We also propose the large volatility prediction procedure and investigate its asymptotic properties.

In the empirical study, in terms of the prediction error and portfolio allocation, the proposed estimator shows the best performance overall except for the MSPE measurement. We note that for MSPE, H-LASSO shows the second-best performance. It reveals that, when predicting large volatility matrices, the proposed estimation method helps handle the heavy-tailedness of financial data and explain the dynamic structure of factor and idiosyncratic volatility matrices. On the other hand, one of the key assumptions in the proposed model is the sparsity condition of the model parameters. Thus, it would be interesting to construct a test procedure for the sparsity condition. To do this, we may need to debias the biased H-LASSO estimator and to derive its asymptotic distribution under the sparsity hypothesis. This is a theoretically demanding task. It would

be interesting and important to develop a tuning parameter choice procedure that has rigorous theoretical properties and works well in practice. However, developing a tuning parameter selection procedure that works well from both practical and theoretical perspectives may be challenging. We leave these topics for future studies.

References

Aït-Sahalia, Y., Fan, J., and Xiu, D. (2010). High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association*, 105(492):1504–1517.

Aït-Sahalia, Y., Mykland, P. A., and Zhang, L. (2011). Ultra high frequency volatility estimation with dependent microstructure noise. *Journal of Econometrics*, 160(1):160–175.

Aït-Sahalia, Y. and Xiu, D. (2016). Increased correlation among asset classes: Are volatility or jumps to blame, or both? *Journal of Econometrics*, 194(2):205–219.

Aït-Sahalia, Y. and Xiu, D. (2017). Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *Journal of Econometrics*, 201(2):384–399.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625.

Andersen, T. G., Thyrsgaard, M., and Todorov, V. (2019). Time-varying periodicity in intraday volatility. *Journal of the American Statistical Association*, 114(528):1695–1707.

Andersen, T. G., Thyrsgaard, M., and Todorov, V. (2021). Recalcitrant betas: Intraday variation in the cross-sectional dispersion of systematic risk. *Quantitative Economics*, 12(2):647–682.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of applied econometrics*, 18(1):1–22.

Bandi, F. M. and Reno, R. (2016). Price and volatility co-jumps. *Journal of Financial Economics*, 119(1):107–146.

Barigozzi, M. and Hallin, M. (2016). Generalized dynamic factor models and volatilities: recovering the market volatility shocks. *The Econometrics Journal*, 19(1):C33–C60.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6):1481–1536.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2011a). Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics*, 162(2):149–169.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2011b). Subsampling realised kernels. *Journal of Econometrics*, 160(1):204–219.

Bibinger, M., Hautsch, N., Malec, P., and Reiß, M. (2014). Estimating the quadratic covariation matrix from noisy observations: Local method of moments and efficiency. *The Annals of Statistics*, 42(4):1312–1346.

Bibinger, M. and Winkelmann, L. (2015). Econometrics of co-jumps in high-frequency data with noise. *Journal of Econometrics*, 184(2):361–378.

Bibinger, M. and Winkelmann, L. (2018). Common price and volatility jumps in noisy high-frequency data. *Electronic Journal of Statistics*, 12(1):2018–2073.

Bollerslev, T., Patton, A. J., and Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, 192(1):1–18.

Bollerslev, T., Patton, A. J., and Quaedvlieg, R. (2018). Modeling and forecasting (un) reliable realized covariances for more reliable financial decisions. *Journal of Econometrics*, 207(1):71–91.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185.

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

Chun, D. and Kim, D. (2022). State heterogeneity analysis of financial volatility using high-frequency financial data. *Journal of Time Series Analysis*, 43(1):105–124.

Cipollini, F., Gallo, G. M., and Otranto, E. (2021). Realized volatility forecasting: Robustness to measurement errors. *International Journal of Forecasting*, 37(1):44–57.

Connor, G., Korajczyk, R. A., and Linton, O. (2006). The common and specific components of dynamic volatility. *Journal of Econometrics*, 132(1):231–255.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196.

Davydov, Y. A. (1968). Convergence of distributions generated by stationary stochastic processes. *Theory of Probability & Its Applications*, 13(4):691–696.

De Nard, G., Ledoit, O., and Wolf, M. (2021). Factor models for portfolio selection in large dimensions: The good, the better and the ugly. *Journal of Financial Econometrics*, 19(2):236–257.

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3):339–350.

Engle, R. F., Ledoit, O., and Wolf, M. (2019). Large dynamic covariance matrices. *Journal of Business & Economic Statistics*, 37(2):363–375.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197.

Fan, J., Furger, A., and Xiu, D. (2016). Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high frequency data. *Journal of Business & Economic Statistics*, 34:489–503.

Fan, J., Guo, Y., and Jiang, B. (2019). Adaptive huber regression on markov-dependent data. *Stochastic Processes and their Applications*.

Fan, J. and Kim, D. (2018). Robust high-dimensional volatility matrix estimation for high-frequency factor model. *Journal of the American Statistical Association*, 113(523):1268–1283.

Fan, J. and Kim, D. (2019). Structured volatility matrix estimation for non-synchronized high-frequency financial data. *Journal of Econometrics*, 209(1):61–78.

Fan, J., Li, Q., and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 79(1):247.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J., Li, Y., Xia, N., and Zheng, X. (2024). Tests for principal eigenvalues and eigenvectors. *arXiv preprint arXiv:2405.06939*.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680.

Fan, J., Liu, H., Sun, Q., and Zhang, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Annals of statistics*, 46(2):814.

Fan, J. and Wang, Y. (2007). Multi-scale jump and volatility analysis for high-frequency financial data. *Journal of the American Statistical Association*, 102(480):1349–1362.

Hansen, P. R., Huang, Z., and Shek, H. H. (2012). Realized garch: a joint model for returns and realized measures of volatility. *Journal of Applied Econometrics*, 27(6):877–906.

Hansen, P. R. and Lunde, A. (2006a). Consistent ranking of volatility models. *Journal of Econometrics*, 131(1-2):97–121.

Hansen, P. R. and Lunde, A. (2006b). Realized variance and market microstructure noise. *Journal of Business & Economic Statistics*, 24(2):127–161.

Hansen, P. R. and Lunde, A. (2014). Estimating the persistence and the autocorrelation function of a time series that is measured with error. *Econometric Theory*, 30(1):60–93.

Herskovic, B., Kelly, B., Lustig, H., and Van Nieuwerburgh, S. (2016). The common factor in idiosyncratic volatility: Quantitative asset pricing implications. *Journal of Financial Economics*, 119(2):249–283.

Hetland, S., Pedersen, R. S., and Rahbek, A. (2023). Dynamic conditional eigenvalue garch. *Journal of Econometrics*, 237(2):105175.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Jacod, J., Klüppelberg, C., and Müller, G. (2017a). Testing for non-correlation between price and volatility jumps. *Journal of Econometrics*, 197(2):284–297.

Jacod, J., Li, Y., Mykland, P. A., Podolskij, M., and Vetter, M. (2009). Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic processes and their applications*, 119(7):2249–2276.

Jacod, J., Li, Y., and Zheng, X. (2017b). Statistical properties of microstructure noise. *Econometrica*, 85(4):1133–1174.

Jacod, J. and Todorov, V. (2010). Do price and volatility jump together? *The Annals of Applied Probability*, 20(4):1425–1469.

Kim, D. (2022). Exponential realized garch-itô volatility models. *Econometric Theory*, pages 1–37.

Kim, D. and Fan, J. (2019). Factor garch-itô models for high-frequency data with application to large volatility matrix prediction. *Journal of Econometrics*, 208(2):395–417.

Kim, D., Liu, Y., and Wang, Y. (2018). Large volatility matrix estimation with factor-based diffusion model for high-frequency financial data. *Bernoulli*, 24(4B):3657–3682.

Kim, D., Oh, M., Song, X., and Wang, Y. (2023). Factor overnight garch-itˆo models. *Available at SSRN 4342551*.

Kim, D., Song, X., and Wang, Y. (2022). Unified discrete-time factor stochastic volatility and continuous-time itô models for combining inference based on low-frequency and high-frequency. *Journal of Multivariate Analysis*, 192:105091.

Kim, D. and Wang, Y. (2016). Unified discrete-time and continuous-time models and statistical inferences for merged low-frequency and high-frequency financial data. *Journal of Econometrics*, 194:220–230.

Kim, D., Wang, Y., and Zou, J. (2016). Asymptotic theory for large volatility matrix estimation based on high-frequency financial data. *Stochastic Processes and their Applications*, 126:3527—-3577.

Koike, Y. (2016). Quadratic covariation estimation of an irregularly observed semimartingale with jumps and noise. *Bernoulli*, 22(3):1894–1936.

Kong, X.-B. (2018). On the systematic and idiosyncratic volatility with large panel high-frequency data. *Annals of Statistics*, 46(3):1077–1108.

Kong, X.-B., Lin, J.-G., Liu, C., and Liu, G.-Y. (2021). Discrepancy between global and local principal component analysis on large-panel high-frequency data. *Journal of the American Statistical Association*, pages 1–12.

Laurent, S., Rombouts, J. V., and Violante, F. (2013). On loss functions and ranking forecasting performances of multivariate volatility models. *Journal of Econometrics*, 173(1):1–10.

Ledoit, O. and Wolf, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions. *Journal of Multivariate Analysis*, 139:360–384.

Ledoit, O. and Wolf, M. (2022). The power of (non-) linear shrinking: A review and guide to covariance matrix estimation. *Journal of Financial Econometrics*, 20(1):187–218.

Li, Z. M. and Linton, O. (2022). A ReMeDI for microstructure noise. *Econometrica*, 90(1):367–389.

Li, Z. M. and Linton, O. B. (2020). Robust estimation of integrated volatility. *Available at SSRN 3702143*.

Liebscher, E. (2005). Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. *Journal of Time Series Analysis*, 26(5):669–689.

Mao, G. and Zhang, Z. (2018). Stochastic tail index model for high frequency financial data with bayesian analysis. *Journal of Econometrics*, 205(2):470–487.

Merlevède, F., Peligrad, M., and Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: The Luminy volume*, volume 5, pages 273–292. Institute of Mathematical Statistics.

Minsker, S. (2018). Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903.

Oh, D. H. and Patton, A. J. (2016). High-dimensional copula-based distributions with mixed frequency data. *Journal of Econometrics*, 193(2):349–366.

Park, S., Hong, S. Y., and Linton, O. (2016). Estimating the quadratic covariation matrix for asynchronously observed high frequency stock returns corrupted by additive measurement error. *Journal of Econometrics*, 191(2):325–347.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256.

Patton, A. J. and Sheppard, K. (2009). Evaluating volatility and correlation forecasts. In *Handbook of financial time series*, pages 801–838. Springer.

Shephard, N. and Sheppard, K. (2010). Realising the future: forecasting with high-frequency-based volatility (heavy) models. *Journal of Applied Econometrics*, 25(2):197–231.

Shin, M., Kim, D., and Fan, J. (2023). Adaptive robust large volatility matrix estimation based on high-frequency financial data. *Journal of Econometrics*, 237(1):105514.

Song, X., Kim, D., Yuan, H., Cui, X., Lu, Z., Zhou, Y., and Wang, Y. (2021). Volatility analysis with realized garch-itô models. *Journal of Econometrics*, 222(1):393–410.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460):1167–1179.

Sun, Q., Zhou, W.-X., and Fan, J. (2020). Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Ubukata, M. and Oya, K. (2009). Estimation and testing for dependence in market microstructure noise. *Journal of Financial Econometrics*, 7(2):106–151.

