Statistica Sinica Preprint No: SS-2024-0308

Title	Heterogeneous Autoregressive Modeling with Flexible Cascade Structures
Manuscript ID	SS-2024-0308
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0308
Complete List of Authors	Huiling Yuan, Guodong Li, Kexin Lu, Alan T.K. Wan and Yong Zhou
Corresponding Authors	Guodong Li
E-mails	gdli@hku.hk
Notice: Accepted author version.	

Heterogeneous Autoregressive Modeling with Flexible Cascade Structures

Huiling Yuan[1], Kexin Lu[2], Guodong Li[2], Alan T.K. Wan[3], and Yong Zhou[1]

[1] *East China Normal University,* [2] *University of Hong Kong, and* [3] *City University of Hong Kong*

Abstract: Advancements in technology have led to increasingly complex structures in high-frequency data, necessitating the development of efficient models for accurately forecasting realized measures. This paper introduces a novel approach known as the multilinear low-rank heterogeneous autoregressive (MLRHAR) model. Distinguishing itself from the conventional heterogeneous autoregressive (HAR) model, our model utilizes a data-driven method to replace the fixed heterogenous volatility components. To address the calendar effect, we utilize the fourth-order tensor technique, which simultaneously reduces dimensions in the response, predictor, and short-term and calendar temporal directions. This not only reduces the parameter space but also enables the automatic selection of heterogeneous components from both temporal directions. Moreover, we establish the non-asymptotic properties of high-dimensional HAR models, and a projected gradient descent algorithm is proposed with theoretical justifications for parameter estimation. Through simulation experiments, we evaluate the efficiency of the proposed model. We apply our method to financial data on the constituent stocks of the S&P 500 Index. The results obtained from both the simulation and real studies convincingly demonstrate the significant forecasting advantages offered by our approach.

Key words and phrases: Calendar effect, Heterogeneous autoregressive model, High-dimensional analysis, High-frequency data, Non-asymptotic property, Tensor technique.

1. Introduction

Volatility analysis has been extensively explored in financial econometrics and statistics. Traditionally, the focus has primarily been on daily data for modeling volatility. The autoregressive conditional heteroskedasticity (ARCH) model introduced by Engle (1982) and its generalization, the generalized autoregressive conditional heteroskedasticity (GARCH) models developed by Bollerslev (1986), have gained significant recognition for analyzing data at this frequency level. However, there are limitations in capturing rapid changes in volatility using solely daily or lower frequency data. Andersen et al. (2003) highlighted the need for timely adjustment to adapt to new volatility levels. Advancements in technology have made it possible to collect high-frequency data, such as tick-by-tick asset prices, which provide more detailed financial information. With the increasing availability of high-frequency data, the focus has shifted towards modeling realized volatilities, constructed from intraday observations at higher frequencies. Several models have been developed to tackle this challenge, including the realized GARCH model (Hansen et al., 2012), the high-frequency-based volatility model (Shephard and Sheppard, 2010), the heterogeneous autoregressive (HAR) model (Corsi, 2009), the multiplicative error model (Engle and Gallo, 2006), and the mixed data sampling model (Ghysels et al., 2006). Among these models, the HAR model has gained popularity due to its straightforward cascade structure (Corsi, 2009), incorporating volatility components in three intervals: daily, weekly, and monthly. Despite its simple autoregressive structure, the HAR model often exhibits remarkable accuracy in forecasting performance.

Since its introduction, the HAR model has received considerable attention in the literature, with various techniques developed to further enhance its forecast accuracy. These techniques typically involve applying logarithmic transformations (Corsi, 2009) or Box-Cox transformations (Taylor, 2017) to realized volatilities prior to estimation. Additionally, estimation methods such as ordinary least squares, weighted least squares (Patton and Sheppard, 2015) and robust regression (Clements and Preve, 2021) have been proposed for estimating HAR models. When dealing with realized volatilities for multiple assets ($N > 1$), it is natural to extend the univariate HAR model to a vector HAR (VHAR) model (Bubák et al., 2011; Souček and Todorova, 2013). The VHAR model for N assets can be expressed as:

$$\mathbf{y}_n^{(d)} = \mathbf{\Phi}^{(d)}\mathbf{y}_{n-1}^{(d)} + \mathbf{\Phi}^{(w)}\mathbf{y}_{n-1}^{(w)} + \mathbf{\Phi}^{(m)}\mathbf{y}_{n-1}^{(m)} + \boldsymbol{\varepsilon}_n, \tag{1.1}$$

where $\mathbf{y}_n^{(d)}$ represents the N-dimensional realized measure for assets on day n, $\mathbf{y}_{n-1}^{(w)} = (1/5) \sum_{j=1}^{5} \mathbf{y}_{n-j}^{(d)}$ denotes the weekly realized measure, while $\mathbf{y}_{n-1}^{(m)} = (1/22) \sum_{j=1}^{22} \mathbf{y}_{n-j}^{(d)}$ represents the monthly realized measure. When $N = 1$, the VHAR model reduces to the univariate HAR model, and the daily, weekly, and monthly volatility components are also reduced to their respective univariate realized measures. Here, the lag structure of HAR models is typically set to $(1, 5, 22)$, considering one-day, five-day, and twenty-two-day lags.

Hong et al. (2020) extended the VHAR model (1.1) by incorporating additional volatility components, and established the asymptotic normality of the least squares estimators for models with independent and autocorrelated errors. However, when the number of assets N is large, the VHAR model would involve a substantial number of parameters in the order of $O(N^2)$. Therefore, dimension reduction becomes essential in practical applications. One approach is to utilize the vector HAR index model introduced by Cubadda et al. (2017), which assumes a low-rank structure in the row space of the coefficient matrices, effectively reducing the number of parameters to $O(N)$. By reducing the dimensionality of VHAR models, it becomes computationally more feasible and provides better estimation efficiency.

HAR-type models, despite their advantages, have faced criticisms regarding their flexibility in empirical studies. Ongoing discussions have raised questions about whether the average value truly represents the optimal choice for characterizing the relationship between multi-period historical volatilities and the daily forecasting target in HAR models. Researchers such as Chen et al. (2010) have demonstrated that incorporating a structural break along with a low-order autoregressive model can outperform traditional HAR models. Audrino and Knaus (2016) evaluated the inclusion of daily, weekly, and monthly components in the HAR model empirically using the LASSO method (Tibshirani, 1996) based on data of realized variance of nine U.S. assets. They found that the LASSO often selected low-order models, suggesting potential shortcomings in including the daily, weekly, and monthly components. Additionally, Hong et al. (2020) reported that a lag structure of $(1, 5, 6)$ outperformed the classic lag structure of $(1, 5, 22)$ in the HAR model when forecasting the daily closing prices of Gold and S&P 500 index based on a three-year dataset.

The calendar effect is a well-recognized phenomenon observed in financial time series, where calendar-related factors influence stock market and financial volatilities. Factors such as the day of the week, time of the month, and time of the year have been found to have an impact. The study of the calendar effect dates back to the 1980s, with pioneering work by Cleveland and Devlin (1980, 1982) emphasizing the need for calendar adjustment in monthly time series. Subsequent empirical studies by researchers such as Sullivan et al. (2001); Levy and Joseph (2012); Proietti and Pedregal (2023) consistently support the existence of the calendar effects.

This paper introduces a novel approach called the multi-linear low-rank HAR (MLRHAR) model. One notable aspect of this approach is replacing the fixed heterogeneous volatility components in the HAR model with a data-driven alternative. To address the calendar effect effectively, a special case of the MLRHAR model, called the multi-linear low-rank fourth-order tensor HAR (MLR-FT-HAR) model is considered. This model decomposes the temporal direction into two distinct components: the short temporal and calendar temporal directions. The coefficient matrices are transformed into a fourth-order tensor denoted by $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N\times N\times S\times Q}$, assuming low multi-linear ranks of (r_1, r_2, r_3, r_4). This transformation significantly reduces the number of parameters to $r_1r_2r_3r_4 + Nr_1 + Nr_2 + Sr_3 + Qr_4$. In this framework, the mode-1, -2, -3, and -4 matricization of $\boldsymbol{\mathcal{A}}$ correspond to the column, row, calendar temporal, and short temporal directions, respectively. The MLR-FT-HAR model not only reduces dimensionality but also determines the components related to shorter and calendar temporal volatilities, offering a comprehensive solution to the calendar effect. For more detailed explanations, readers may refer to Section 2.1.

If the calendar effect is disregarded, the MLR-FT-HAR model simplifies to the multilinear low-rank third-order tensor HAR (MLR-TT-HAR) model. This model incorporates low-rank assumptions on the column space, row space, and lag space of coefficient matrices. By transforming the coefficient matrices into a third-order tensor, the response, predictor, and temporal directions are represented through the mode-1, -2, and -3 matricizations of the tensor, respectively. This reduction effectively reduces the parameter dimension to $r_1r_2r_3 + Nr_1 + Nr_2 + Pr_3$. Notably, the benchmark VHAR model (1.1) can be seen as a special case of the MLR-TT-HAR model. Readers may refer to Section 2.2 for further details.

To estimate the parameters of the newly proposed MLR-HAR model, the least squares method is considered in Section 3, and non-asymptotic properties are derived for the high-dimensional estimation. Moreover, a projected gradient descent algorithm with theoretical justifications is suggested to search for estimates. In many practical scenarios, it is important to quantify the estimation error in addition to point estimations, referred to as uncertainty quantification or statistical inference (Xia et al., 2022; Agterberg and Zhang, 2024), and we also provide the statistical inference of high-dimensional estimation. See Corollary 1 for details. In addition, the finite-sample properties of the proposed MLR-HAR model are assessed through a simulation study and a real data analysis by utilizing the realized variance of the constituent stocks of the S&P 500 Index. Both analyses demonstrate that the MLR-HAR model surpasses the forecasting performance of VHAR index models (Cubadda et al., 2017) and the benchmark VHAR model (1.1). Also, the MLR-FT-HAR model exhibits superior forecasting performance compared to the MLR-TT-HAR model in real data analysis, highlighting the significance of splitting into two temporal directions when forecasting the realized measure.

The paper is organized as follows: Section 2 presents an introduction to the MLR-FT-HAR model and its third-order counterpart. In Section 3, we develop a high-dimensional HAR modeling approach specifically designed for these models. The results from simulation experiments are discussed in Section 4, and real data examples are provided in Section 5 to demonstrate the practical utility of our method. We conclude the paper in Section 6 with a concise summary. The technical proofs of theorems can be found in the supplementary file.