Wang, Y. and Zou, J. (2010). Vast volatility matrix estimation for high-frequency financial data. *The Annals of Statistics*, 38:943–978.

Wong, K. C., Li, Z., and Tewari, A. (2020). Lasso guarantees for β-mixing heavy-tailed time series. *Annals of Statistics*, 48(2):1124–1142.

Xiu, D. (2010). Quasi-maximum likelihood estimation of volatility with high frequency data. *Journal of Econometrics*, 159(1):235–250.

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323.

Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli*, 12(6):1019–1043.

Zhang, L. (2011). Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics*, 160(1):33–47.

Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472):1394–1411.

A Appendix

A.1 A simulation setup

We considered the following jump diffusion process with the FIVAR(h)-Itô model in Definition 1:

$$d\mathbf{X}(t) = \mathbf{Q}_F \boldsymbol{\lambda}_F^{1/2}(t)d\mathbf{W}(t) + \mathbf{Q}_I \boldsymbol{\lambda}_I^{1/2}(t)\mathbf{W}^*(t) + \mathbf{J}(t)d\boldsymbol{\Lambda}(t),$$

$$\lambda_{t,i}(\boldsymbol{\theta}_i) = (\lceil t \rceil - t)\lambda_{\lceil t-1 \rceil,i}(\boldsymbol{\theta}_i) + \sum_{j=1}^{p+r} \zeta_{1,i,j} \left\{ \int_{\lceil t-1 \rceil}^{t} \lambda_{s,j}(\boldsymbol{\theta}_j)\,ds \right\} + (\lceil t \rceil - t)\int_{\lceil t-1 \rceil}^{t} J_{\lambda,i}(s)d\widetilde{\Lambda}_{\lambda,i}(s)$$

$$+ (t - \lceil t \rceil + 1)\left(a_i + \sum_{k=2}^{h}\sum_{j=1}^{p+r} \zeta_{k,i,j} \left\{ \int_{\lceil t \rceil - k}^{\lceil t \rceil - k+1} \lambda_{s,j}(\boldsymbol{\theta}_j)\,ds \right\} \right) + (\lceil t \rceil - t)Z_{i,t}^2,$$

where $\boldsymbol{\lambda}_F(t) = \mathrm{Diag}(p\lambda_{t,1}(\boldsymbol{\theta}_1), ..., p\lambda_{t,r}(\boldsymbol{\theta}_r))$, $\boldsymbol{\lambda}_I(t) = \mathrm{Diag}(\lambda_{t,r+1}(\boldsymbol{\theta}_{r+1}), ..., \lambda_{t,p+r}(\boldsymbol{\theta}_{p+r}))$, and $\mathbf{W}(t)$ and $\mathbf{W}^*(t)$ are r-dimensional and p-dimensional independent Brownian motions, respectively, $\mathbf{J}(t) = (J_1(t), \ldots, J_p(t))^\top$ is the jump size vector, and $\boldsymbol{\Lambda}(t) = (\Lambda_1(t), \ldots, \Lambda_p(t))^\top$ is the Poisson process with intensity $\mathbf{I}(t) = (5, \ldots, 5)^\top$. The jump size $J_i(t)$ was obtained from the independent Gaussian distribution with mean zero and standard deviation $0.05\sqrt{\int_0^1 \gamma_{ii}(t)dt}$. For $t \in [d-1, d)$, $i = 1, \ldots, p+r$, and $d = 1, \ldots, n$, we set $J_{\lambda,i}(t) = J_{\lambda,i}(d-1)$, and $J_{\lambda,i}(d)$'s were generated from independent $\mathrm{unif}(-a_i/100, a_i/100)$. Also, the compensated Poisson process $\widetilde{\Lambda}_{\lambda,i}(t)$ has the intensity $I_{\lambda,i}(t) = (10, \ldots, 10)^\top$. To obtain the eigenvector matrix for the factor part, $\mathbf{Q}_F$, we first generated the symmetric p by p matrix whose elements were obtained from i.i.d. $\mathrm{unif}(0, 1)$. Then, we chose its first r eigenvectors as $\mathbf{Q}_F$. We chose the eigenvector matrix for the idiosyncratic part, $\mathbf{Q}_I$, as the p-dimensional identity matrix. For $t \in [d-1, d)$, $i = 1, \ldots, p+r$, and $d = 1, \ldots, n$, we set $z_{i,t} = z_{i,d-1}$. Let $v_1 = 0.6$, $v_i = 0.3$ for $i = 2, \ldots, r$, and $v_i = 0.1$ for $i = r+1, \ldots, p+r$. For the heavy-tailed process, $z_{i,d}$'s were obtained from v_i times independent t-distribution with degrees of freedom 9, while for the sub-Gaussian process, $z_{i,d}$'s were generated from v_i times independent $\mathrm{unif}(-2, 2)$. We chose $p = 200$, $r = 3$, $h = 1$, $m^{all} = 2000$, and we varied m from 250 to 2000. The model parameters are chosen as follows. We set $a_1 = 0.8$, $a_2 = a_3 = 0.6$, $a_i = (14 - i)/10$ for

$4 \leq i \leq 13$, $a_i = 0.1$ for $14 \leq i \leq 203$,

$$(\zeta_{1,i,j})_{1\leq i,j\leq 3} = \begin{pmatrix} 0.5 & 0.15 & 0 \\ 0 & 0.45 & 0.1 \\ 0 & 0 & 0.4 \end{pmatrix}, \quad (\zeta_{1,i,j})_{2k\leq i,j\leq 2k+1} = \begin{pmatrix} 0.19 - 0.02k & 0.02 \\ 0.02 & 0.18 - 0.02k \end{pmatrix}$$

for $2 \leq k \leq 6$, and $(\zeta_{1,i,j})_{14\leq i,j\leq 203}$ as 0.05 times 190-dimensional identity matrix. Other elements of $(\zeta_{1,i,j})$ were set as zero. We took $\mathbf{X}(0) = (0,\ldots,0)^\top$ and $\lambda_{0,i}(\boldsymbol{\theta}_i) = \mathrm{E}(\lambda_{1,i}(\boldsymbol{\theta}_i))$.

We calculated the jump adjusted pre-averaging volatility matrix (PAVM) estimator (Aït-Sahalia and Xiu, 2016; Christensen et al., 2010; Jacod et al., 2009) as follows:

$$\widehat{\Gamma}_{d,ij} = \frac{1}{\psi K} \sum_{k=1}^{m-K+1} \left\{ \bar{Y}_i(t_{d,k}) \bar{Y}_j(t_{d,k}) - \frac{1}{2}\widehat{Y}_{i,j}(t_{d,k}) \right\} \mathbf{1}\left\{ |\bar{Y}_i(t_{d,k})| \leq u_{i,m} \right\} \mathbf{1}\left\{ |\bar{Y}_j(t_{d,k})| \leq u_{j,m} \right\}, \tag{A.1}$$

where

$$\bar{Y}_i(t_{d,k}) = \sum_{l=1}^{K-1} g\left(\frac{l}{K}\right) (Y_i(t_{d,k+l}) - Y_i(t_{d,k+l-1})),$$

$$\widehat{Y}_{i,j}(t_{d,k}) = \sum_{l=1}^{K} \left[\left\{ g\left(\frac{l}{K}\right) - g\left(\frac{l-1}{K}\right) \right\}^2 \right.$$
$$\left. \times (Y_i(t_{d,k+l-1}) - Y_i(t_{d,k+l-2})) (Y_j(t_{d,k+l-1}) - Y_j(t_{d,k+l-2})) \right],$$

$\psi = \int_0^1 g(t)^2 dt$, $\mathbf{1}\{\cdot\}$ is an indicator function, and $u_{i,m} = c_{i,u} m^{-0.235}$ is a truncation parameter for some constant $c_{i,u}$. We chose the bandwidth parameter $K = \lfloor m^{1/2} \rfloor$, weight function $g(x) = x \wedge (1-x)$, and $c_{i,u}$ as 7 times the sample standard deviation for the pre-averaged variables $m^{1/4}\bar{Y}_i(t_{d,k})$.

A.2 Empirical study for the constant eigenvector hypothesis

In this section, we conducted a hypothesis test for the constant eigenvector assumption based on the procedure in Fan et al. (2024). For each day, we first splited the return data into two groups

as follows:

$$\Delta\mathbf{Y}_1^{(1)},\ldots,\Delta\mathbf{Y}_{T_1}^{(1)}, \quad \text{and} \quad \Delta\mathbf{Y}_1^{(2)},\ldots,\Delta\mathbf{Y}_{T_2}^{(2)},$$

where $\Delta\mathbf{Y}_i^{(j)} = \left(\Delta Y_{1,i}^{(j)},\ldots,\Delta Y_{p,i}^{(j)}\right)^\top$, $\Delta Y_{k,i}^{(j)}$ is the i-th return for the k-th asset and j-th group, and T_1 and T_2 are the numbers of returns for each group. With these return data, we obtained the sample covariance matrices as follows:

$$\widehat{\boldsymbol{\Gamma}}^{(j)} = \frac{1}{T_j} \sum_{i=1}^{T_j} \Delta\mathbf{Y}_i^{(j)}(\Delta\mathbf{Y}_i^{(j)})^\top \quad \text{for } j = 1, 2.$$

Then, we decomposed the above sample covariance matrices and obtained eigenvalue and eigenvector estimates for the two groups. With these estimates, we conducted the hypothesis test for the constancy of the eigenvector process. The null hypothesis is that the eigenvectors of the two groups are the same. Detailed test procedure is presented in Section 4.3 in Fan et al. (2024). The tests were conducted for three principal eigenvectors over 1498 days from 2015 to 2020. In this paper, to mitigate the effect of noises, we used 5-min log-returns and chose $T_1 = T_2 = 38$. As shown in Figure 6, we found that the constant eigenvector hypothesis is often rejected at 5% significance level.

A.3   Empirical study for the structural break test

In this section, we conducted the structural break test for the eigenvalue process based on the procedure in Bai and Perron (2003). Specifically, we considered the following linear regression model with L breaks:

$$1 = \mathbf{b}_k\widehat{\boldsymbol{\xi}}_d^\top + e_d, \quad d = d_{k-1} + 1,\ldots,d_k, \quad k = 1,\ldots,L+1,$$

where k denotes the segment index, $\mathbf{b}_k$ is the regression coefficient for the k-th segment, $\widehat{\boldsymbol{\xi}}_d = \left(\widehat{\xi}_{d,1},\ldots,\widehat{\xi}_{d,p+r}\right)^\top$ is a non-parametric integrated eigenvalue estimator for $\boldsymbol{\xi}_d$ whose estimation procedure is presented in Section 4, e_d is a error term for the d-th day, $d_0 = 0$, and d_{L+1} represents the end date of the data. The dependent variable was set as 1 to investigate the break in the mean
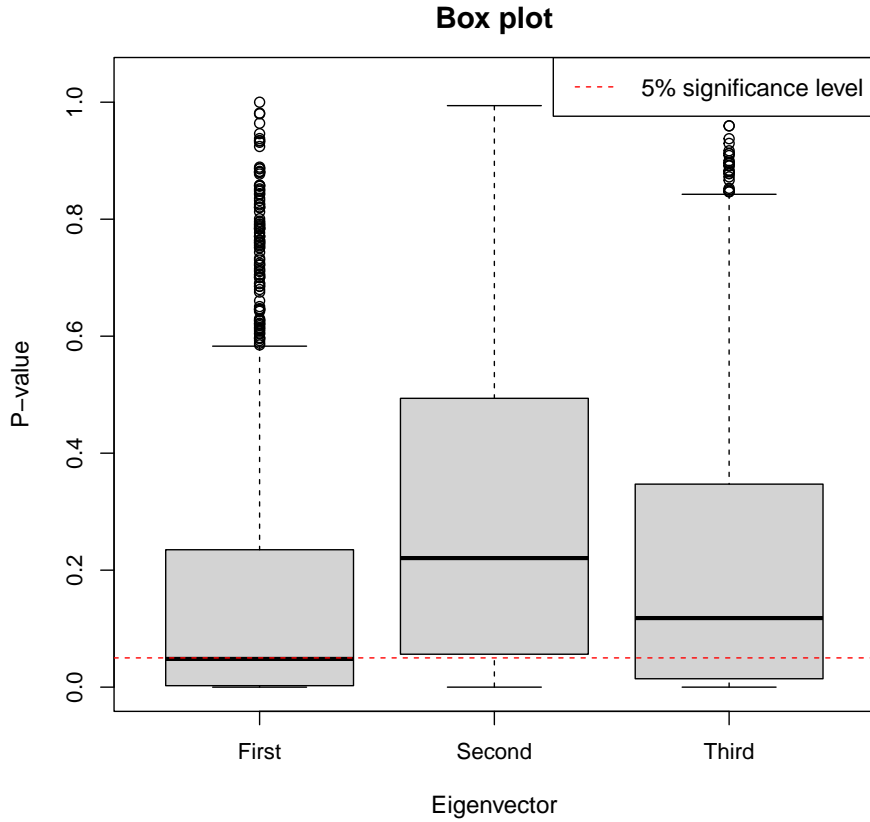
Figure 6: The box plots of the p-values for the constant eigenvector hypothesis for three principal eigenvectors over 1498 days from 2015 to 2020. We used 5-min log-returns of the top 200 large trading volume assets in the S&P 500 index. The red dash represents the 5% significance level.

of the eigenvalue process. For each $L = 1, \ldots, 5$, we estimated the breakpoints, $d_k, k = 1, \ldots, L$, by minimizing the residual sum of squares (RSS). Then, we chose $L \in \{0, \ldots, 5\}$ that minimizes the corresponding Bayesian information criterion (BIC). Details can be found in Bai and Perron (2003). The result is $L = 1$ with the breakpoint $d_1 = 1035$ that corresponds to March 10, 2020. This may be due to the covid sell-off in 2020. Thus, in the empirical study, we excluded the year 2020 to avoid the non-stationarity. We note that the proposed FIVAR-Itô model is based on the stationary condition and it is hard to apply the parametric model to the non-stationary period.

A.4 Proof of Proposition 1

Proof of Proposition 1. For non-negative integer $l \in \mathbb{N}_0$, let

$$\mathbf{R}(l) = (R_1(l), \ldots, R_{p+r}(l))^\top \text{ with } R_i(l) = \int_{d-1}^d \frac{(d-t)^l}{l!} \lambda_{t,i}(\boldsymbol{\theta}_i) \, dt$$

and $\mathbf{R}(0)$ is the quantity that we would like to obtain. Using the Itô's lemma and (2.3), we have

$$
\begin{aligned}
R_i(l) &= \frac{a_i}{(l+2)!} + \frac{\lambda_{d-1,i}(\boldsymbol{\theta}_i)}{l!(l+2)} + \sum_{j=1}^{p+r} \zeta_{1,i,j} \int_{d-1}^d \frac{(d-t)^{l+1}}{(l+1)!} \lambda_{t,j}(\boldsymbol{\theta}_j) \, dt \\
&\quad + \sum_{k=2}^h \sum_{j=1}^{p+r} \frac{\zeta_{k,i,j}}{(l+2)!} \int_{d-k}^{d-k+1} \lambda_{t,j}(\boldsymbol{\theta}_j) \, dt + \int_{d-1}^d \frac{(d-t)^{l+2}}{l!(l+2)} z_{i,t}^2 \, dt \\
&\quad + \int_{d-1}^d \frac{(d-t)^{l+2}}{l!(l+2)} J_{\lambda,i}(t) d\widetilde{\Lambda}_{\lambda,i}(t) + 2 \int_{d-1}^d \frac{(d-t)^{l+2}}{l!(l+2)} \int_{d-1}^t z_{i,s} dW_{i,s} z_{i,t} dW_{i,t} \\
&= \frac{a_i}{(l+1)!} + \sum_{k=1}^{h-1} \sum_{j=1}^{p+r} \left(\frac{\zeta_{k,i,j}}{l!(l+2)} + \frac{\zeta_{k+1,i,j}}{(l+2)!} \right) \int_{d-k-1}^{d-k} \lambda_{t,j}(\boldsymbol{\theta}_j) \, dt \\
&\quad + \sum_{j=1}^{p+r} \frac{\zeta_{h,i,j}}{l!(l+2)} \int_{d-h-1}^{d-h} \lambda_{t,j}(\boldsymbol{\theta}_j) \, dt + \int_{d-1}^d \frac{(d-t)^{l+2}}{l!(l+2)} z_{i,t}^2 \, dt \\
&\quad + 2 \int_{d-1}^d \frac{(d-t)^{l+2}}{l!(l+2)} \int_{d-1}^t z_{i,s} dW_{i,s} z_{i,t} dW_{i,t} + \int_{d-1}^d \frac{(d-t)^{l+2}}{l!(l+2)} J_{\lambda,i}(t) d\widetilde{\Lambda}_{\lambda,i}(t) \\
&\quad + \sum_{j=1}^{p+r} \zeta_{1,i,j} R_j(l+1) \text{ a.s.}
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\mathbf{R}(l) &= \frac{\mathbf{a}}{(l+1)!} + \sum_{k=1}^{h-1} \left(\frac{\boldsymbol{\zeta}_k}{l!(l+2)} + \frac{\boldsymbol{\zeta}_{k+1}}{(l+2)!} \right) \int_{d-k-1}^{d-k} \boldsymbol{\lambda}_t(\boldsymbol{\theta}) \, dt + \frac{\boldsymbol{\zeta}_h}{l!(l+2)} \int_{d-h-1}^{d-h} \boldsymbol{\lambda}_t(\boldsymbol{\theta}) \, dt \\
&\quad + \left[\int_{d-1}^d \frac{(d-t)^{l+2}}{l!(l+2)} z_{i,t}^2 \, dt + \int_{d-1}^d \frac{(d-t)^{l+2}}{l!(l+2)} J_{\lambda,i}(t) d\widetilde{\Lambda}_{\lambda,i}(t) \right. \\
&\qquad\qquad\qquad\qquad \left. + 2 \int_{d-1}^d \frac{(d-t)^{l+2}}{l!(l+2)} \int_{d-1}^t z_{i,s} dW_{i,s} z_{i,t} dW_{i,t} \right]^\top_{i=1,\ldots,p+r} \\
&\quad + \boldsymbol{\zeta}_1 \mathbf{R}(l+1) \text{ a.s.,}
\end{aligned}
$$

where $\mathbf{a} = (a_1, \ldots, a_{p+r})^\top$. Define

$$\boldsymbol{\pi}_1 = \sum_{l=0}^\infty \frac{\boldsymbol{\zeta}_1^l}{(l+1)!} = \boldsymbol{\zeta}_1^{-1}\left(e^{\boldsymbol{\zeta}_1} - \mathbf{I}_{p+r}\right) \quad \text{and} \quad \boldsymbol{\pi}_2 = \sum_{l=0}^\infty \frac{\boldsymbol{\zeta}_1^l}{(l+2)!} = \boldsymbol{\zeta}_1^{-2}(e^{\boldsymbol{\zeta}_1} - \mathbf{I}_{p+r} - \boldsymbol{\zeta}_1),$$

where $\mathbf{I}_{p+r}$ is the $(p+r)$-dimensional identity matrix and $e^{\boldsymbol{\zeta}_1} = \sum_{l=0}^\infty \boldsymbol{\zeta}_1^l/l!$. Then, iterativing the above formula, we have

$$
\begin{aligned}
\mathbf{R}(0) &= \int_{d-1}^d \boldsymbol{\lambda}_t\left(\boldsymbol{\theta}\right) dt \\
&= \boldsymbol{\pi}_1\mathbf{a} + \sum_{k=1}^{h-1}\left((\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2)\boldsymbol{\zeta}_k + \boldsymbol{\pi}_2\boldsymbol{\zeta}_{k+1}\right)\int_{d-k-1}^{d-k}\boldsymbol{\lambda}_t\left(\boldsymbol{\theta}\right) dt + (\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2)\boldsymbol{\zeta}_h\int_{d-h-1}^{d-h}\boldsymbol{\lambda}_t\left(\boldsymbol{\theta}\right) dt \\
&\quad + \sum_{l=0}^\infty \boldsymbol{\zeta}_1^l \left[\int_{d-1}^d \frac{(d-t)^{l+2}}{l!\,(l+2)} z_{i,t}^2 dt + \int_{d-1}^d \frac{(d-t)^{l+2}}{l!(l+2)} J_{\lambda,i}(t)d\widetilde{\Lambda}_{\lambda,i}(t)\right. \\
&\qquad\qquad\qquad\qquad \left. + 2\int_{d-1}^d \frac{(d-t)^{l+2}}{l!\,(l+2)}\int_{d-1}^t z_{i,s}dW_{i,s}z_{i,t}dW_{i,t}\right]^\top_{i=1,\ldots,p+r} \\
&= \boldsymbol{\nu} + \sum_{k=1}^h \mathbf{A}_k\boldsymbol{\xi}_{d-k} + \boldsymbol{\epsilon}_d \quad \text{a.s.,}
\end{aligned}
$$

where

$$\mathbf{A}_k = \left((\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2)\boldsymbol{\zeta}_k + \boldsymbol{\pi}_2\boldsymbol{\zeta}_{k+1}\right) \text{ for } 1 \leq k \leq h-1,$$

$$\mathbf{A}_h = (\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2)\boldsymbol{\zeta}_h,$$

$$\boldsymbol{\nu} = \boldsymbol{\pi}_1\mathbf{a} + \sum_{l=0}^\infty \boldsymbol{\zeta}_1^l\left[\int_{d-1}^d\left(\frac{(d-t)^{l+2}}{(l+1)!} - \frac{(d-t)^{l+2}}{(l+2)!}\right)\mathrm{E}\left[z_{i,t}^2\right]dt\right]^\top_{i=1,\ldots,p+r}, \quad \text{and}$$

$$
\begin{aligned}
\boldsymbol{\epsilon}_d &= \sum_{l=0}^\infty \boldsymbol{\zeta}_1^l\left[\int_{d-1}^d\left(\frac{(d-t)^{l+2}}{(l+1)!} - \frac{(d-t)^{l+2}}{(l+2)!}\right)\left(z_{i,t}^2 - \mathrm{E}\left[z_{i,t}^2\right]\right)dt + \int_{d-1}^d\frac{(d-t)^{l+1}}{(l+1)!}J_{\lambda,i}(t)d\widetilde{\Lambda}_{\lambda,i}(t)\right. \\
&\qquad\qquad \left. + 2\int_{d-1}^d\left(\frac{(d-t)^{l+2}}{(l+1)!} - \frac{(d-t)^{l+2}}{(l+2)!}\right)\int_{d-1}^t z_{i,s}dW_{i,s}z_{i,t}dW_{i,t}\right]^\top_{i=1,\ldots,p+r}. \qquad (\text{A.2})
\end{aligned}
$$