Throughout the paper, tensors are denoted by calligraphic capital letters, e.g., $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}$, etc.; see the supplementary file for a brief introduction to tensor notations and Tucker decomposition. Vectors are denoted by small boldface letters, such as $\mathbf{a}$ and $\mathbf{b}$. Let $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_j a_j b_j$ and $\|\boldsymbol{a}\|_2 = \sqrt{\langle \boldsymbol{a}, \boldsymbol{a} \rangle}$ be the inner product and ℓ_2-norm, respectively. Matrices are represented using capital letters, for example, $\mathbf{A}$ and $\mathbf{B}$. For a matrix $\mathbf{A} \in \mathbb{R}^{p_1 \times p_2}$, let $\mathbf{A}^\top, \operatorname{rank}(\mathbf{A}), \sigma_{\max}(\mathbf{A}), \sigma_{\min}(\mathbf{A}), \lambda_{\max}(\mathbf{A})$, $\lambda_{\min}(\mathbf{A})$, $\|\mathbf{A}\|_{\text{op}} = \sigma_{\max}(\mathbf{A})$ and $\|\mathbf{A}\|_{\text{F}} = \sqrt{\sum_{i,j} \mathbf{A}_{ij}^2}$ be its transpose, rank, largest, smallest non-zero singular value, largest, smallest eigenvalue, operator norm and Frobenius norm, respectively. Moreover, for any $p_1 \geqslant p_2$, the set of orthonormal matrices is denoted by $\mathcal{O}^{p_1 \times p_2} := \left\{ \mathbf{O} \in \mathbb{R}^{p_1 \times p_2} \mid \mathbf{O}^\top \mathbf{O} = \mathbf{I}_{p_2} \right\}$, where $\mathbf{I}_{p_2}$ is a $p_2 \times p_2$ identity matrix. Finally, for any two sequences a_n and b_n, we denote $a_n \lesssim b_n$ if there exists an absolute constant $C > 0$ such that $a_n \leqslant C b_n$, $a_n \gtrsim b_n$ if $a_n \geqslant C b_n$, and write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

2. Multilinear low-rank tensor HAR models

2.1 Multilinear low-rank fourth-order tensor HAR model

Let $\mathbf{y}_n = (y_{1,n}, \cdots, y_{N,n})^\top \in \mathbb{R}^N$ represent the log-transformed realized measures. We define S as the order of the calendar temporal direction and Q as the order of the short temporal direction. In the s-th calendar cycle, $\mathbf{A}_q^{(s)}$ (for $1 \le q \le Q$) represents the $N \times N$ coefficient matrices, and T denotes the sample size. We consider the following model:

$$\begin{aligned}\mathbf{y}_n =& \boldsymbol{\varepsilon}_n + \underbrace{\mathbf{A}_1^{(1)}\mathbf{y}_{n-1} + \cdots + \mathbf{A}_Q^{(1)}\mathbf{y}_{n-Q}}_{Q \text{ terms}} \\ &+ \underbrace{\mathbf{A}_1^{(2)}\mathbf{y}_{n-(Q+1)} + \cdots + \mathbf{A}_Q^{(2)}\mathbf{y}_{n-2Q}}_{Q \text{ terms}} + \underbrace{\cdots + \cdots}_{(S-3)Q \text{ terms}} \\ &+ \underbrace{\mathbf{A}_1^{(S)}\mathbf{y}_{n-((S-1)Q+1)} + \cdots + \mathbf{A}_Q^{(S)}\mathbf{y}_{n-SQ}}_{Q \text{ terms}},\end{aligned} \tag{2.1}$$

where $\{\boldsymbol{\varepsilon}_n\}$ with $\boldsymbol{\varepsilon}_n = (\varepsilon_{1,n}, \cdots, \varepsilon_{N,n})^\top \in \mathbb{R}^N$ are independent and identically distributed (i.i.d.) random vectors. The innovations satisfy $\mathbb{E}(\boldsymbol{\varepsilon}_n) = 0$ and $\text{var}(\boldsymbol{\varepsilon}_n) < \infty$. Model (2.1) distinguishes the calendar and short-term temporal effects. The matrix polynomial for this model is defined as $\boldsymbol{\mathcal{A}}(z) = \mathbf{I}_N - \mathbf{A}_1^{(1)}z - \cdots - \mathbf{A}_Q^{(1)}z^Q - \cdots - \mathbf{A}_1^{(S)}z^{(S-1)Q+1} - \cdots - \mathbf{A}_Q^{(S)}z^{SQ}$, where $z \in \mathbb{C}$, with $\mathbb{C}$ representing the complex space. We introduce the following assumption:

Assumption 1. The determinant of $\boldsymbol{\mathcal{A}}(z)$ is non-zero for all $|z| < 1$.

Assumption 1 serves as a necessary and sufficient condition for achieving strict stationarity in a vector autoregression. It guarantees that the sequence $\{\mathbf{y}_n\}$ is strictly stationary.

To address the potentially large number of parameters in model (2.1), which amounts to N^2SQ parameters, we employ the fourth-order tensor technique for dimension reduction. The approach is depicted in Figure 1 and can be summarized as follows. In each calendar cycle s, the coefficient matrices $(\mathbf{A}_1^{(s)}, \cdots, \mathbf{A}_Q^{(s)})$ are organized into a third-order tensor. This tensorization process is repeated for $1 \leq s \leq S$. The resulting S third-order tensors are then combined to form a fourth-order tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N \times N \times S \times Q}$, where $\boldsymbol{\mathcal{A}}_{(1)} = \left(\mathbf{A}_1^{(1)}, \cdots, \mathbf{A}_Q^{(1)}, \cdots, \mathbf{A}_1^{(S)}, \cdots, \mathbf{A}_Q^{(S)}\right)$.

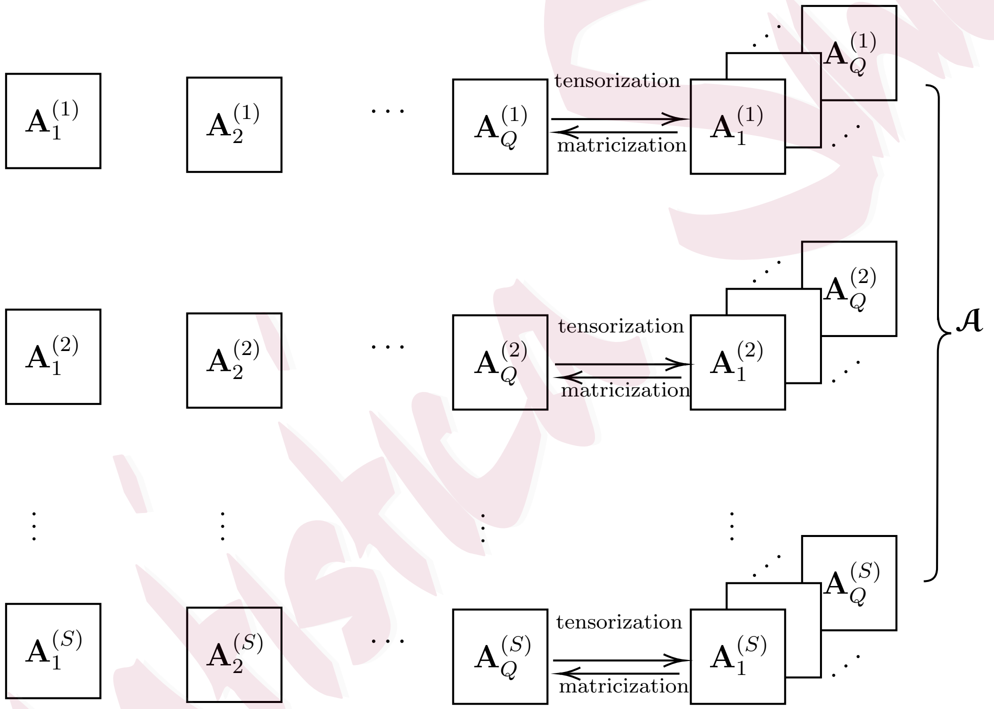

Figure 1: Rearranging $\mathbf{A}_q^{(s)}$ into a fourth-order tensor $\boldsymbol{\mathcal{A}}$.

We introduce the assumption of multilinear low ranks (r_1, r_2, r_3, r_4) for the coefficient tensor $\boldsymbol{\mathcal{A}}$. This allows us to represent $\boldsymbol{\mathcal{A}}$ using a Tucker decomposition (De Lathauwer et al., 2000) as follows:

$$\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{G}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 \times_4 \mathbf{U}_4. \tag{2.2}$$

Here, $\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3 \times r_4}$ represents the core tensor, and $\mathbf{U}_1 \in \mathbb{R}^{N \times r_1}$, $\mathbf{U}_2 \in \mathbb{R}^{N \times r_2}$, $\mathbf{U}_3 \in \mathbb{R}^{S \times r_3}$, and $\mathbf{U}_4 \in \mathbb{R}^{Q \times r_4}$ are factor matrices. For simplicity, we refer to model (2.1) with the Tucker decomposition (2.2) as the multilinear low-rank fourth-order tensor HAR (MLR-FT-HAR) model. The MLR-FT-HAR model significantly reduces the number of parameters to $r_1 r_2 r_3 r_4 + N r_1 + N r_2 + S r_3 + Q r_4$.

We observe that $\boldsymbol{\mathcal{A}}_{(2)} = \left({\mathbf{A}_1^{(1)}}^\top, \cdots, {\mathbf{A}_Q^{(1)}}^\top, \cdots, {\mathbf{A}_1^{(S)}}^\top, \cdots, {\mathbf{A}_Q^{(S)}}^\top\right)$, $\boldsymbol{\mathcal{A}}_{(3)} = \left(\mathrm{vec}\left(\mathbf{A}_1^{(1)}, \cdots, \mathbf{A}_Q^{(1)}\right), \cdots, \mathrm{vec}\left(\mathbf{A}_1^{(S)}, \cdots, \mathbf{A}_Q^{(S)}\right)\right)^\top$, and $\boldsymbol{\mathcal{A}}_{(4)} = \left(\mathrm{vec}\left(\mathbf{A}_1^{(1)}, \cdots, \mathbf{A}_1^{(S)}\right), \cdots, \mathrm{vec}\left(\mathbf{A}_Q^{(1)}, \cdots, \mathbf{A}_Q^{(S)}\right)\right)^\top$. The spaces spanned by $\boldsymbol{\mathcal{A}}_{(1)}$, $\boldsymbol{\mathcal{A}}_{(2)}$, $\boldsymbol{\mathcal{A}}_{(3)}$, and $\boldsymbol{\mathcal{A}}_{(4)}$ correspond to the column space, row space, calendar space, and short temporal space of the coefficient matrices, respectively. This decomposition allows us to analyze and interpret the different structural aspects of the tensor model in terms of these spaces.