∎

A.5   Proof of Theorem 1

We note that the model (2.4) can be written in a VAR(1) form as follows:

$$\widetilde{\boldsymbol{\xi}}_d = \widetilde{\boldsymbol{\nu}} + \widetilde{\mathbf{A}}\widetilde{\boldsymbol{\xi}}_{d-1} + \widetilde{\boldsymbol{\epsilon}}_d,$$

where

$$\widetilde{\boldsymbol{\xi}}_d = \begin{pmatrix} \boldsymbol{\xi}_d \\ \boldsymbol{\xi}_{d-1} \\ \vdots \\ \boldsymbol{\xi}_{d-h+1} \end{pmatrix}_{h(p+r)\times 1}, \quad \widetilde{\boldsymbol{\nu}} = \begin{pmatrix} \boldsymbol{\nu} \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{h(p+r)\times 1},$$

$$\widetilde{\mathbf{A}} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_{h-1} & \mathbf{A}_h \\ \mathbf{I}_{p+r} & 0 & 0 & 0 & 0 \\ 0 & \mathbf{I}_{p+r} & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{I}_{p+r} & 0 \end{pmatrix}_{h(p+r)\times h(p+r)}, \quad \widetilde{\boldsymbol{\epsilon}}_d = \begin{pmatrix} \boldsymbol{\epsilon}_d \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{h(p+r)\times 1}. \qquad (A.3)$$

Lemma 1. *Under the model (2.4) and Assumption 1(a)–(b), suppose that s_β is bounded by some positive constant and $\max_{1\le i\le p+r} \mathrm{E}(|\epsilon_{d,i}|^{c_\epsilon}) < \infty$. Then, we have $\max_{1\le i\le p+r} \mathrm{E}(|\xi_{d,i}|^{c_\epsilon}) < \infty$.*

Proof of Lemma 1. Since $\rho(\widetilde{\mathbf{A}}) < 1$, by Gelfand's formula, we have

$$\lim_{l\to\infty} \left\|\widetilde{\mathbf{A}}^l\right\|_\infty^{1/l} = \rho(\widetilde{\mathbf{A}}) < 1.$$

Thus, there exists a positive integer k such that $\left\|\widetilde{\mathbf{A}}^k\right\|_\infty < 1$. Note that for any fixed matrix $\mathbf{M} \in \mathbb{R}^{p_1\times p_2}$ and multivariate random variable $\mathbf{x} \in \mathbb{R}^{p_2}$, we have

$$\sup_{j\le p_1} \left\|(\mathbf{Mx})_j\right\|_{L_{c_\epsilon}} \le \|\mathbf{M}\|_\infty \sup_{j\le p_1} \left\|(\mathbf{x})_j\right\|_{L_{c_\epsilon}},$$

where for any vector $\mathbf{x}$, $(\mathbf{x})_j$ is the j-th element of $\mathbf{x}$. Hence, by the fact that

$$\widetilde{\boldsymbol{\xi}}_d = \widetilde{\mathbf{A}}^k \widetilde{\boldsymbol{\xi}}_{d-k} + \sum_{i=1}^{k} \widetilde{\mathbf{A}}^{k-i} \left(\widetilde{\boldsymbol{\nu}} + \widetilde{\boldsymbol{\epsilon}}_{d-k+i} \right),$$

we have

$$\sup_{j \leq h(p+r)} \left\| \left(\widetilde{\boldsymbol{\xi}}_d \right)_j \right\|_{L_{c_\epsilon}}$$

$$\leq \left\| \widetilde{\mathbf{A}}^k \right\|_\infty \sup_{j \leq h(p+r)} \left\| \left(\widetilde{\boldsymbol{\xi}}_{d-k} \right)_j \right\|_{L_{c_\epsilon}} + \sum_{i=1}^{k} \left\| \widetilde{\mathbf{A}}^{k-i} \right\|_\infty \sup_{j \leq h(p+r)} \left\| \left(\widetilde{\boldsymbol{\nu}} + \widetilde{\boldsymbol{\epsilon}}_{d-k+i} \right)_j \right\|_{L_{c_\epsilon}}.$$

Then, we have

$$\sup_{j \leq h(p+r)} \left\| \left(\widetilde{\boldsymbol{\xi}}_d \right)_j \right\|_{L_{c_\epsilon}} \leq \frac{\sup_{j \leq h(p+r)} \left\| \left(\widetilde{\boldsymbol{\nu}} + \widetilde{\boldsymbol{\epsilon}}_d \right)_j \right\|_{L_{c_\epsilon}}}{1 - \left\| \widetilde{\mathbf{A}}^k \right\|_\infty} \left(\sum_{i=1}^{k} \left\| \widetilde{\mathbf{A}}^{k-i} \right\|_\infty \right) \leq C, \tag{A.4}$$

where the first inequality is from the stationarity and the last inequality is due to the boundedness of s_β. $\blacksquare$

Proposition 3. *Under the assumptions in Theorem 1, we have for $i \in \{1, \ldots, r\}$, with probability at least $1 - 3(hr + 1)e^{-\delta}$,*

$$\left\| \nabla \mathcal{L}_{\tau,\varpi}^{F,i}(\boldsymbol{\beta}_{i0}) \right\|_\infty \leq \eta_F / 2. \tag{A.5}$$

Also, we have for $i \in \{r+1, \ldots, p+r\}$, with probability at least $1 - 3\left(h\left(p+r\right)+1\right)e^{-\delta}$,

$$\left\| \nabla \mathcal{L}_{\tau,\varpi}^{I,i}(\boldsymbol{\beta}_{i0}) \right\|_\infty \leq \eta_I / 2. \tag{A.6}$$

Proof of Proposition 3. For the simplicity, we assume that $h = 1$ and omit the intercept term $\boldsymbol{\nu}$. Due to the similarity, we only provide the arguments for $\left\| \nabla \mathcal{L}_{\tau,\varpi}^{I,i}(\boldsymbol{\beta}_{i0}) \right\|_\infty$. Note that $\widehat{\boldsymbol{\xi}}_d^I = (\widehat{\xi}_{d,1}, \ldots, \widehat{\xi}_{d,p+r})^\top$. For each $1 \leq j \leq p+r$, we have

$$\left| \nabla_j \mathcal{L}_{\tau,\varpi}^{I,i}(\boldsymbol{\beta}_{i0}) \right| = \left| \frac{\partial \mathcal{L}_{\tau,\varpi}^{I,i}(\boldsymbol{\beta}_{i0})}{\partial \boldsymbol{\beta}_j} \right| \leq (I)_j + (II)_j, \tag{A.7}$$

where

$$(I)_j = \frac{1}{n-1}\left|\sum_{d=2}^{n}\psi_{\tau_I}\left(\xi_{d,i}-\langle\psi_{\varpi_I}(\boldsymbol{\xi}_{d-1}),\boldsymbol{\beta}_{i0}\rangle\right)\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right|,$$

$$(II)_j = \frac{1}{n-1}\left|\sum_{d=2}^{n}\left[\psi_{\tau_I}\left(\widehat{\xi}_{d,i}-\langle\psi_{\varpi_I}(\widehat{\boldsymbol{\xi}}_{d-1}^{I}),\boldsymbol{\beta}_{i0}\rangle\right)\psi_{\varpi_I}\left(\widehat{\xi}_{d-1,j}\right)\right.\right.$$
$$\left.\left.-\psi_{\tau_I}\left(\xi_{d,i}-\langle\psi_{\varpi_I}(\boldsymbol{\xi}_{d-1}),\boldsymbol{\beta}_{i0}\rangle\right)\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right]\right|.$$

We first consider $(I)_j$. Let $\mathbf{y}_d = (y_{d,1},\ldots,y_{d,p+r})^{\top}$, $y_{d,k}$ be the k-th element of $\boldsymbol{\xi}_d-\psi_{\varpi_I}(\boldsymbol{\xi}_d)$ for $k\in S_i$, and $y_{d,k}=0$ for $k\in S_i^c$, where S_i is defined in Assumption 1(g). Also, let $\epsilon'_{d,i}=\epsilon_{d,i}+\langle\mathbf{y}_{d-1},\boldsymbol{\beta}_{i0}\rangle$. Then, we have

$$(I)_j = \frac{1}{n-1}\left|\sum_{d=2}^{n}\psi_{\tau_I}\left(\epsilon'_{d,i}\right)\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right| \leq (I)_j^{(1)}+(I)_j^{(2)},$$

where

$$(I)_j^{(1)} = \frac{1}{n-1}\sum_{d=2}^{n}\left|\mathrm{E}\left\{\psi_{\tau_I}\left(\epsilon'_{d,i}\right)\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right\}\right|,$$

$$(I)_j^{(2)} = \frac{1}{n-1}\left|\sum_{d=2}^{n}\left[\psi_{\tau_I}\left(\epsilon'_{d,i}\right)\psi_{\varpi_I}\left(\xi_{d-1,j}\right)-\mathrm{E}\left\{\psi_{\tau_I}\left(\epsilon'_{d,i}\right)\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right\}\right]\right|.$$

For $(I)_j^{(1)}$, we have

$$\mathrm{E}\left\{\psi_{\tau_I}\left(\epsilon'_{d,i}\right)\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right\} = \mathrm{E}\left\{\langle\mathbf{y}_{d-1},\boldsymbol{\beta}_{i0}\rangle\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right\}-\mathrm{E}\left\{\left[\epsilon'_{d,i}-\psi_{\tau_I}\left(\epsilon'_{d,i}\right)\right]\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right\}.$$

Let $v_3 = \max_{1\leq k\leq p+r}\mathrm{E}(|\epsilon_{d,k}|^3)$, $K_2 = \max_{1\leq k\leq p+r}\mathrm{E}(|\xi_{d,k}|^2)$, and $K_4 = \max_{1\leq k\leq p+r}\mathrm{E}(|\xi_{d,k}|^4)$. Since

$$|y_{d-1,k}| \leq |\xi_{d-1,k}|\,\mathbb{1}\left(|\xi_{d-1,k}|>\varpi_I\right) \leq \varpi_I^{-2}|\xi_{d-1,k}|^3 \quad \text{a.s.}$$

for $k\in\{1,\ldots,p+r\}$ and

$$\left|\epsilon'_{d,i}-\psi_{\tau_I}\left(\epsilon'_{d,i}\right)\right| \leq |\epsilon'_{d,i}|\,\mathbb{1}\left(|\epsilon'_{d,i}|>\tau_I\right) \leq \tau_I^{-2}|\epsilon'_{d,i}|^3 \quad \text{a.s.,}$$

we have

$$
\begin{aligned}
&\left| \mathrm{E}\left\{ \psi_{\tau_I}\left(\epsilon'_{d,i}\right)\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right\}\right| \\
&\leq \left| \mathrm{E}\left\{ \langle \mathbf{y}_{d-1}, \boldsymbol{\beta}_{i0}\rangle \psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right\}\right| + \tau_I^{-2}\mathrm{E}\left\{ \left|\epsilon'_{d,i}\right|^3 \left|\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right|\right\} \\
&\leq \|\boldsymbol{\beta}_{i0}\|_1 K_4 \varpi_I^{-2} + \tau_I^{-2}\mathrm{E}\left\{ \left|\epsilon'_{d,i}\right|^3 \left|\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right|\right\} \\
&\leq \|\boldsymbol{\beta}_{i0}\|_2 K_4 s_\beta^{1/2}\varpi_I^{-2} + 4\tau_I^{-2}\left[\mathrm{E}\left\{ |\epsilon_{d,i}|^3 \left|\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right|\right\} + \|\boldsymbol{\beta}_{i0}\|_2^3 \mathrm{E}\left\{ \|\mathbf{y}_{d-1}\|_2^3 \left|\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right|\right\}\right] \\
&\leq \|\boldsymbol{\beta}_{i0}\|_2 K_4 s_\beta^{1/2}\varpi_I^{-2} + 4\tau_I^{-2}\left\{ v_3 K_2^{1/2} + \|\boldsymbol{\beta}_{i0}\|_2^3 K_4 s_\beta^{3/2}\right\}.
\end{aligned}
\tag{A.8}
$$