Remark 1. Consider $\mathcal{H} = \mathcal{G} \times_3 \mathbf{U}_3 \times_4 \mathbf{U}_4 \in \mathbb{R}^{r_1 \times r_2 \times S \times Q}$, and let $\mathbf{H}_j \in \mathbb{R}^{r_1 \times r_2}$ denote its j-th frontal slice for $1 \leq j \leq SQ$. In other words, $\mathcal{H}_{(1)} = (\mathbf{H}_1, \cdots, \mathbf{H}_Q, \cdots, \mathbf{H}_{(S-1)Q+1}, \cdots \mathbf{H}_{SQ}) \in \mathbb{R}^{r_1 \times r_2 SQ}$. Thus, we can express $\boldsymbol{\mathcal{A}}$ as the tensor product $\boldsymbol{\mathcal{A}} = \mathcal{H} \times_1 \mathbf{U}_1 \times {}_2\mathbf{U}_2$. By rewriting model (2.1), we obtain:

$$\mathbf{y}_n = \mathbf{U}_1 \sum_{j=1}^{SQ} \mathbf{H}_j \mathbf{U}_2^\top \mathbf{y}_{n-j} + \boldsymbol{\varepsilon}_n \quad \text{or} \quad \mathbf{U}_1^\top \mathbf{y}_n = \sum_{j=1}^{SQ} \mathbf{H}_j \mathbf{U}_2^\top \mathbf{y}_{n-j} + \mathbf{U}_1^\top \boldsymbol{\varepsilon}_n,$$

where $\mathbf{U}_1^\top \mathbf{y}_n$ represents r_1 response factors across the N variables of $\mathbf{y}_n$, and $\mathbf{U}_2^\top \mathbf{y}_{n-j}$ represents r_2 predictor factors across the N variables of $\mathbf{y}_{n-j}$.

Remark 2. Let $\mathcal{S} = \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_4 \mathbf{U}_4 \in \mathbb{R}^{N \times N \times r_3 \times Q}$, and let $\mathbf{S}_j \in \mathbb{R}^{N \times N}$ represents its j-th frontal slice for $1 \leq j \leq Q r_3$. In other words, $\mathbf{S}^{(k)} = \left(\mathbf{S}_{(k-1) \times Q+1}, \mathbf{S}_{(k-1) \times Q+2}, \ldots, \mathbf{S}_{k \times Q}\right) \in \mathbb{R}^{N \times NQ}$ for $1 \leq k \leq r_3$, and $\mathcal{S}_{(1)} = (\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, \cdots, \mathbf{S}^{(r_3)})$. Hence $\boldsymbol{\mathcal{A}} = \mathcal{S} \times_3 \mathbf{U}_3$. Let

$\mathbf{U}_3 = (\mathbf{u}_3^{(1)}, \ldots, \mathbf{u}_3^{(r_3)}) \in \mathbb{R}^{S\times r_3}$, where $\mathbf{u}_3^{(k)} = (u_{3,1}^{(k)}, \ldots, u_{3,S}^{(k)})^\top \in \mathbb{R}^S$ for $1 \le k \le r_3$. We introduce the following notations: $\mathbf{y}_1^{(Q)} = \left(\mathbf{y}_{n-1}^\top, \mathbf{y}_{n-2}^\top, \ldots, \mathbf{y}_{n-Q}^\top\right)^\top \in \mathbb{R}^{NQ}$, $\mathbf{y}_2^{(Q)} = \left(\mathbf{y}_{n-Q-1}^\top, \mathbf{y}_{n-Q-2}^\top, \ldots, \mathbf{y}_{n-2Q}^\top\right)^\top \in \mathbb{R}^{NQ}$, $\cdots$, $\mathbf{y}_S^{(Q)} = \left(\mathbf{y}_{n-(S-1)Q-1}^\top, \mathbf{y}_{n-(S-1)Q-2}^\top, \ldots, \mathbf{y}_{n-SQ}^\top\right)^\top \in \mathbb{R}^{NQ}$. By reformulating model (2.1), we obtain:

$$\mathbf{y}_n = \mathbf{S}^{(1)}\mathbf{x}_{Q,n}^{(1)} + \cdots + \mathbf{S}^{(r_3)}\mathbf{x}_{Q,n}^{(r_3)} + \boldsymbol{\varepsilon}_n \quad \text{with} \quad \mathbf{x}_{Q,n}^{(k)} = \sum_{j=1}^{S} u_{3,j}^{(k)}\mathbf{y}_j^{(Q)}. \tag{2.3}$$

In the above equation, $\mathbf{x}_{Q,n}^{(k)}$ represents the summarized factors along the calendar temporal direction. These factors can be interpreted as r_3 heterogeneous volatility components, which are automatically selected during the estimation process.

Remark 3. Let $\boldsymbol{\mathcal{Q}} = \boldsymbol{\mathcal{G}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 \in \mathbb{R}^{N\times N\times S\times r_4}$, and let $\mathbf{Q}_j \in \mathbb{R}^{N\times N}$ represents its j-th frontal slice for $1 \le j \le Sr_4$. In other words, $\mathbf{Q}^{(k)} = \left(\mathbf{Q}_k, \mathbf{Q}_{Q+k}, \ldots, \mathbf{Q}_{(S-1)\times Q+k}\right) \in \mathbb{R}^{N\times NS}$ for $1 \le k \le r_4$, and $\boldsymbol{\mathcal{Q}}_{(1)} = (\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}, \cdots, \mathbf{Q}^{(r_4)})$. Hence $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{Q}} \times_4 \mathbf{U}_4$. Let $\mathbf{U}_4 = (\mathbf{u}_4^{(1)}, \ldots, \mathbf{u}_4^{(r_4)}) \in \mathbb{R}^{Q\times r_4}$, where $\mathbf{u}_4^{(k)} = (u_{4,1}^{(k)}, \ldots, u_{4,Q}^{(k)})^\top \in \mathbb{R}^Q$ for $1 \le k \le r_4$. We also introduce the following notations. $\mathbf{y}_1^{(S)} = \left(\mathbf{y}_{n-1}^\top, \mathbf{y}_{n-Q-1}^\top, \ldots, \mathbf{y}_{n-(S-1)Q-1}^\top\right)^\top \in \mathbb{R}^{NS}$, $\mathbf{y}_2^{(S)} = \left(\mathbf{y}_{n-2}^\top, \mathbf{y}_{n-Q-2}^\top, \ldots, \mathbf{y}_{n-(S-1)Q-2}^\top\right)^\top \in \mathbb{R}^{NS}$, $\cdots$, $\mathbf{y}_Q^{(S)} = \left(\mathbf{y}_{n-Q}^\top, \mathbf{y}_{n-2Q}^\top, \ldots, \mathbf{y}_{n-SQ}^\top\right)^\top \in \mathbb{R}^{NS}$. Using these notations in model (2.1), we have

$$\mathbf{y}_n = \mathbf{Q}^{(1)}\mathbf{x}_{S,n}^{(1)} + \cdots + \mathbf{Q}^{(r_4)}\mathbf{x}_{S,n}^{(r_4)} + \boldsymbol{\varepsilon}_n \quad \text{with} \quad \mathbf{x}_{S,n}^{(k)} = \sum_{j=1}^{Q} u_{4,j}^{(k)}\mathbf{y}_j^{(S)}, \tag{2.4}$$

where $\mathbf{x}_{S,n}^{(k)}$ represents the summarized factors along the short temporal direction. These factors can be interpreted as r_4 heterogeneous volatility components, which are automatically selected during the estimation process.

2.2 Multilinear low-rank third-order tensor HAR model

In this section, we discuss the multilinear low-rank HAR model without the calendar effect. When the calendar effect is absent, the model (2.1) can be represented as follows:

$$\mathbf{y}_n = \sum_{j=1}^{P} \mathbf{A}_j \mathbf{y}_{n-j} + \boldsymbol{\varepsilon}_n. \tag{2.5}$$

Here, $\mathbf{y}_n$ represents the N-dimensional log-transformed realized measures, P denotes the order, $\mathbf{A}_j$'s are $N \times N$ coefficient matrices, and $\{\boldsymbol{\varepsilon}_n\}$ satisfy the same conditions as in (2.1).

The matrix polynomial for model (2.5) is defined as $\boldsymbol{\mathcal{A}}(z) = \mathbf{I}_N - \mathbf{A}_1 z - \cdots - \mathbf{A}_P z^P$, where $z \in \mathbb{C}$ and $\mathbb{C}$ represents the complex space. Similar to Assumption 1, to ensure strict stationarity of $\{\mathbf{y}_n\}$, the determinant of $\boldsymbol{\mathcal{A}}(z)$ should be nonzero for all $|z| < 1$. In model (2.5), the total number of parameters is $N^2 P$, which can be prohibitively large. To address this issue, we employ third-order tensor techniques to achieve dimension reduction in the parameter space. Specifically, we rearrange the coefficient matrices into a third-order tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N \times N \times P}$ such that $\boldsymbol{\mathcal{A}}_{(1)} = (\mathbf{A}_1, \ldots, \mathbf{A}_P)$, as illustrated in Figure 2.

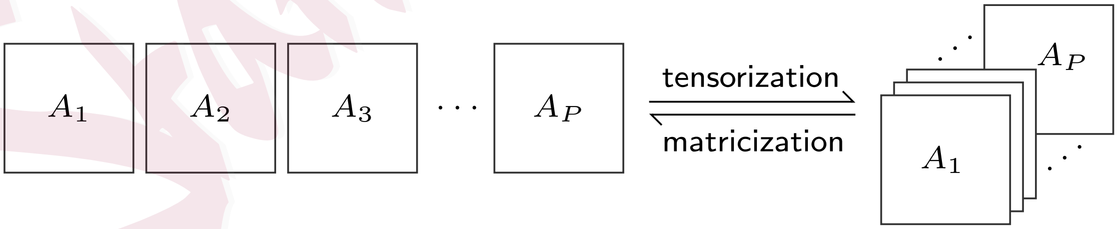

Figure 2: Rearranging $\mathbf{A}_j$s into a third-order tensor $\boldsymbol{\mathcal{A}}$.

We introduce the assumption of multilinear low ranks (r_1, r_2, r_3) for the coefficient tensor

$\boldsymbol{\mathcal{A}}$. Consequently, we can express it using a Tucker decomposition:

$$\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{G}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3. \tag{2.6}$$

Here, $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ represents the core tensor, and $\mathbf{U}_1 \in \mathbb{R}^{N \times r_1}$, $\mathbf{U}_2 \in \mathbb{R}^{N \times r_2}$, and $\mathbf{U}_3 \in \mathbb{R}^{P \times r_3}$ are the factor matrices. We refer to (2.5) with the Tucker decomposition (2.6) as the multilinear low-rank third-order tensor HAR (MLR-TT-HAR) model for simplicity.

It is important to note that $\boldsymbol{\mathcal{A}}_{(2)} = (\mathbf{A}_1^\top, \cdots, \mathbf{A}_P^\top)$ and $\boldsymbol{\mathcal{A}}_{(3)} = (\text{vec}(\mathbf{A}_1), \ldots, \text{vec}(\mathbf{A}_P))^\top$. The spaces spanned by $\boldsymbol{\mathcal{A}}_{(1)}$, $\boldsymbol{\mathcal{A}}_{(2)}$, and $\boldsymbol{\mathcal{A}}_{(3)}$ correspond to the column, row, and temporal spaces of the coefficient matrices, respectively. Consequently, the low-rank assumption in (2.6) restricts the parameter space from three directions simultaneously. The number of parameters is then reduced to $r_1 r_2 r_3 + N r_1 + N r_2 + P r_3$, which is generally larger than that of the MLR-FT-HAR model.

Let $\boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{G}} \times_3 \mathbf{U}_3$, and $\mathbf{H}_j \in \mathbb{R}^{r_1 \times r_2}$ represents its j-th frontal slice for $1 \leq j \leq P$, i.e. $\boldsymbol{\mathcal{H}}_{(1)} = (\mathbf{H}_1, \mathbf{H}_2, \cdots, \mathbf{H}_P)$. Thus, we can express $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{H}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2$, and model (2.5) can be rewritten as:

$$\mathbf{y}_n = \mathbf{U}_1 \sum_{j=1}^{P} \mathbf{H}_j \mathbf{U}_2^\top \mathbf{y}_{n-j} + \boldsymbol{\varepsilon}_n \quad \text{or} \quad \mathbf{U}_1^\top \mathbf{y}_n = \sum_{j=1}^{P} \mathbf{H}_j \mathbf{U}_2^\top \mathbf{y}_{n-j} + \mathbf{U}_1^\top \boldsymbol{\varepsilon}_n.$$

Here, $\mathbf{U}_1^\top \mathbf{y}_n$ and $\mathbf{U}_2^\top \mathbf{y}_{n-j}$ represent the summarized factors of responses and predictors, respectively.

It is worth noting that the vector HAR index model in Cubadda et al. (2017) corresponds to the case with $r_1 = N$ and $\mathbf{U}_1$ being an identity matrix.

Remark 4. Let $\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{G}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \in \mathbb{R}^{N \times N \times r_3}$, and $\mathbf{S}_j \in \mathbb{R}^{N \times N}$ be its j-th frontal slice for $1 \leq j \leq r_3$, i.e. $\boldsymbol{\mathcal{S}}_{(1)} = (\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_{r_3})$. Denote $\mathbf{U}_3 = (\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(r_3)}) \in \mathbb{R}^{P \times r_3}$ and

$\mathbf{u}^{(k)} = (u_1^{(k)}, \ldots, u_P^{(k)})^\top \in \mathbb{R}^P$, where $1 \leq k \leq r_3$. Hence we can express $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{S}} \times_3 \mathbf{U}_3$, and model (2.5) can be reformulated as

$$\mathbf{y}_n = \mathbf{S}_1 \mathbf{x}_n^{(1)} + \cdots + \mathbf{S}_{r_3} \mathbf{x}_n^{(r_3)} + \boldsymbol{\varepsilon}_n \quad \text{with} \quad \mathbf{x}_n^{(k)} = \sum_{j=1}^{P} u_j^{(k)} \mathbf{y}_{n-j}. \tag{2.7}$$

The variables $\mathbf{x}_n^{(k)}$ represent the summarized factors along the temporal direction and can be interpreted as r_3 heterogeneous volatility components, which are automatically selected by the estimation method. Furthermore, $\mathbf{U}_3$ is the corresponding loading matrix, and model (2.7) reduces to the VHAR model (Bubák et al., 2011; Souček and Todorova, 2013) when $\mathbf{U}_3 = \mathbf{U}_\mathrm{C}$, where

$$\mathbf{U}_\mathrm{C}^\top = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & \cdots & 0 \\ 1/22 & 1/22 & 1/22 & 1/22 & 1/22 & 1/22 & \cdots & 1/22 \end{pmatrix} \in \mathbb{R}^{3\times 22}.$$

From the above statements, the traditional VHAR model (Bubák et al., 2011; Souček and Todorova, 2013) is a specific case of our proposed MLR-TT-HAR model, where the factor matrices are set to fixed values ($\mathbf{U}_1 = \mathbf{I}_N, \mathbf{U}_2 = \mathbf{I}_N, \mathbf{U}_3 = \mathbf{U}_\mathrm{C}$). In contrast, the MLR-TT-HAR model is capable of automatically selecting heterogeneous volatility components, and the MLR-FT-HAR model can identify heterogeneous volatility components in both calendar and short temporal directions. Additionally, the MLR-FT-HAR and MLR-TT-HAR models can also reduce the parameter space from the column and row directions simultaneously, functioning as powerful supervised factor models. As a result, these proposed models are fundamentally based on a robust and data-driven methodology.

3. High-dimensional HAR modeling

3.1 Estimation and theoretical properties

In practice, it is common to deal with a large number of assets. That means the size of N can be very large and may grow with the sample size T at an arbitrary rate. In our context, this corresponds to high-dimensional HAR modeling, and this section focuses on establishing the non-asymptotic properties of the estimators for MLR-FT-HAR models, and similar results for the MLR-TT-HAR model.

Suppose that we have prior knowledge of multilinear ranks (r_1, r_2, r_3, r_4) of the coefficient tensor $\boldsymbol{\mathcal{A}}$. Consider an observed vector time series $\{\mathbf{y}_{-SQ+1}, \dots, \mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_T\}$, generated by model (2.1), with low-rank constraint at (2.2), we can define an Ordinary Least Squares (OLS) estimator for parameters of the MLR-FT-HAR model as follows:

$$\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}} \equiv [\![\widehat{\boldsymbol{\mathcal{G}}}; \widehat{\mathbf{U}}_1, \widehat{\mathbf{U}}_2, \widehat{\mathbf{U}}_3, \widehat{\mathbf{U}}_4]\!] = \operatorname{argmin} L(\boldsymbol{\mathcal{G}}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{U}_4),$$

where $\mathbf{x}_n = (\mathbf{y}_{n-1}^\top, \dots, \mathbf{y}_{n-SQ}^\top)^\top$, and

$$L(\boldsymbol{\mathcal{G}}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{U}_4) = \frac{1}{T}\sum_{n=1}^{T} ||\mathbf{y}_n - (\boldsymbol{\mathcal{G}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 \times_4 \mathbf{U}_4)_{(1)}\mathbf{x}_n||_2^2.$$

It should be noted that although the components of the Tucker decomposition, $\boldsymbol{\mathcal{G}}$, $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$, and $\mathbf{U}_4$, are not individually identifiable, the coefficient tensor $\boldsymbol{\mathcal{A}}$ can be uniquely identified.

Assumption 2. Model error $\boldsymbol{\varepsilon}_n$ follows a structured form of $\boldsymbol{\varepsilon}_n = \boldsymbol{\Sigma}_\varepsilon^{1/2}\boldsymbol{\xi}_n$. Here, $\{\boldsymbol{\xi}_n\}$ represents a sequence of i.i.d. random vectors satisfying $\mathbb{E}(\boldsymbol{\xi}_n) = 0$, $\text{var}(\boldsymbol{\xi}_n) = \mathbf{I}_N$, and $\boldsymbol{\Sigma}_\varepsilon$ is a positive definite matrix representing the variance-covariance matrix of $\boldsymbol{\varepsilon}_n$. As well, the individual entries $(\boldsymbol{\xi}_{in})_{1\leq i\leq N}$ of the vector $\boldsymbol{\xi}_n$ are mutually independent, and each entry is assumed to be κ^2-sub-Gaussian.

The assumption of sub-Gaussianity, as mentioned above, is commonly employed in the literature for high-dimensional settings (Wainwright, 2019). In the following, we will derive non-asymptotic error bounds, which will depend on the temporal and cross-sectional dependence of $\mathbf{y}_n$ (Basu and Michailidis, 2015).

To begin, let us start by introducing two dependence measures as follows:

$$\mu_{\min}(\boldsymbol{\mathcal{A}}) := \min_{|z|=1} \lambda_{\min}(\boldsymbol{\mathcal{A}}^*(z)\boldsymbol{\mathcal{A}}(z)) \quad \text{and} \quad \mu_{\max}(\boldsymbol{\mathcal{A}}) := \max_{|z|=1} \lambda_{\max}(\boldsymbol{\mathcal{A}}^*(z)\boldsymbol{\mathcal{A}}(z)),$$

where $\boldsymbol{\mathcal{A}}^*(z)$ represents the conjugate transpose of $\boldsymbol{\mathcal{A}}(z)$. Let us further define two additional quantities: $\kappa_L = \lambda_{\min}(\boldsymbol{\Sigma}_\varepsilon)/\mu_{\max}(\boldsymbol{\mathcal{A}})$ and $\kappa_U = \lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)/\mu_{\min}(\boldsymbol{\mathcal{A}})$. Moreover, we introduce $d_{\mathcal{M}}$, which represents the model complexity for MLR-FT-HAR models and is defined as $d_{\mathcal{M}} = r_1r_2r_3r_4 + Nr_1 + Nr_2 + Sr_3 + Qr_4$.

Theorem 1. *Suppose that the sample size $T \gtrsim \max(\kappa^2, \kappa^4)(\kappa_U/\kappa_L)^2 d_{\mathcal{M}}$. If Assumptions 1 and 2 hold, then*

$$\|\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}} - \boldsymbol{\mathcal{A}}\|_{\text{F}} \leq \frac{C}{\kappa_L}\kappa^2\sqrt{\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)\kappa_U}\sqrt{\frac{d_{\mathcal{M}}}{T}}$$

and

$$\frac{1}{T}\sum_{n=1}^{T}\|(\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}})_{(1)}\mathbf{x}_n - \boldsymbol{\mathcal{A}}_{(1)}\mathbf{x}_n\|_2^2 \leq \frac{C}{\kappa_L}\left(\kappa^2\sqrt{\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)\kappa_U}\sqrt{\frac{d_{\mathcal{M}}}{T}}\right)^2,$$

with a probability at least $1 - \exp(-Cd_{\mathcal{M}}) - 2\exp\left(Cd_{\mathcal{M}} - CT(\kappa_L/\kappa_U)^2\ \min\{\kappa^{-2}, \kappa^{-4}\}\right)$. Here, C is a positive constant defined in the proof.