Thus, we have

$$
(I)_j^{(1)} \leq C\left(s_\beta \varpi_I^{-2} + s_\beta^3 \tau_I^{-2}\right).
\tag{A.9}
$$

For $(I)_j^{(2)}$, note that the process $(\widetilde{\boldsymbol{\xi}}_d)_{d=1,2,\ldots}$ is geometrically α-mixing and $\widetilde{\boldsymbol{\epsilon}}_d = \widetilde{\boldsymbol{\xi}}_d - \widetilde{\mathbf{A}}\widetilde{\boldsymbol{\xi}}_{d-1}$. Since each $\psi_{\tau_I}\left(\epsilon'_{d,i}\right)\psi_{\varpi_I}\left(\xi_{d-1,j}\right) - \mathrm{E}\left[\psi_{\tau_I}\left(\epsilon'_{d,i}\right)\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right]$ is a measurable function of $\widetilde{\boldsymbol{\xi}}_d$ and $\widetilde{\boldsymbol{\xi}}_{d-1}$, $\left\{ \psi_{\tau_I}\left(\epsilon'_{d,i}\right)\psi_{\varpi_I}\left(\xi_{d-1,j}\right) - \mathrm{E}\left[\psi_{\tau_I}\left(\epsilon'_{d,i}\right)\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right]\right\}$ is also geometrically α-mixing with the coefficients satisfying Assumption 1(d). Therefore, by applying Theorem 2 in Merlevède et al. (2009), we have, for $t \geq 0$,

$$
\Pr\left\{ (I)_j^{(2)} \geq t\right\} \leq \exp\left\{ -\frac{Cn^2 t^2}{V^2 n + \tau_I^2 \varpi_I^2 + t\tau_I \varpi_I n \left(\log n\right)^2}\right\},
\tag{A.10}
$$

where

$$
V^2 = \mathrm{Var}\left[\psi_{\tau_I}\left(\epsilon'_{2,i}\right)\psi_{\varpi_I}\left(\xi_{1,j}\right)\right] + 2\sum_{d=3}^{\infty}\left|\mathrm{Cov}\left[\psi_{\tau_I}\left(\epsilon'_{2,i}\right)\psi_{\varpi_I}\left(\xi_{1,j}\right), \psi_{\tau_I}\left(\epsilon'_{d,i}\right)\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right]\right|.
$$

Since the $\frac{c_\epsilon}{2}$-th moment of $\psi_{\tau_I}\left(\epsilon'_{d,i}\right)\psi_{\varpi_I}\left(\xi_{d-1,j}\right)$ is bounded by $Cs_\beta^{c_\epsilon/2}$, by the inequality (2.2) in Davydov (1968), we have

$$
V^2 \leq Cs_\beta^2 \sum_{k=1}^{\infty}\varphi^{[1-(4/c_\epsilon)]k} \leq Cs_\beta^2,
\tag{A.11}
$$

which implies

$$
\Pr\left\{ (I)_j^{(2)} \leq C\left(\frac{\tau_I \varpi_I \left(\log n\right)^2 \delta + s_\beta \sqrt{n\delta}}{n}\right)\right\} \geq 1 - e^{-\delta}.
\tag{A.12}
$$

50

Combining (A.9) and (A.12), we obtain that with probability at least $1 - (p+r)\, e^{-\delta}$,

$$\max_{1\leq j\leq p+r} (I)_j \leq C \left(s_\beta^3 \tau_I^{-2} + s_\beta \varpi_I^{-2} + \frac{\tau_I \varpi_I (\log n)^2 \delta + s_\beta \sqrt{n\delta}}{n} \right). \tag{A.13}$$

Now, consider $(II)_j$. Note that for any $x, y \in \mathbb{R}$,

$$|\psi_{\varpi_I} (x) - \psi_{\varpi_I} (y)| \leq |x - y|.$$

Hence, by Assumption 1(e), we have

$$\begin{aligned}
&\left| \psi_{\tau_I}\left(\widehat{\xi}_{d,i} - \langle \psi_{\varpi_I}(\widehat{\boldsymbol{\xi}}^I_{d-1}), \boldsymbol{\beta}_{i0}\rangle \right) - \psi_{\tau_I}\left(\xi_{d,i} - \langle \psi_{\varpi_I}(\boldsymbol{\xi}_{d-1}), \boldsymbol{\beta}_{i0}\rangle \right) \right| \\
&\leq \left| \widehat{\xi}_{d,i} - \xi_{d,i} - \langle \psi_{\varpi_I}(\widehat{\boldsymbol{\xi}}^I_{d-1}) - \psi_{\varpi_I}(\boldsymbol{\xi}_{d-1}), \boldsymbol{\beta}_{i0}\rangle \right| \\
&\leq C \left(b^F_{m,n,p} + s_\beta b^I_{m,n,p} \right) \quad \text{a.s.}
\end{aligned}$$

Thus, by using the fact that

$$|x_1 y_1 - x_2 y_2| \leq |(x_1 - x_2)(y_1 - y_2)| + |(x_1 - x_2) y_2| + |x_2 (y_1 - y_2)|$$

for any $x_1, x_2, y_1, y_2 \in \mathbb{R}$, we have

$$\begin{aligned}
(II)_j \ \leq\ & C\Big[\big(b^F_{m,n,p} + s_\beta b^I_{m,n,p} \big) \max\big(b^F_{m,n,p}, b^I_{m,n,p} \big) \\
&+ \frac{b^F_{m,n,p} + s_\beta b^I_{m,n,p}}{n-1} \sum_{d=2}^{n} \{ |\psi_{\varpi_I} (\xi_{d-1,j})| \} \\
&+ \frac{\max\big(b^F_{m,n,p}, b^I_{m,n,p} \big)}{n-1} \sum_{d=2}^{n} \{ |\psi_{\tau_I} (\xi_{d,i} - \langle \psi_{\varpi_I}(\boldsymbol{\xi}_{d-1}), \boldsymbol{\beta}_{i0}\rangle)| \} \Big] \\
\leq\ & C\Big[\big(b^F_{m,n,p} + s_\beta b^I_{m,n,p} \big) \max\big(b^F_{m,n,p}, b^I_{m,n,p} \big) \\
&+ \big(b^F_{m,n,p} + s_\beta b^I_{m,n,p} \big) \big\{ (II)_j^{(1)} + (II)_j^{(2)} \big\} \\
&+ \max\big(b^F_{m,n,p}, b^I_{m,n,p} \big) \big\{ (II)_j^{(3)} + (II)_j^{(4)} \big\} \Big], \tag{A.14}
\end{aligned}$$

where

$$(II)_j^{(1)} = \frac{1}{n-1} \left| \sum_{d=2}^{n} \left[\left| \psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right| - \mathrm{E}\left|\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right| \right] \right|,$$

$$(II)_j^{(2)} = \frac{1}{n-1} \sum_{d=2}^{n} \mathrm{E}\left|\psi_{\varpi_I}\left(\xi_{d-1,j}\right)\right|,$$

$$(II)_j^{(3)} = \frac{1}{n-1} \left| \sum_{d=2}^{n} \left[\left| \psi_{\tau_I}\left(\xi_{d,i} - \langle \psi_{\varpi_I}(\boldsymbol{\xi}_{d-1}), \boldsymbol{\beta}_{i0}\rangle\right)\right| - \mathrm{E}\left|\psi_{\tau_I}\left(\xi_{d,i} - \langle \psi_{\varpi_I}(\boldsymbol{\xi}_{d-1}), \boldsymbol{\beta}_{i0}\rangle\right)\right| \right] \right|,$$

$$(II)_j^{(4)} = \frac{1}{n-1} \sum_{d=2}^{n} \mathrm{E}\left|\psi_{\tau_I}\left(\xi_{d,i} - \langle \psi_{\varpi_I}(\boldsymbol{\xi}_{d-1}), \boldsymbol{\beta}_{i0}\rangle\right)\right|.$$

Consider $(II)_j^{(1)}$ and $(II)_j^{(3)}$. Similar to the proofs of $(I)_j^{(2)}$, we can show

$$\Pr\left\{ (II)_j^{(1)} \leq C \left(\frac{\varpi_I \left(\log n\right)^2 \delta + s_\beta \sqrt{n\delta}}{n} \right) \right\} \geq 1 - e^{-\delta} \tag{A.15}$$

and

$$\Pr\left\{ (II)_j^{(3)} \leq C \left(\frac{\tau_I \left(\log n\right)^2 \delta + s_\beta \sqrt{n\delta}}{n} \right) \right\} \geq 1 - e^{-\delta}. \tag{A.16}$$

Also, we have

$$(II)_j^{(2)} \leq C \text{ and } (II)_j^{(4)} \leq \frac{1}{n-1} \sum_{d=2}^{n} \mathrm{E}\left|\xi_{d,i} - \langle \psi_{\varpi_I}(\boldsymbol{\xi}_{d-1}), \boldsymbol{\beta}_{i0}\rangle\right| \leq C s_\beta. \tag{A.17}$$

By (A.14)–(A.17), we have

$$\Pr\left\{ \max_{1 \leq j \leq p+r} (II)_j \leq C s_\beta \left(b_{m,n,p}^F + b_{m,n,p}^I \right) \right\} \geq 1 - 2 \left(p + r\right) e^{-\delta}. \tag{A.18}$$

Combining (A.7), (A.13), and (A.18), we obtain that with probability at least $1 - 3\left(p+r\right)e^{-\delta}$,

$$\left\| \nabla \mathcal{L}_{\tau,\varpi}^{I,i}(\boldsymbol{\beta}_{i0}) \right\|_\infty \leq C \left[s_\beta \left(b_{m,n,p}^F + b_{m,n,p}^I \right) + s_\beta^3 \tau_I^{-2} \right.$$

$$\left. + s_\beta \varpi_I^{-2} + \frac{\tau_I \varpi_I \left(\log n\right)^2 \delta + s_\beta \sqrt{n\delta}}{n} \right]. \tag{A.19}$$

∎

52

Proof of Theorem 1. By Proposition 3, we prove the statements under (A.5) and (A.6). First, we consider $\left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0}\right\|_2$ for $i \in \{1, \ldots, r\}$. Suppose that

$$\left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0}\right\|_2 > \frac{(hr+1)^{1/2}\,\eta_F}{\kappa}. \tag{A.20}$$

By the optimality of $\widehat{\boldsymbol{\beta}}_i$ and the integral form of the Taylor expansion, we have

$$
\begin{aligned}
0 \;\geq\; & \mathcal{L}_{\tau,\varpi}^{F,i}(\widehat{\boldsymbol{\beta}}_i) - \mathcal{L}_{\tau,\varpi}^{F,i}(\boldsymbol{\beta}_{i0}) \\
=\; & \langle \nabla \mathcal{L}_{\tau,\varpi}^{F,i}(\boldsymbol{\beta}_{i0}), \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0} \rangle \\
& + \int_0^1 (1-t)\,(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0})^\top \nabla^2 \mathcal{L}_{\tau,\varpi}^{F,i}(\boldsymbol{\beta}_{i0} + t(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0}))(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0})\,dt.
\end{aligned} \tag{A.21}
$$