The above theorem provides upper bounds for both estimation and prediction errors. When κ_L and κ_U are bounded away from zero and infinity, the estimation error $\|\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}} - \boldsymbol{\mathcal{A}}\|_{\text{F}} = O_P(\sqrt{d_{\mathcal{M}}/T})$, where $d_{\mathcal{M}}$ measures the complexity of the MLR-FT-HAR model. Similarly, the prediction error $T^{-1}\sum_{n=1}^{T}\|(\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}})_{(1)}\mathbf{x}_n - \boldsymbol{\mathcal{A}}_{(1)}\mathbf{x}_n\|_2^2 = O_P(d_{\mathcal{M}}/T)$, and their consistency can be achieved as $T \to \infty$ and $d_{\mathcal{M}}/T \to 0$.

Remark 5. For the MLR-TT-HAR model, we can derive similar results to Theorem 1. In this case, the model complexity is given by $d_{\mathcal{M}} = r_1r_2r_3+Nr_1+Nr_2+Pr_3$, where r_1, r_2, r_3, N, and P denote the ranks and dimensions of coefficient tensors, respectively. The tensors $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}}$ and $\boldsymbol{\mathcal{A}}$ are also third-order tensors in this context. The matrix $\boldsymbol{\Sigma}_\varepsilon$, $\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)$, κ_L and κ_U are defined accordingly for the MLR-TT-HAR model. Moreover, the proposed method can continue to work if the three volatility components are fixed as daily, weekly, and monthly volatilities, i.e. $\mathbf{U}_3 = \mathbf{U}_{\text{C}}$ at Section 2.2. Note that the model complexity is given by $d_{\mathcal{M}} = 3r_1r_2 + Nr_1 + Nr_2$ and the parameter space can be reduced from the column and row directions, and the temporal direction simultaneously with ranks r_1, r_2, and $r_3 = 3$. Compared to the number of parameters $3N^2$ in the VHAR model, the proposed method can reduce the parameter space dramatically. Additionally, if the data deviates from the fixed volatilities' structure, the proposed MLR-TT-HAR model can automatically select the volatility components.

Many applications underscore the need for statistical inference methods capable of handling linear functionals for signal tensors, with loading tensors exhibiting diverse sparsity and structural complexity (Xu et al., 2025). We next provide an inference result for a functional $\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle$ with a prespecified loading tensor $\boldsymbol{\mathcal{B}}$. To this end, a debiased estimator $\widehat{\boldsymbol{\mathcal{A}}}^{u}_{(1)}$ is first introduced,

$$\widehat{\boldsymbol{\mathcal{A}}}^{u}_{(1)} = \widehat{\boldsymbol{\mathcal{A}}}_{(1)} + \frac{1}{T} \sum_{n=1}^{T} (\mathbf{y}_n - \widehat{\boldsymbol{\mathcal{A}}}_{(1)} \mathbf{x}_n) \mathbf{x}_n^\top \mathbf{K}, \tag{3.1}$$

where $\widehat{\boldsymbol{\mathcal{A}}}$ is the initial estimator, and $\mathbf{K}$ is the precision matrix of $\mathbf{x}_n$. Let $\widehat{\boldsymbol{\Delta}} = \boldsymbol{\mathcal{A}} - \widehat{\boldsymbol{\mathcal{A}}}$ be the estimation error. The debiased estimator can be decomposed as

$$\widehat{\boldsymbol{\mathcal{A}}}^{u}_{(1)} = \boldsymbol{\mathcal{A}}_{(1)} + \frac{1}{T} \sum_{n=1}^{T} \boldsymbol{\varepsilon}_n \mathbf{x}_n^\top \mathbf{K} + \frac{1}{T} \sum_{n=1}^{T} (\widehat{\boldsymbol{\Delta}}_{(1)} \mathbf{x}_n \mathbf{x}_n^\top \mathbf{K} - \widehat{\boldsymbol{\Delta}}_{(1)}). \tag{3.2}$$

Equation (3.2) bridges the debiased estimator $\widehat{\boldsymbol{\mathcal{A}}}^{u}_{(1)}$ with the true parameter $\boldsymbol{\mathcal{A}}_{(1)}$, and we next state the asymptotic distribution below.

Corollary 1. *Suppose the sample size* $T^{1/2} \gtrsim {\kappa_U^k}^2 \max(\kappa^2, \kappa^4) d_{\mathcal{M}} d_{\mathcal{B}}$, *and* $\kappa_U^k \asymp \kappa_U \|\mathbf{K}\|_2$. *If Assumptions 1 and 2 hold, then*

$$\sup_{u \in \mathbb{R}} \left| \mathbb{P}\left(\langle \widehat{\boldsymbol{\mathcal{A}}}^{u}, \boldsymbol{\mathcal{B}} \rangle \leq u \right) - \mathbb{P}\left(g \leq u \right) \right| \lesssim C T^{-1/9},$$

where $g \sim \mathcal{N}(0, T^{-1} \mathrm{tr}(\boldsymbol{\mathcal{B}}_{(1)} \mathbf{K} \boldsymbol{\mathcal{B}}_{(1)}^\top \boldsymbol{\Sigma}_{\varepsilon}))$, $\mathbf{K}$ *is the precision matrix of* $\mathbf{x}_n$, *and* $d_{\mathcal{B}}$ *denotes the size of the low-rank space that the prespecified loading tensor* $\boldsymbol{\mathcal{B}}$ *resides.*

The precision matrix $\mathbf{K}$ is unknown in real applications, while a good estimation method for it is still lacking in the literature. We leave it for future research.

3.2 Projected gradient descent algorithm

In contrast to the usual least squares estimator in linear regression, the OLS estimator $\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}}$ in Section 3.1 does not correspond to a convex optimization problem. This non-convex nature presents challenges for both numerical and theoretical parameter estimation.

To overcome these challenges, we propose a projected gradient descent (PGD) algorithm in this subsection, building upon the method outlined in Chen et al. (2019). We provide a theoretical guarantee that establishes the effectiveness of this algorithm. Our focus is primarily on the algorithm for the MLR-FT-HAR model, and similar results also hold for the MLR-TT-HAR model.

Let us consider the parameter space of MLR-FT-HAR models as defined in (2.1) and (2.2):

$$\boldsymbol{\Theta}(r_1, r_2, r_3, r_4) = \{\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N\times N\times S\times Q} : \text{rank}(\boldsymbol{\mathcal{A}}_{(i)}) \le r_i \quad \text{for} \quad 1 \le i \le 4\}.$$

To define a projection operator that maps any tensor $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{N\times N\times S\times Q}$ onto $\boldsymbol{\Theta}(r_1, r_2, r_3, r_4)$, we utilize the matricization operator $\mathcal{M}_i$, which converts a tensor to its mode-i matricization, and its inverse operator $\mathcal{M}_i^{-1}$ that maps a mode-i matricization back to the original tensor. In other words, $\mathcal{M}_i(\boldsymbol{\mathcal{B}}) = \boldsymbol{\mathcal{B}}_{(i)}$ and $\mathcal{M}_i^{-1}(\boldsymbol{\mathcal{B}}_{(i)}) = \boldsymbol{\mathcal{B}}$. In addition, we denote P_r as the projection operator that maps a matrix to its best rank-r approximation. Specifically, P_r performs SVD on the matrix and keeps the r largest singular values while setting the rest to zero.

By utilizing these operators, we can define the projection of a tensor $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{N\times N\times S\times Q}$ onto $\boldsymbol{\Theta}(r_1, r_2, r_3, r_4)$ as follows:

$$\begin{aligned}\widehat{P}_{\boldsymbol{\Theta}(r_1,r_2,r_3)}(\boldsymbol{\mathcal{B}}) := &(\mathcal{M}_4^{-1} \circ P_{r_4} \circ \mathcal{M}_4)\circ(\mathcal{M}_3^{-1} \circ P_{r_3} \circ \mathcal{M}_3)\\ &\circ(\mathcal{M}_2^{-1} \circ P_{r_2} \circ \mathcal{M}_2)\circ(\mathcal{M}_1^{-1} \circ P_{r_1} \circ \mathcal{M}_1)(\boldsymbol{\mathcal{B}}).\end{aligned}$$

Specifically, we first perform the mode-1 matricization of $\boldsymbol{\mathcal{B}}$. Then, we obtain the best rank-

r_1 approximation of the resulting matrix using SVD and fold it back into a fourth-order tensor. This process is repeated sequentially for the second, third, and fourth modes. For more details, refer to Algorithm 1.

It is instructive to mention that the order in which matricizations are performed is not important and does not affect the convergence analysis that follows. Also, $\widehat{P}_{\boldsymbol{\Theta}(r_1,r_2,r_3,r_4)}(\cdot)$ is only an approximate projection onto $\boldsymbol{\Theta}(r_1,r_2,r_3,r_4)$. The exact projection is a well-known NP-hard problem (Hillar and Lim, 2013).

We can incorporate the PGD method to compute the OLS estimator $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}}$. The details are provided in Algorithm 1. To begin, we update the estimate using the gradient descent method. However, since the updated tensor may not have low Tucker ranks, we project it onto $\boldsymbol{\Theta}(r_1,r_2,r_3,r_4)$.

Let (r_1',r_2',r_3',r_4') be the running Tucker ranks used in Algorithm 1. Denote $d_{\mathcal{M}}' = (r_1+r_1')(r_2+r_2')(r_3+r_3')(r_4+r_4')+N(r_1+r_1')+N(r_2+r_2')+S(r_3+r_3')+Q(r_4+r_4')$, and it represents the space size of the estimator that we get each time from the algorithm, plus the size of the true parameter space. The following theorem holds under certain assumptions.

Theorem 2. *Suppose that Assumptions 1 and 2 hold, $\eta = 2/(3\kappa_U)$, and the running Tucker ranks $r_i' \geq \left(\sqrt[4]{1+\kappa_L/(24\kappa_U)}-1\right)^{-2} r_i$ with $1 \leq i \leq 4$. If $T \gtrsim \max(\kappa^2,\kappa^4)\ (\kappa_U/\kappa_L)^2 d_{\mathcal{M}}'$, then*

$$\|\widehat{\boldsymbol{\mathcal{A}}}_K - \boldsymbol{\mathcal{A}}\|_{\text{F}} \leq (1-\frac{\kappa_L}{24\kappa_U})^K \|\widehat{\boldsymbol{\mathcal{A}}}_0 - \boldsymbol{\mathcal{A}}\|_{\text{F}} + \frac{C}{\kappa_L}\left(\kappa^2\sqrt{\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)\kappa_U}\sqrt{\frac{d_{\mathcal{M}}'}{T}}\right),$$

with probability no smaller than $1-\exp(-Cd_{\mathcal{M}}')-2\exp\left(Cd_{\mathcal{M}}' - CT(\kappa_L/\kappa_U)^2\ \min\{\kappa^{-2},\kappa^{-4}\}\right)$, where C is a positive constant. The precise definition of C is given in the proof of this the-

Algorithm 1 Projected gradient descent algorithm for HAR modeling

```
Input : data $\{\mathbf{y}_n\}$, parameter space $\mathbf{\Theta} = \mathbf{\Theta}(r_1, r_2, r_3, r_4)$, iterations $K$, step size $\eta$
Initialize : $k = 0$ and $\widehat{\boldsymbol{\mathcal{A}}}_0 \in \mathbf{\Theta}$.
for $k = 1, 2, \ldots, K$ do
    $\widetilde{\boldsymbol{\mathcal{A}}}_k = \widehat{\boldsymbol{\mathcal{A}}}_{k-1} - \eta \nabla L(\widehat{\boldsymbol{\mathcal{A}}}_{k-1})$ (gradient descent)
    for $j = 1, 2, 3, 4$ do
        $\boldsymbol{B}_j = \mathcal{M}_j(\widetilde{\boldsymbol{\mathcal{A}}}_k)$ (mode-$j$ matricization)
        $\widehat{\boldsymbol{B}}_j = P_{r_j}(\boldsymbol{B}_j)$ (best rank $r_j$ approximation by the SVD)
        $\widehat{\boldsymbol{\mathcal{A}}}_k = \mathcal{M}_j^{-1}(\widehat{\boldsymbol{B}}_j)$ (fold into tensor by reversing the mode-$j$ matricization)
    end for
end for
Output : $\widehat{\boldsymbol{\mathcal{A}}}_K$
```

orem.

The upper bound in the above theorem consists of two terms, which correspond to the optimization and statistical errors, respectively. The statistical error has a similar form to that in Theorem 1. It is worth noting that $\kappa_L < \kappa_U$, implying a linear convergence rate for the optimization error. Specifically, for any $\epsilon > 0$, we can select the number of iterations $K = [\log(\epsilon) - \log \|\widehat{\boldsymbol{\mathcal{A}}}_0 - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}] / \log[1 - \kappa_L/(24\kappa_U)]$ such that the optimization error is smaller than ϵ.

Remark 6. The optimization algorithm is a nonconvex problem in nature. Note that, in

Algorithm 1, the two main steps of gradient descent and projection are performed iteratively. The gradient descent step is a global search algorithm, and the projection step satisfies the contraction property, which guarantees that the estimator after projection has an error bound of the same order as the one before projection; see Lemma 4 in the supplementary file for details. Therefore, the whole algorithm can be viewed as nearly convex. As a result, unlike the local search algorithms in Han et al. (2020); Auddy and Yuan (2023), our algorithm is not sensitive to initial values, and the convergence analysis in Theorem 2 does not require a larger sample size or additional signal-to-noise ratio conditions.

3.3 Implementation issues

We first consider the initialization for Algorithm 1. Although it is not sensitive to initial values as mentioned, a good choice of initial values can result in faster convergence. First, following the discussion in Remark 1, $\{\mathbf{y}_n\}$ generated by the MLR-FT-HAR model follows a static factor model with $\sum_{j=1}^{SQ}\mathbf{H}_j\mathbf{U}_2^\top\mathbf{y}_{n-j}$ being the factors and $\mathbf{U}_1$ being the loading matrix. Therefore, we can find an initial value for the factor matrix $\mathbf{U}_1$ by the PCA method (Stock and Watson, 2002). Secondly, the spectral method is a common initialization for scalar on tensor regressions (Han et al., 2022), and we can adjust it to find the initial values of the remaining factor matrices. Finally, we apply the OLS method to initialize the core tensor. The detailed procedure is given below.

1. Let the covariance $\mathbf{\Sigma}_y = \frac{1}{T}\sum_{n=1}^{T}\mathbf{y}_n\mathbf{y}_n^\top$, with eigenvalue decomposition $\mathbf{\Sigma}_y = \mathbf{U}_1\mathbf{\Lambda}\mathbf{U}_1^\top$, then select the first $2r_1$ columns $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{2r_1}$ of $\mathbf{U}_1$ and get $\widehat{\mathbf{U}}_1^{(2r_1)}$, the MLR-FT-

HAR model can be rewritten as

$$\widehat{\mathbf{U}}_1^{(2r_1)^\top}\mathbf{y}_n = \boldsymbol{\mathcal{A}}_{(1)}^{re}\mathbf{x}_n + \widehat{\mathbf{U}}_1^{(2r_1)^\top}\boldsymbol{\varepsilon}_n, \tag{3.3}$$

where $\boldsymbol{\mathcal{A}}_{(1)}^{re} = \widehat{\mathbf{U}}_1^{(2r_1)^\top}\boldsymbol{\mathcal{A}}_{(1)}$ and $\mathbf{x}_n = (\mathbf{y}_{n-1}^\top, \dots, \mathbf{y}_{n-SQ}^\top)^\top$.

2. In model (3.3), the response vector is of length r_1. Consider it as r_1 scalar on vector regressions in parallel, and then they share the same factor matrices. Therefore, we can average them and utilize the spectral method to get a good estimate of the factor matrices $\mathbf{U}_j, j = 2, 3, 4$. Let $\widehat{\mathbf{y}}_n^{(2r_1)} = \widehat{\mathbf{U}}_1^{(2r_1)}\mathbf{y}_n \in \mathbb{R}^{2r_1}$, then average it and get $\widehat{y}_n^{(2r_1)} = \sum_{i=1}^{2r_1}\widehat{\mathbf{y}}_{in}^{(2r_1)}/2r_1$, where $\widehat{\mathbf{y}}_{in}^{(2r_1)}$ is the ith row of $\widehat{\mathbf{y}}_n^{(2r_1)}$. The model (3.3) can be rewritten as the following form

$$\widehat{y}_n^{(2r_1)} = \boldsymbol{\mathcal{A}}_{(1)}^{(2r_2,2r_3,2r_4)}\mathbf{x}_n + \varepsilon_n^{(2r_1)}, \tag{3.4}$$

where $\varepsilon_n^{(2r_1)} = \sum_{i=1}^{2r_1}\widehat{\boldsymbol{\varepsilon}}_{in}^{(2r_1)}/2r_1$ and $\widehat{\boldsymbol{\varepsilon}}_{in}^{(2r_1)}$ is the ith row of $\widehat{\boldsymbol{\varepsilon}}_n^{(2r_1)} = \widehat{\mathbf{U}}_1^{(2r_1)^\top}\boldsymbol{\varepsilon}_n$.

3. Compute the estimator $\widehat{\boldsymbol{\mathcal{A}}}_{(1)}^{(2r_2,2r_3,2r_4)} = \frac{1}{n}\sum_{n=1}^{T}\mathbf{x}_n^\top\widehat{y}_n^{(2r_1)}$, then fold it into a third-oder tensor $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{N\times S\times Q}$.

4. Input tensor $\boldsymbol{\mathcal{B}}$, ranks $2r_2, 2r_3, 2r_4$ and iteration, and Higher Order Orthogonal Iteration (HOOI) is employed to search the factor matrices $\widehat{\mathbf{U}}_2^{(2r_2)}$, $\widehat{\mathbf{U}}_3^{(2r_3)}$ and $\widehat{\mathbf{U}}_4^{(2r_4)}$.

5. Plug the factor matrices $\widehat{\mathbf{U}}_2^{(2r_2)}$, $\widehat{\mathbf{U}}_3^{(2r_3)}$ and $\widehat{\mathbf{U}}_4^{(2r_4)}$ into the model (3.3), and we can get an estimate of core tensor $\widehat{\boldsymbol{\mathcal{G}}}^{(2r_1,2r_2,2r_3,2r_4)} \in \mathbb{R}^{2r_1\times 2r_2\times 2r_3\times 2r_4}$ using OLS method. Together with $\widehat{\mathbf{U}}_1^{(2r_1)}$, then the fourth-order tensor $\widehat{\boldsymbol{\mathcal{A}}}^{(2r_1,2r_2,2r_3,2r_4)}$ can be easily gotten.

6. Input tensor $\widehat{\boldsymbol{\mathcal{A}}}^{(2r_1,2r_2,2r_3,2r_4)}$, ranks r_1, r_2, r_3, r_4 and iteration, and HOOI is employed to search the factor matrices $\widehat{\mathbf{U}}_1^{(r_1)}$, $\widehat{\mathbf{U}}_2^{(r_2)}$, $\widehat{\mathbf{U}}_3^{(r_3)}$, $\widehat{\mathbf{U}}_4^{(r_4)}$, and $\widehat{\boldsymbol{\mathcal{G}}}^{(r_1,r_2,r_3,r_4)}$. As a result, the fourth-order tensor $\widehat{\boldsymbol{\mathcal{A}}}^{(r_1,r_2,r_3,r_4)}$ can be easily gotten,

7. Let $\widehat{\mathbf{U}}_1 = \widehat{\mathbf{U}}_1^{(r_1)}$, $\widehat{\mathbf{U}}_2 = \widehat{\mathbf{U}}_1^{(r_2)}$, $\widehat{\mathbf{U}}_3 = \widehat{\mathbf{U}}_1^{(r_3)}$, $\widehat{\mathbf{U}}_4 = \widehat{\mathbf{U}}_1^{(r_4)}$ and $\widehat{\boldsymbol{\mathcal{G}}} = \widehat{\boldsymbol{\mathcal{G}}}^{(r_1,r_2,r_3,r_4)}$, and then the fourth-order tensor $\widehat{\boldsymbol{\mathcal{A}}} = \widehat{\boldsymbol{\mathcal{A}}}^{(r_1,r_2,r_3,r_4)}$. Take the initial $\widehat{\boldsymbol{\mathcal{A}}}_0$ to be $\widehat{\boldsymbol{\mathcal{A}}}$.

Moreover, the Tucker ranks of $\boldsymbol{\mathcal{A}}$ are often unknown in real-world applications. They can be chosen empirically or determined using the following high-dimensional BIC:

$$\mathrm{BIC}(\mathbf{r}) = \log\left\{\frac{1}{T}\sum_{n=1}^{T}\|\mathbf{y}_n - \widehat{\boldsymbol{\mathcal{A}}}(\mathbf{r})\mathbf{x}_n\|_2^2\right\} + \frac{\lambda d_{\mathcal{M}}(\mathbf{r})\log(T)}{T}, \tag{3.5}$$

where $\widehat{\boldsymbol{\mathcal{A}}}(\mathbf{r})$ represents $\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}}$ with Tucker ranks $\mathbf{r} = (r_1, r_2, r_3, r_4)$. The term $d_{\mathcal{M}}(\mathbf{r}) = r_1r_2r_3r_4 + Nr_1 + Nr_2 + Sr_3 + Qr_4$ denotes the model complexity, and λ is a tuning parameter.