Since $\left\|\nabla \mathcal{L}_{\tau,\varpi}^{F,i}(\boldsymbol{\beta}_{i0})\right\|_\infty \leq \eta_F/2$, we have

$$
\begin{aligned}
\left| \langle \nabla \mathcal{L}_{\tau,\varpi}^{F,i}(\boldsymbol{\beta}_{i0}), \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0} \rangle \right| \;\leq\; & \left\|\nabla \mathcal{L}_{\tau,\varpi}^{F,i}(\boldsymbol{\beta}_{i0})\right\|_\infty \left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0}\right\|_1 \\
\leq\; & \frac{(hr+1)^{1/2}\eta_F}{2}\left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0}\right\|_2 .
\end{aligned} \tag{A.22}
$$

By (A.20), we have

$$ z = \frac{(hr+1)^{1/2}\eta_F}{\kappa \left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0}\right\|_2} < 1. $$

Then, for any $0 \leq t \leq z$, we have

$$ \left\|[\boldsymbol{\beta}_{i0} + t(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0})] - \boldsymbol{\beta}_{i0}\right\|_1 \leq t(hr+1)^{1/2}\left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0}\right\|_2 \leq \frac{(hr+1)\eta_F}{\kappa}. $$

Hence, we have

$$
\begin{aligned}
& \int_0^1 (1-t)\,(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0})^\top \nabla^2 \mathcal{L}_{\tau,\varpi}^{F,i}(\boldsymbol{\beta}_{i0} + t(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0}))(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0})\,dt \\
& \geq \int_0^z (1-t)\kappa \left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0}\right\|_2^2 \, dt \\
& = (hr+1)^{1/2}\eta_F \left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0}\right\|_2 - \frac{(hr+1)\eta_F^2}{2\kappa},
\end{aligned} \tag{A.23}
$$

where the first inequality is due to Assumption 1(g). Combining (A.21)–(A.23), we have

$$0 \geq \frac{(hr+1)^{1/2}\eta_F \left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0}\right\|_2}{2} - \frac{(hr+1)\eta_F^2}{2\kappa},$$

which contradicts to (A.20). Thus, (3.5) is showed.

Now, consider $\left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0}\right\|_2$ for $i \in \{r+1, \ldots, p+r\}$. By Proposition 1 in Fan et al. (2019) and Proposition 3, we can show (3.6). ∎

A.6   Proof of Theorem 3.

Proof of Proposition 2. Similar to the proofs of Theorem 3 in Fan and Kim (2018), we can show (4.6) and (4.7) under the event

$$E = \left\{ \max_{1 \leq d \leq n} \max_{1 \leq i,j \leq p} \left| \widehat{\Gamma}_{d,ij} - \Gamma_{d,ij} \right| \leq C\sqrt{\log\left(pn \vee m\right)/m^{1/2}} \right\}.$$

By Weyl's theorem, (4.6) implies (4.5). Thus, it is enough to show (4.4) under the event E. Without loss of generality, we assume that sign $\left(\langle \widehat{\mathbf{q}}_i^F, \mathbf{q}_i^F \rangle\right) = 1$ for $i = 1, \ldots, r$. We have for each $d \in \{1, \ldots, n\}$ and $i \in \{1, \ldots, r\}$,

$$\begin{aligned}
&\left| \widehat{\xi}_{d,i} - \xi_{d,i} \right| \\
&\leq p^{-1} \left| \left(\mathbf{q}_i^F\right)^\top \left(\widehat{\boldsymbol{\Gamma}}_d - \boldsymbol{\Gamma}_d\right) \mathbf{q}_i^F \right| + p^{-1} \left| \left(\widehat{\mathbf{q}}_i^F\right)^\top \widehat{\boldsymbol{\Gamma}}_d \widehat{\mathbf{q}}_i^F - \left(\mathbf{q}_i^F\right)^\top \widehat{\boldsymbol{\Gamma}}_d \mathbf{q}_i^F \right| + p^{-1} \left| \left(\mathbf{q}_i^F\right)^\top \boldsymbol{\Gamma}_d \mathbf{q}_i^F - \xi_{d,i} \right| \\
&= (I) + (II) + (III).
\end{aligned} \tag{A.24}$$

For (I), we have

$$(I) \leq p^{-1} \left\|\widehat{\boldsymbol{\Gamma}}_d - \boldsymbol{\Gamma}_d\right\|_F \leq C\sqrt{\log\left(pn \vee m\right)/m^{1/2}}. \tag{A.25}$$

For (II), we have

$$\begin{aligned}
\left\|\mathbf{q}_i^F - \widehat{\mathbf{q}}_i^F\right\|_2 &\leq Cp^{-1} \left\| \sum_{k=n-\ell+1}^{n} \left(\widehat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Psi}_k\right)/\ell \right\|_2 \\
&\leq Cp^{-1}\ell^{-1} \sum_{k=n-\ell+1}^{n} \left(\left\|\widehat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}_k\right\|_F + \|\boldsymbol{\Sigma}_k\|_1 \right)
\end{aligned}$$

$$\leq C \left(\sqrt{\log\left(pn \vee m\right)/m^{1/2}} + p^{-1} \max_{1 \leq i \leq p} \sum_{j=1}^{p} |\Sigma_{d,ij}|^{\Upsilon} (\Sigma_{d,ii}\Sigma_{d,jj})^{(1-\Upsilon)/2} \right)$$

$$\leq C \left(\sqrt{\log\left(pn \vee m\right)/m^{1/2}} + s_I/p \right),$$

where the first inequality is by Theorem 2 in Yu et al. (2015). Hence, we have

(II)

$$\leq p^{-1} \left| \left(\mathbf{q}_i^F - \widehat{\mathbf{q}}_i^F\right)^\top \widehat{\boldsymbol{\Gamma}}_d \left(\mathbf{q}_i^F - \widehat{\mathbf{q}}_i^F\right) \right| + p^{-1} \left| \left(\mathbf{q}_i^F - \widehat{\mathbf{q}}_i^F\right)^\top \widehat{\boldsymbol{\Gamma}}_d \mathbf{q}_i^F \right| + p^{-1} \left| \left(\mathbf{q}_i^F\right)^\top \widehat{\boldsymbol{\Gamma}}_d \left(\mathbf{q}_i^F - \widehat{\mathbf{q}}_i^F\right) \right|$$

$$\leq p^{-1} \|\widehat{\boldsymbol{\Gamma}}_d\|_F \|\mathbf{q}_i^F - \widehat{\mathbf{q}}_i^F\|_2^2 + 2p^{-1} \|\widehat{\boldsymbol{\Gamma}}_d\|_F \|\mathbf{q}_i^F - \widehat{\mathbf{q}}_i^F\|_2$$

$$\leq C\sqrt{\log\left(pn \vee m\right)/m^{1/2}}. \tag{A.26}$$

For (III), we have

$$(III) = p^{-1} \left(\mathbf{q}_i^F\right)^\top \boldsymbol{\Sigma}_d \mathbf{q}_i^F \leq p^{-1} \|\boldsymbol{\Sigma}_d\|_2 \leq p^{-1} \|\boldsymbol{\Sigma}_d\|_\infty \leq p^{-1} \max_{1 \leq i \leq p} \sum_{j=1}^{p} |\Sigma_{d,ij}|^{\Upsilon} (\Sigma_{d,ii}\Sigma_{d,jj})^{(1-\Upsilon)/2}$$

$$\leq Cs_I/p. \tag{A.27}$$

Combining (A.24)–(A.27), we have

$$\left| \widehat{\xi}_{d,i} - \xi_{d,i} \right| \leq C \left(\sqrt{\log\left(pn \vee m\right)/m^{1/2}} + s_I/p \right),$$

which completes the proof. ∎

Proof of Theorem 3. We show the statements (4.10)–(4.12) under (A.5)–(A.6) and (4.4)–(4.9). For simplicity, we assume that $h = 1$ and omit the intercept term $\boldsymbol{\nu}$. Note that $\widehat{\boldsymbol{\beta}}_i = \left(\widehat{A}_{1,i1}, \ldots, \widehat{A}_{1,ir}\right)^\top$ for $1 \leq i \leq r$ and $\widehat{\boldsymbol{\beta}}_i = \left(\widehat{A}_{1,i1}, \ldots, \widehat{A}_{1,i(p+r)}\right)^\top$ for $r+1 \leq i \leq p+r$. First, we consider (4.10). By (4.4) and (4.8), we have for any $1 \leq i \leq r$,

$$\left| \widehat{\xi}_{n+1,i} - \mathrm{E}\left(\xi_{n+1,i}|\mathcal{F}_n\right) \right| = \left| \sum_{j=1}^{r} \left(\widehat{A}_{1,ij}\widehat{\xi}_{n,j} - A_{1,ij}\xi_{n,j} \right) \right|$$

$$\leq \sum_{j=1}^{r} \left[\left| \left(\widehat{A}_{1,ij} - A_{1,ij}\right) \left(\widehat{\xi}_{n,j} - \xi_{n,j}\right) \right| \right.$$

$$\left. + \left| A_{1,ij} \left(\widehat{\xi}_{n,j} - \xi_{n,j}\right) \right| + \left| \left(\widehat{A}_{1,ij} - A_{1,ij}\right) \xi_{n,j} \right| \right]$$

$$\leq C \left[H_m + (\log n)^2 \sqrt{\log p/n} \right].$$

Consider (4.11). Similar to the proofs of Proposition 1 in Fan et al. (2019), we can show $\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0} \in \mathcal{W}_i$ for any $r+1 \leq i \leq p+r$, where $\mathcal{W}_i$ is defined in Assumption 1(g). Thus, we have

$$\max_{r+1 \leq i \leq p+r} \sum_{j=1}^{p+r} \left| \widehat{A}_{1,ij} - A_{1,ij} \right| \leq \max_{r+1 \leq i \leq p+r} 4 \sum_{j \in S_i} \left| \widehat{A}_{1,ij} - A_{1,ij} \right|$$
$$\leq C \left\{ s_\beta^2 s_I H_m^{1-\Upsilon} + s_\beta^2 (\log n)^2 \sqrt{\log p/n} \right\},$$

where the last inequality is due to the Cauchy–Schwarz inequality and (4.9). Then, by (4.5), we have, for any $r+1 \leq i \leq p+r$,

$$\left| \widehat{\xi}_{n+1,i} - \mathrm{E}\left(\xi_{n+1,i}|\mathcal{F}_n\right) \right| = \left| \sum_{j=1}^{p+r} \left(\widehat{A}_{1,ij} \widehat{\xi}_{n,j} - A_{1,ij} \xi_{n,j} \right) \right|$$
$$\leq \sum_{j=1}^{p+r} \left[\left| \left(\widehat{A}_{1,ij} - A_{1,ij} \right) \left(\widehat{\xi}_{n,j} - \xi_{n,j} \right) \right| \right.$$
$$\left. + \left| A_{1,ij} \left(\widehat{\xi}_{n,j} - \xi_{n,j} \right) \right| + \left| \left(\widehat{A}_{1,ij} - A_{1,ij} \right) \xi_{n,j} \right| \right]$$
$$\leq C \left[s_\beta^2 s_I H_m^{1-\Upsilon} + s_\beta^2 (\log n)^2 \sqrt{\log p/n} \right].$$

For (4.12), we have

$$\left\| \widetilde{\boldsymbol{\Gamma}}_{n+1} - \mathrm{E}\left(\boldsymbol{\Gamma}_{n+1}|\mathcal{F}_n\right) \right\|_{\boldsymbol{\Gamma}^*} \leq \left\| \widehat{\boldsymbol{\Psi}}_{n+1} - \mathrm{E}\left(\boldsymbol{\Psi}_{n+1}|\mathcal{F}_n\right) \right\|_{\boldsymbol{\Gamma}^*}$$
$$+ \left\| \widehat{\boldsymbol{\Sigma}}_{n+1} - \mathrm{E}\left(\boldsymbol{\Sigma}_{n+1}|\mathcal{F}_n\right) \right\|_{\boldsymbol{\Gamma}^*}$$
$$= (I) + (II). \tag{A.28}$$