4. Simulation studies

This section conducts two simulation experiments to evaluate the finite sample performance of MLR-FT-HAR models. The experiments for MLR-TT-HAR models are given at Supplementary material due to the limited space.

The high-frequency data is generated from the following settings. The log prices $\mathbf{p}_t = (p_{1,t}, \ldots, p_{N,t})^\top \in \mathbb{R}^N$ is governed by the standard continuous stochastic process:

$$d\mathbf{p}_t = \boldsymbol{\mu}_t dt + \boldsymbol{\sigma}_t^\top d\mathbf{B}_t, \tag{4.1}$$

where $\boldsymbol{\mu}_t \in \mathbb{R}^N$ is the drift term, $\boldsymbol{\sigma}_t \in \mathbb{R}^{N\times N}$ is the volatility matrix of $\mathbf{p}_t$ which satisfies $\boldsymbol{\gamma}(t) = \boldsymbol{\sigma}_t^T\boldsymbol{\sigma}_t$. We fix the initial log prices $p_{i,0} = \log(10)$ for all $1 \leq i \leq N$, and set the time interval to $\Delta = 1/78$. For the Brownian motions $\mathbf{B}_t$, we let $\mathbf{B}_{t+\Delta} - \mathbf{B}_t$ follow multivariate normal distributions with mean zero and variance matrix $\Delta \cdot \mathbf{I}_N$. For convenience, we let $\boldsymbol{\mu}_t = \mathbf{0}$ and the volatility $\boldsymbol{\sigma}_t$ be a Cholesky decomposition of $\boldsymbol{\gamma}(t)$ with

$$\boldsymbol{\gamma}(t) = (\gamma_{ij}(t)), \quad \gamma_{ij}(t) = \sqrt{\tau_i \tau_j} \kappa^{|i-j|},$$

where $\{\tau_i, i = 1, \ldots, N\}$ are independently generated from a uniform distribution on $[0, 1]$, and κ is set to 0.5. Then, we calculate the realized volatilities $\mathbf{y}_n = (y_{1,n}, \ldots, y_{N,n})^\top$,

$$y_{i,n} = \sum_{j=0}^{m} (p_{t_{i,j}} - p_{t_{i,j-1}})^2 \quad \text{with} \quad 1 \le i \le N \quad \text{and} \quad 1 \le j \le m, \tag{4.2}$$

where $n = 1, \ldots, T$ and $m = 78$ represents the number of high-frequency observations per trading day.

The first experiment aims to validate the non-asymptotic estimation error bound stated in Theorem 1 for MLR-FT-HAR models. The realized volatilities are generated by using the model at (4.1) and (4.2), and the coefficient tensor is generated from (2.1) with $S = 4$ and $Q = 5$. The coefficient tensor $\boldsymbol{\mathcal{A}}$ takes the form $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{G}} \times \mathbf{U}_1 \times \mathbf{U}_2 \times \mathbf{U}_3 \times \mathbf{U}_4 \in \mathbb{R}^{N \times N \times S \times Q}$, where the entries of the core tensor $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{r_1 \times r_2 \times r_3 \times r_4}$ are independently generated from the standard normal distribution and then rescaled such that $\|\boldsymbol{\mathcal{G}}\|_{\mathrm{F}} = 1$. The factor matrices $\mathbf{U}_i$ are generated by extracting the first r_i left singular vectors of Gaussian random matrices, while ensuring the stationary condition in Assumption 1. The error terms $\{\boldsymbol{\varepsilon}_n\}$ follow a multivariate standard normal distribution.

For simplicity, Tucker ranks are set to be equal, i.e., $r_1 = r_2 = r_3 = r_4 = r$. Note that, from Theorem 1, $\|\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}} = O_p(\sqrt{d_{\mathcal{M}}/T})$ with $d_{\mathcal{M}} = r_1 r_2 r_3 r_4 + N r_1 + N r_2 + S r_3 + Q r_4 =$

$r^4 + r(2N + S + Q)$, and hence it is roughly linear with respect to T^{-1} and N, given fixed values of r, S and Q. We consider four settings to verify the relationship: (a) (S, Q, r) is fixed at $(4, 5, 2)$, the dimensionality is limited to $N = 10, 15, 25$, while the sample size T varies among the set of $\{700, 800, 900, 1000, 1100\}$ such that the values of T^{-1} are approximately and evenly spaced from 9×10^{-4} to 14×10^{-4}; (b) (S, Q, r) is fixed at $(4, 5, 3)$ with the dimensionality and sample size the same as in (a); (c) (S, Q, r) is fixed at $(4, 5, 2)$, the sample size is limited to $T = 100, 300$ and 500, while the dimensionality N varies among the set to $\{10, 15, 25, 40, 60\}$ such that the values of $\sqrt{N}$ are approximately and evenly spaced from 3.1 to 7.8; (d) (S, Q, r) is fixed at $(4, 5, 3)$ with the dimensionality and sample size the same as in (c).

Algorithm 1 is employed to search $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}}$ with a step size of 5×10^{-5}, tolerance of 10^{-8}, and initial value $\widehat{\boldsymbol{\mathcal{A}}}_0$ is given by Section 3.3. Both Figure 3 and 4 display the average estimation error $\|\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}} - \boldsymbol{\mathcal{A}}\|_{\text{F}}$ over 500 replications. Figure 3 indicates that $\|\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}} - \boldsymbol{\mathcal{A}}\|_{\text{F}}$ is proportional to $1/\sqrt{T}$, while Figure 4 implies that $\|\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}} - \boldsymbol{\mathcal{A}}\|_{\text{F}}$ is proportional to $\sqrt{N}$, and the theoretical findings in Theorem 1 are hence confirmed. It can also be seen that these lines have different slopes, and it may be due to the fact that these constant terms in the error bound in Theorem 1, such as κ, κ_L, and κ_U, vary for different dimensions of N.

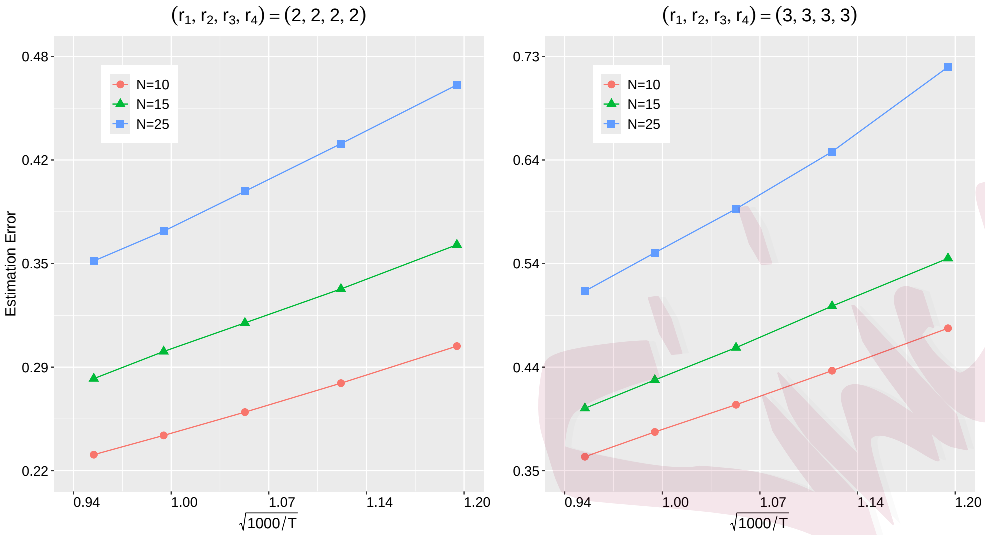

Figure 3: Estimation errors $\|\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}$ against with $\sqrt{1000/T}$. The ranks are $(r_1, r_2, r_3, r_4) = (2, 2, 2, 2)$ in the left panel, and $(3, 3, 3, 3)$ in the right panel.

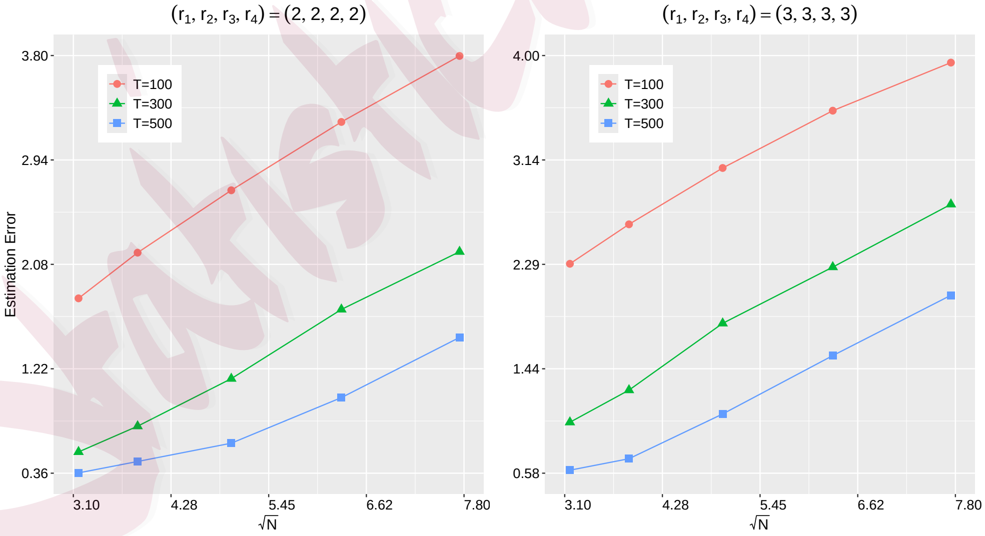

Figure 4: Estimation errors $\|\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}$ against with $\sqrt{N}$. The ranks are $(r_1, r_2, r_3, r_4) = (2, 2, 2, 2)$ in the left panel, and $(3, 3, 3, 3)$ in the right panel.

Moreover, we report the estimation accuracy as the dimensionality N varies among the set $\{10, 15, 25, 40, 60\}$, while T is fixed at 100, 300 and 500, respectively. The estimation accuracy is evaluated by the averaged mean squared error (MSE), calculated as $\|\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}} - \boldsymbol{\mathcal{A}}\|_{\text{F}}^2$ over 500 replications. The corresponding runtime (RT) results of first 50 interations in a single replication are also presented. All the results are summarized in Table 1. It can be seen that the MSEs tend to increase as N increases, while they decrease as T increases. These results further support the theoretical findings in Theorem 1. Moreover, RTs tend to increase as N or T increases, which aligns with the common pattern of algorithms.

Table 1: MSEs and RTs (seconds) with $N = 10, 15, 25, 40, 60$, while holding $T = 100, 300, 500$.