Consider (I). We have

$$\left\| \widehat{\boldsymbol{\Psi}}_{n+1} - \mathrm{E}\left(\boldsymbol{\Psi}_{n+1}|\mathcal{F}_n\right) \right\|_F = p \left\| \sum_{i=1}^{r} \left\{ \widehat{\xi}_{n+1,i} \widehat{\mathbf{q}}_i^F \left(\widehat{\mathbf{q}}_i^F \right)^\top - \mathrm{E}\left(\xi_{n+1,i}|\mathcal{F}_n\right) \mathbf{q}_i^F \left(\mathbf{q}_i^F \right)^\top \right\} \right\|_F$$
$$\leq p \left[\left\| \sum_{i=1}^{r} \left\{ \widehat{\xi}_{n+1,i} - \mathrm{E}\left(\xi_{n+1,i}|\mathcal{F}_n\right) \right\} \widehat{\mathbf{q}}_i^F \left(\widehat{\mathbf{q}}_i^F \right)^\top \right\|_F \right.$$
$$\left. + \left\| \sum_{i=1}^{r} \mathrm{E}\left(\xi_{n+1,i}|\mathcal{F}_n\right) \left\{ \widehat{\mathbf{q}}_i^F \left(\widehat{\mathbf{q}}_i^F \right)^\top - \mathbf{q}_i^F \left(\mathbf{q}_i^F \right)^\top \right\} \right\|_F \right]$$

56

$$
\begin{aligned}
&\leq\; p\left[\sum_{i=1}^{r}\left|\widehat{\xi}_{n+1,i}-\mathrm{E}\left(\xi_{n+1,i}|\mathcal{F}_n\right)\right|\left\|\widehat{\mathbf{q}}_i^F\left(\widehat{\mathbf{q}}_i^F\right)^\top\right\|_F\right.\\
&\qquad\left.+\sum_{i=1}^{r}\mathrm{E}\left(\xi_{n+1,i}|\mathcal{F}_n\right)\left\|\left\{\widehat{\mathbf{q}}_i^F\left(\widehat{\mathbf{q}}_i^F\right)^\top-\mathbf{q}_i^F\left(\mathbf{q}_i^F\right)^\top\right\}\right\|_F\right]\\
&\leq\; Cp\left[H_m+(\log n)^2\sqrt{\log p/n}+\sum_{i=1}^{r}\left\|\widehat{\mathbf{q}}_i^F\left(\widehat{\mathbf{q}}_i^F\right)^\top-\mathbf{q}_i^F\left(\mathbf{q}_i^F\right)^\top\right\|_F\right],
\end{aligned}
$$

where the last inequality is due to (4.10). For the last term, by Theorem 2 in Yu et al. (2015), we have

$$
\begin{aligned}
\left\|\widehat{\mathbf{q}}_i^F\left(\widehat{\mathbf{q}}_i^F\right)^\top-\mathbf{q}_i^F\left(\mathbf{q}_i^F\right)^\top\right\|_F &\leq\; Cp^{-1}\left\|\sum_{d=n-\ell+1}^{n}\left(\widehat{\mathbf{\Gamma}}_d-\mathbf{\Psi}_d\right)/\ell\right\|_2\\
&\leq\; Cp^{-1}\ell^{-1}\sum_{d=n-\ell+1}^{n}\left(\left\|\widehat{\mathbf{\Gamma}}_d-\mathbf{\Gamma}_d\right\|_F+\|\mathbf{\Sigma}_d\|_1\right)\\
&\leq\; C\left(\sqrt{\log\left(pn\vee m\right)/m^{1/2}}+s_I/p\right).
\end{aligned}
$$

Thus, we have

$$
\left\|\widehat{\mathbf{\Psi}}_{n+1}-\mathrm{E}\left(\mathbf{\Psi}_{n+1}|\mathcal{F}_n\right)\right\|_F\leq Cp\left[H_m+(\log n)^2\sqrt{\log p/n}\right].
$$

Then, similar to the proofs of Theorem 4.1 in Fan and Kim (2018), we can show

$$
\begin{aligned}
(I)\;&\leq\; C\left[\left\{p^{-3/2}+p^{-1}\widehat{\xi}_{n+1,1}\right\}\left\|\widehat{\mathbf{\Psi}}_{n+1}-\mathrm{E}\left(\mathbf{\Psi}_{n+1}|\mathcal{F}_n\right)\right\|_F\right.\\
&\qquad\left.+p^{-3/2}\widehat{\xi}_{n+1,1}\left\|\widehat{\mathbf{\Psi}}_{n+1}-\mathrm{E}\left(\mathbf{\Psi}_{n+1}|\mathcal{F}_n\right)\right\|_F^2\right]\\
&\leq\; C\left[H_m+p^{1/2}H_m^2+(\log n)^2\sqrt{\log p/n}+p^{1/2}\log p\,(\log n)^4/n\right]. \qquad (A.29)
\end{aligned}
$$

Consider (II). We have

$$
\begin{aligned}
(II)^2\;&\leq\; p^{-1}\left\|\left\{\widehat{\mathbf{\Sigma}}_{n+1}-\mathrm{E}\left(\mathbf{\Sigma}_{n+1}|\mathcal{F}_n\right)\right\}\mathbf{\Gamma}^{-1}\right\|_F^2\\
&\leq\; p^{-1}\left\|\widehat{\mathbf{\Sigma}}_{n+1}-\mathrm{E}\left(\mathbf{\Sigma}_{n+1}|\mathcal{F}_n\right)\right\|_F^2\left\|\mathbf{\Gamma}^{-1}\right\|_2^2\\
&\leq\; p^{-1}\left\|\widehat{\mathbf{\Sigma}}_{n+1}-\mathrm{E}\left(\mathbf{\Sigma}_{n+1}|\mathcal{F}_n\right)\right\|_F^2\\
&=\; p^{-1}\left\|\sum_{i=1}^{p}\left\{\widehat{\xi}_{n+1,i+r}\widehat{\mathbf{q}}_i^I\left(\widehat{\mathbf{q}}_i^I\right)^\top-\mathrm{E}\left(\xi_{n+1,i+r}|\mathcal{F}_n\right)\mathbf{q}_i^I\left(\mathbf{q}_i^I\right)^\top\right\}\right\|_F^2
\end{aligned}
$$

$$\leq \ Cp^{-1}\left[\left\|\sum_{i=1}^{p}\left\{\widehat{\xi}_{n+1,i+r}-\mathrm{E}\left(\xi_{n+1,i+r}|\mathcal{F}_n\right)\right\}\widehat{\mathbf{q}}_i^I\left(\widehat{\mathbf{q}}_i^I\right)^{\top}\right\|_F^2\right.$$

$$\left.+\left\|\sum_{i=1}^{p}\mathrm{E}\left(\xi_{n+1,i+r}|\mathcal{F}_n\right)\left\{\widehat{\mathbf{q}}_i^I\left(\widehat{\mathbf{q}}_i^I\right)^{\top}-\mathbf{q}_i^I\left(\mathbf{q}_i^I\right)^{\top}\right\}\right\|_F^2\right]$$

$$= \ (III)+(IV). \tag{A.30}$$

For (III), we have

$$(III)$$
$$=Cp^{-1}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{tr}\left(\{\widehat{\xi}_{n+1,i+r}-\mathrm{E}\left(\xi_{n+1,i+r}|\mathcal{F}_n\right)\}\{\widehat{\xi}_{n+1,j+r}-\mathrm{E}\left(\xi_{n+1,j+r}|\mathcal{F}_n\right)\}\widehat{\mathbf{q}}_i^I\left(\widehat{\mathbf{q}}_i^I\right)^{\top}\widehat{\mathbf{q}}_j^I\left(\widehat{\mathbf{q}}_j^I\right)^{\top}\right)$$
$$=Cp^{-1}\sum_{i=1}^{p}\mathrm{tr}\left(\{\widehat{\xi}_{n+1,i+r}-\mathrm{E}\left(\xi_{n+1,i+r}|\mathcal{F}_n\right)\}^2\widehat{\mathbf{q}}_i^I\left(\widehat{\mathbf{q}}_i^I\right)^{\top}\right)$$
$$=Cp^{-1}\sum_{i=1}^{p}\{\widehat{\xi}_{n+1,i+r}-\mathrm{E}\left(\xi_{n+1,i+r}|\mathcal{F}_n\right)\}^2$$
$$\leq C\left(s_\beta^2 s_I H_m^{1-\Upsilon}+s_\beta^2\left(\log n\right)^2\sqrt{\log p/n}\right)^2, \tag{A.31}$$

where the last inequality is due to (4.11). For (IV), we have

$$(IV)$$
$$=Cp^{-1}\left\|\sum_{i=1}^{p}\{\mathrm{E}\left(\xi_{n+1,i+r}|\mathcal{F}_n\right)-\mathrm{E}\left(\xi_{n+1,p+r}|\mathcal{F}_n\right)\}\left\{\widehat{\mathbf{q}}_i^I\left(\widehat{\mathbf{q}}_i^I\right)^{\top}-\mathbf{q}_i^I\left(\mathbf{q}_i^I\right)^{\top}\right\}\right\|_F^2$$
$$\leq Cp^{-1}\sum_{i=1}^{p}\mathrm{tr}\left[\{\mathrm{E}\left(\xi_{n+1,i+r}|\mathcal{F}_n\right)-\mathrm{E}\left(\xi_{n+1,p+r}|\mathcal{F}_n\right)\}^2\left\{\widehat{\mathbf{q}}_i^I\left(\widehat{\mathbf{q}}_i^I\right)^{\top}-\mathbf{q}_i^I\left(\mathbf{q}_i^I\right)^{\top}\right\}^2\right]$$
$$+Cp^{-1}\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\mathrm{tr}\left[\{\mathrm{E}\left(\xi_{n+1,i+r}|\mathcal{F}_n\right)-\mathrm{E}\left(\xi_{n+1,p+r}|\mathcal{F}_n\right)\}\{\mathrm{E}\left(\xi_{n+1,j+r}|\mathcal{F}_n\right)-\mathrm{E}\left(\xi_{n+1,p+r}|\mathcal{F}_n\right)\}\right.$$
$$\left.\times\left\{\widehat{\mathbf{q}}_i^I\left(\widehat{\mathbf{q}}_i^I\right)^{\top}-\mathbf{q}_i^I\left(\mathbf{q}_i^I\right)^{\top}\right\}\left\{\widehat{\mathbf{q}}_j^I\left(\widehat{\mathbf{q}}_j^I\right)^{\top}-\mathbf{q}_j^I\left(\mathbf{q}_j^I\right)^{\top}\right\}\right]$$
$$\leq Cp^{-1}\sum_{i=1}^{p}\{\mathrm{E}\left(\xi_{n+1,i+r}|\mathcal{F}_n\right)-\mathrm{E}\left(\xi_{n+1,p+r}|\mathcal{F}_n\right)\}^2\left\|\widehat{\mathbf{q}}_i^I\left(\widehat{\mathbf{q}}_i^I\right)^{\top}-\mathbf{q}_i^I\left(\mathbf{q}_i^I\right)^{\top}\right\|_F^2,$$

where the first equality is due to the fact that $\sum_{i=1}^{p}\widehat{\mathbf{q}}_i^I\left(\widehat{\mathbf{q}}_i^I\right)^{\top}=\sum_{i=1}^{p}\mathbf{q}_i^I\left(\mathbf{q}_i^I\right)^{\top}$ and the last inequality is from the positiveness of $\mathrm{E}\left(\xi_{n+1,i+r}|\mathcal{F}_n\right)-\mathrm{E}\left(\xi_{n+1,p+r}|\mathcal{F}_n\right)$ for $1\leq i\leq p-1$ and (A.32)

below. For $i \neq j$, we have

$$\text{tr}\left[\left\{\widehat{\mathbf{q}}_i^I (\widehat{\mathbf{q}}_i^I)^\top - \mathbf{q}_i^I (\mathbf{q}_i^I)^\top\right\}\left\{\widehat{\mathbf{q}}_j^I (\widehat{\mathbf{q}}_j^I)^\top - \mathbf{q}_j^I (\mathbf{q}_j^I)^\top\right\}\right]$$

$$= -\text{tr}\left[\widehat{\mathbf{q}}_i^I (\widehat{\mathbf{q}}_i^I)^\top \mathbf{q}_j^I (\mathbf{q}_j^I)^\top + \mathbf{q}_i^I (\mathbf{q}_i^I)^\top \widehat{\mathbf{q}}_j^I (\widehat{\mathbf{q}}_j^I)^\top\right]$$

$$= -\text{tr}\left[(\widehat{\mathbf{q}}_i^I)^\top \mathbf{q}_j^I (\mathbf{q}_j^I)^\top \widehat{\mathbf{q}}_i^I + (\mathbf{q}_i^I)^\top \widehat{\mathbf{q}}_j^I (\widehat{\mathbf{q}}_j^I)^\top \mathbf{q}_i^I\right]$$

$$= -\text{tr}\left[\left\{(\widehat{\mathbf{q}}_i^I)^\top \mathbf{q}_j^I\right\}^2 + \left\{(\mathbf{q}_i^I)^\top \widehat{\mathbf{q}}_j^I\right\}^2\right] \leq 0. \tag{A.32}$$

Thus, we have

$$
\begin{aligned}
(IV) &\leq Cp^{-1}\sum_{i=1}^{p}\chi^{2(i+r)}\left\|\widehat{\mathbf{q}}_i^I (\widehat{\mathbf{q}}_i^I)^\top - \mathbf{q}_i^I (\mathbf{q}_i^I)^\top\right\|_F^2 \\
&\leq Cp^{-1}\sum_{i=1}^{p}\chi^{2(i+r)}\left\|\sum_{d=n-\ell+1}^{n}\left(\widehat{\mathbf{\Sigma}}_d - \mathbf{\Sigma}_d\right)/\ell\right\|_2^2 / \chi^{2(i+r)} \\
&\leq Cs_I^2 H_m^{2-2\Upsilon},
\end{aligned}
\tag{A.33}
$$

where the first inequality is due to Assumption 3, the second inequality is from Theorem 2 in Yu et al. (2015), and the last inequality is due to (4.6). By (A.30), (A.31), and (A.33), we have

$$(II) \leq C\left(s_\beta^2 s_I H_m^{1-\Upsilon} + s_\beta^2 (\log n)^2 \sqrt{\log p/n}\right). \tag{A.34}$$

Combining (A.28), (A.29), and (A.34), we have

$$
\begin{aligned}
\|\widetilde{\mathbf{\Gamma}}_{n+1} - \text{E}\left(\mathbf{\Gamma}_{n+1}|\mathcal{F}_n\right)\|_{\mathbf{\Gamma}^*} &\leq C\Big[p^{1/2}H_m^2 + p^{1/2}\log p (\log n)^4 /n \\
&\quad + s_\beta^2 s_I H_m^{1-\Upsilon} + s_\beta^2 (\log n)^2 \sqrt{\log p/n}\Big].
\end{aligned}
\tag{A.35}
$$

∎

A.7 Miscellaneous materials

Algorithm 1 Parameter estimation procedure

Step 1 Decompose the input volatility matrix:

$$\widehat{\mathbf{\Gamma}}_d = \sum_{k=1}^{p} \bar{\xi}_{d,k} \bar{\mathbf{q}}_{d,k} \bar{\mathbf{q}}_{d,k}^\top,$$

where $\bar{\xi}_{d,k}$ is the k-th largest eigenvalue of $\widehat{\mathbf{\Gamma}}_d$ and $\bar{\mathbf{q}}_{d,k}$ is its corresponding eigenvector.

Step 2 (factor components) Calculate r eigenvectors, $\widehat{\mathbf{q}}_1^F, \ldots, \widehat{\mathbf{q}}_r^F$, of $\frac{1}{\ell} \sum_{d=n-l+1}^{n} \widehat{\mathbf{\Gamma}}_d$ and obtain the eigenvalues $\widehat{\xi}_{d,i} = \left(\widehat{\mathbf{q}}_i^F\right)^\top \widehat{\mathbf{\Gamma}}_d \widehat{\mathbf{q}}_i^F / p$ for $d = 1, \ldots, n$ and $i = 1, \ldots, r$.

Step 3 Obtain the input idiosyncratic volatility matrix estimator:

$$\bar{\mathbf{\Sigma}}_d = (\bar{\Sigma}_{d,ij})_{1 \leq i,j \leq p} = \widehat{\mathbf{\Gamma}}_d - \sum_{k=1}^{r} \bar{\xi}_{d,k} \bar{\mathbf{q}}_{d,k} \bar{\mathbf{q}}_{d,k}^\top.$$

Step 4 Threshold the input idiosyncratic volatility matrix estimator:

$$\widehat{\Sigma}_{d,ij} = \begin{cases} \bar{\Sigma}_{d,ij} \vee 0 & \text{if } i = j \\ g_{ij}(\bar{\Sigma}_{d,ij}) \mathbf{1}(|\bar{\Sigma}_{d,ij}| \geq \upsilon_{ij}) & \text{if } i \neq j \end{cases} \quad \text{and} \quad \widehat{\mathbf{\Sigma}}_d = (\widehat{\Sigma}_{d,ij})_{1 \leq i,j \leq p},$$

where $g_{ij}(\cdot)$ satisfies $|g_{ij}(x) - x| \leq \upsilon_{ij}$, and $\upsilon_{ij} = \upsilon_m \sqrt{(\bar{\Sigma}_{d,ii} \vee 0)(\bar{\Sigma}_{d,jj} \vee 0)}$.

Step 5 (idiosyncratic components) Calculate p eigenvectors, $\widehat{\mathbf{q}}_1^I, \ldots, \widehat{\mathbf{q}}_p^I$, of $\frac{1}{\ell} \sum_{d=n-l+1}^{n} \widehat{\mathbf{\Sigma}}_d$ and obtain $\widehat{\xi}_{d,i+r} = \left(\widehat{\mathbf{q}}_i^I\right)^\top \widehat{\mathbf{\Sigma}}_d \widehat{\mathbf{q}}_i^I$ for $d = 1, \ldots, n$ and $i = 1, \ldots, p$.

Step 6 Estimate the factor coefficient:

$$\widehat{\boldsymbol{\beta}}_i = \arg \min_{\boldsymbol{\beta}_i \in \mathbb{R}^{hr+1}} \mathcal{L}_{\tau,\varpi}^{F,i}(\boldsymbol{\beta}_i) \quad \text{for } i = 1, \ldots, r,$$

where $\mathcal{L}_{\tau,\varpi}^{F,i}(\boldsymbol{\beta}_i)$ is defined in (3.4).

Step 7 Estimate the idiosyncratic coefficient:

$$\widehat{\boldsymbol{\beta}}_i = \arg \min_{\boldsymbol{\beta}_i \in \mathbb{R}^{h(p+r)+1}} \mathcal{L}_{\tau,\varpi}^{I,i}(\boldsymbol{\beta}_i) + \eta_I \|\boldsymbol{\beta}_i\|_1 \quad \text{for } i = r+1, \ldots, p+r,$$

where $\eta_I > 0$ is the regularization parameter, and $\mathcal{L}_{\tau,\varpi}^{I,i}(\boldsymbol{\beta}_i)$ is defined in (3.2).

Step 8 Calculate the future eigenvalue and conditional expected volatility matrix:

$$\widehat{\boldsymbol{\xi}}_{n+1} = \left(\widehat{\xi}_{n+1,1}, \ldots, \widehat{\xi}_{n+1,p+r}\right)^\top = \widehat{\boldsymbol{\nu}} + \sum_{k=1}^{h} \widehat{\mathbf{A}}_k \widehat{\boldsymbol{\xi}}_{n+1-k} \quad \text{and}$$

$$\widetilde{\mathbf{\Gamma}}_{n+1} = \widehat{\mathbf{\Psi}}_{n+1} + \widehat{\mathbf{\Sigma}}_{n+1} = p \sum_{i=1}^{r} \widehat{\xi}_{n+1,i} \widehat{\mathbf{q}}_i^F \left(\widehat{\mathbf{q}}_i^F\right)^\top + \sum_{i=1}^{p} \widehat{\xi}_{n+1,i+r} \widehat{\mathbf{q}}_i^I \left(\widehat{\mathbf{q}}_i^I\right)^\top.$$
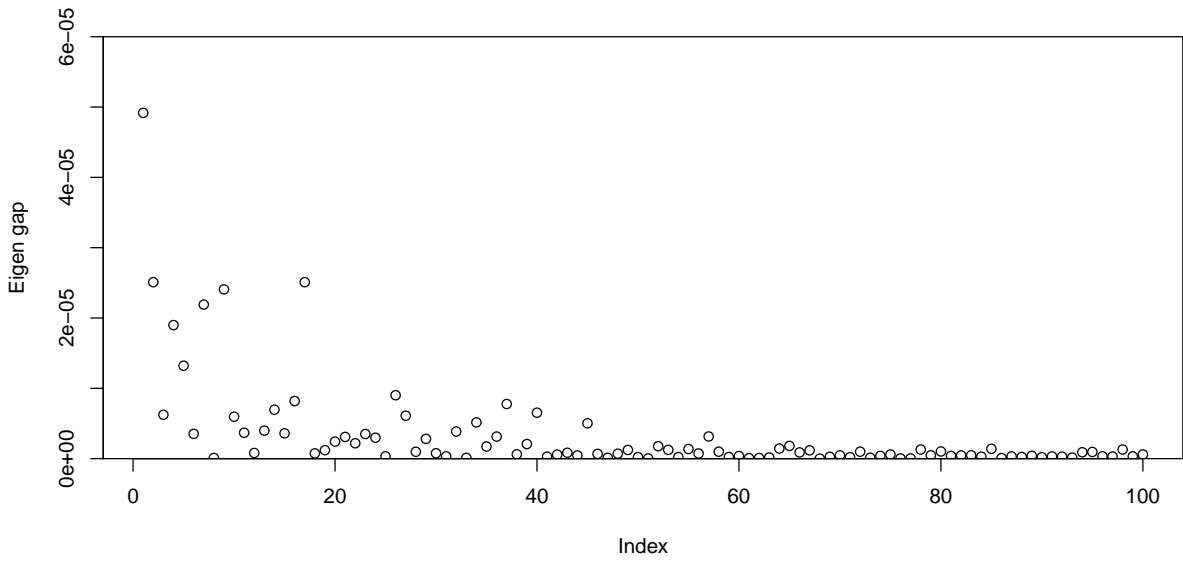

Figure 7: The plot of the first 100 differences between the consecutive eigenvalues of the average of 997 idiosyncratic volatility matrix estimators. We estimated the rank r based on the rank estimation procedure in (5.1) with $n = 997$. The result is $\hat{r} = 4$. We used 1-min log-returns of the top 200 large trading volume stocks among the S&P 500 from January 2016 to December 2019.