	$(r_1, r_2, r_3, r_4) = (2, 2, 2, 2)$					$(r_1, r_2, r_3, r_4) = (3, 3, 3, 3)$				
N	10	15	25	40	60	10	15	25	40	60
					$T = 100$					
MSE	3.34	4.84	7.37	10.76	14.60	5.29	6.88	9.51	12.63	15.58
RT	0.14	0.22	0.53	1.38	3.74	0.14	0.21	0.54	1.49	4.05
					$T = 300$					
MSE	0.30	0.61	1.36	2.98	4.82	1.02	1.62	3.28	5.16	7.76
RT	0.16	0.26	0.63	1.71	4.69	0.16	0.26	0.62	1.86	4.63
					$T = 500$					
MSE	0.13	0.22	0.38	1.02	2.28	0.37	0.49	1.16	2.41	4.17
RT	0.19	0.29	0.71	1.92	5.04	0.19	0.29	0.70	1.96	5.36

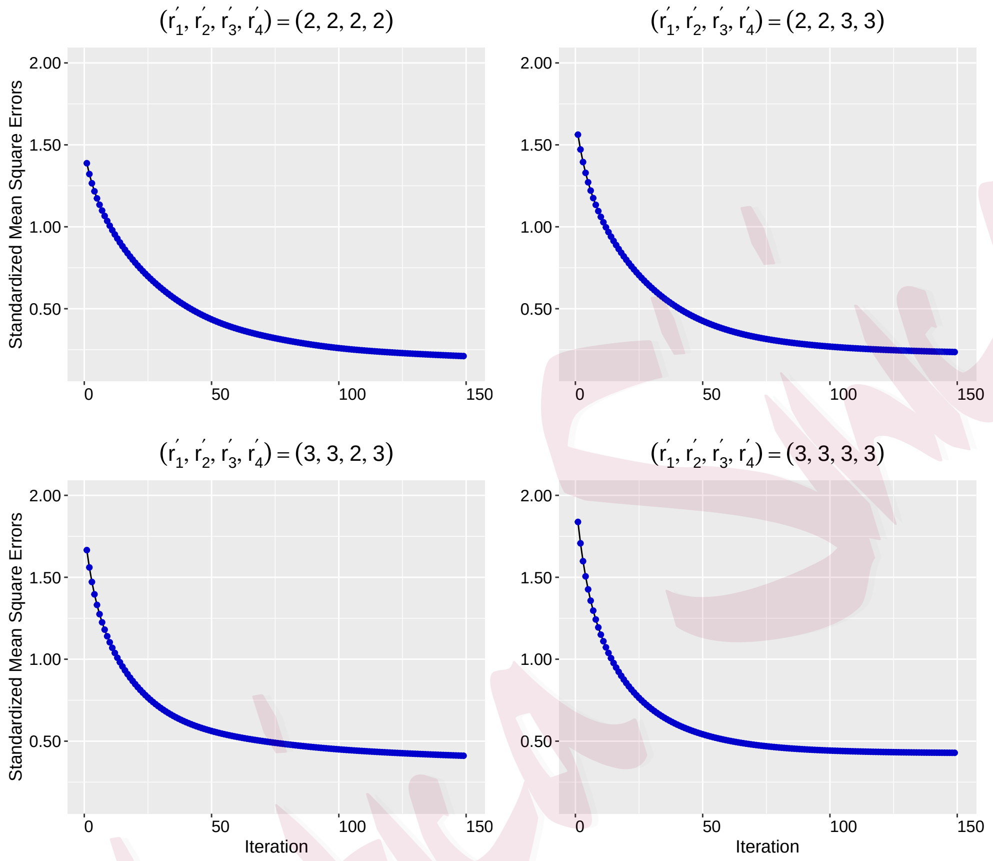

Figure 5: Standardized mean squares errors $\|\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}} - \boldsymbol{\mathcal{A}}\|_{\text{F}}/\|\boldsymbol{\mathcal{A}}\|_{\text{F}}$ for the first 150 iterations with running ranks $(r_1', r_2', r_3', r_4') = (2, 2, 2, 2)$, $(2, 2, 3, 3)$, $(3, 3, 2, 3)$ or $(3, 3, 3, 3)$.

The second experiment aims to assess the convergence performance of Algorithm 1 in Section 3.2 for the MLR-FT-HAR model. The sample points are generated using the same data generation process as in the first experiment, with dimensions $(N, T) = (20, 800)$. The true Tucker ranks are $(r_1, r_2, r_3, r_4) = (2, 2, 2, 2)$. In the algorithm, we consider four different running ranks: $(r_1', r_2', r_3', r_4') = (2, 2, 2, 2)$, $(2, 2, 3, 3)$, $(3, 3, 2, 3)$, and $(3, 3, 3, 3)$. We use a step size of 5×10^{-5}, a tolerance of 10^{-8}, and initial value $\widehat{\boldsymbol{\mathcal{A}}}_0$ in Section 3.3. Figure 5 illustrates the average standardized root mean square errors $\|\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}/\|\boldsymbol{\mathcal{A}}\|_{\mathrm{F}}$ over 500 replications for the first 150 iterations. The plot reveals that all cases exhibit a similar decay pattern, while lower estimation errors can be achieved by specifying more accurate ranks in advance.

5. Real data analysis

This section applies the proposed methodology to analyze the high-frequency trading data of the constituent stocks of the S&P 500 Index with two distinct periods. The long period spans from April 1, 2009, to December 30, 2013, deliberately excluding the first quarter of 2009 to mitigate the potential influence of structural breaks. The short period encompasses January 1, 2011, to December 30, 2013. Our analysis focuses on a selected set of either $N = 60$ or 90 stocks, chosen based on their highest trading volumes on January 2, 2013. The data from 2013 is used to evaluate the out-of-sample performance of our method in both periods. Consequently, we have $T = 937$ days available for estimation during the long period and 497 days for estimation during the short period. Additionally, we have $M = 249$ days dedicated to prediction.

The daily trading data, obtained from the Wharton Research Data Services, covers the time interval from 9:30 am to 4:00 pm. We utilize the commonly employed five-minute returns in the literature (Liu et al., 2015), which leads to a total of $m = 78$ intraday observations. Overnight returns are excluded from our analysis due to their tendency to contain jumps influenced by external factors. To estimate the integrated volatility, we employ the realized measure, RV, which is first transformed into a logarithmic form and then centered to yield a mean of zero. This process ensures that the data aligns with standard volatility estimation techniques.

For the sake of comparison, we evaluate the performance of four models: the proposed MLR-FT-HAR, MLR-TT-HAR, vector HAR (VHAR), and vector HAR-index (VHARI) models. The MLR-FT-HAR and MLR-TT-HAR models allow for the selection of volatility components in the temporal direction, whereas the latter two models lack this flexibility and utilize predetermined volatility components. We explore the volatility components of the MLR-FT-HAR with order $Q = 22$ and $S = 3$, and the MLR-TT-HAR, VHAR, and VHARI models with orders either $P = 22$ or $P = 66$.

The VHAR model does not involve any additional dimension reduction, while the VHARI model assumes a low rank of $r < N$ for the row space of the coefficient matrices. Consequently, the VHAR, VHARI, MLR-TT-HAR and MLR-FT-HAR models have numbers of parameters given by $3N^2$, $4Nr - r^2$, $r_1r_2r_3 + Nr_1 + Nr_2 + Pr_3$, and $r_1r_2r_3r_4 + Nr_1 + Nr_2 + Sr_3 + Qr_4$, respectively.

To determine the Tucker ranks of the MLR-FT-HAR model, we use BIC at (3.5) with a tuning parameter $\lambda = 0.5 \times 10^{-4}$. The ranks are selected within the ranges $1 \leq r_1, r_2, r_4 \leq 10$ and $1 \leq r_3 \leq 3$. The modified BIC is also employed to select the ranks of the MLR-TT-HAR and VHARI models, and the results are given in Supplementary Material.

To evaluate the out-of-sample performance of the four models, a rolling window forecast procedure is employed. The historical data's ending point iterates within the out-of-sample period of 2013, with a fixed window size of $T = 937$ or 497. At each iteration, a one-step ahead prediction is conducted. The ranks of the MLR-FT-HAR and MLR-TT-HAR, and VHARI models are fixed during forecasting. The popular empirical quasi-likelihood (QLIKE) is used to evaluate the forecasting accuracy,

$$\text{QLIKE}_i = \frac{1}{M}\sum_{n=1}^{M}\left(\frac{y_{i,n}}{\widehat{y}_{i,n}} - \log\left(\frac{y_{i,n}}{\widehat{y}_{i,n}}\right) - 1\right) \quad \text{with} \quad 1 \le i \le N,$$

where $\widehat{y}_{i,n}$ and $y_{i,n}$ represent the predicted and calculated realized measures, respectively, for the i-th asset on the n-th trading day in 2013. There are $M = 249$ trading days in 2013.

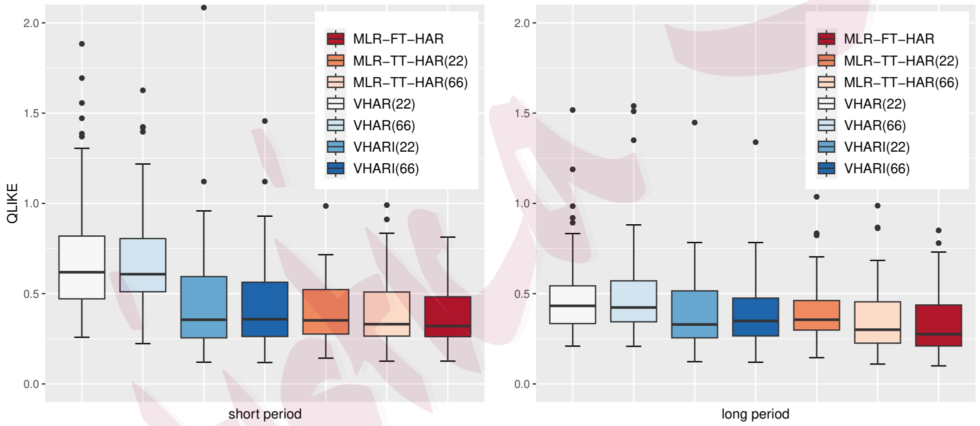

Figure 6: Boxplots of QLIKEs for the VHAR, VHARI and MLR-TT-HAR models with order $P = 22$, $P = 66$, and MLR-FT-HAR model with $Q = 22$ and $S = 3$ for $N = 60$ stocks during the short period $(2011.01 - 2013.12)$ and long period $(2009.04 - 2013.12)$. RV is used as the realized measure.

Figures 6 and 7 illustrate the QLIKE scores for the VHAR, VHARI, MLR-TT-HAR models with order $P = 22$ and $P = 66$, and the MLR-FT-HAR model with $Q = 22$ and $S = 3$ during the short and long periods. The following findings can be observed:

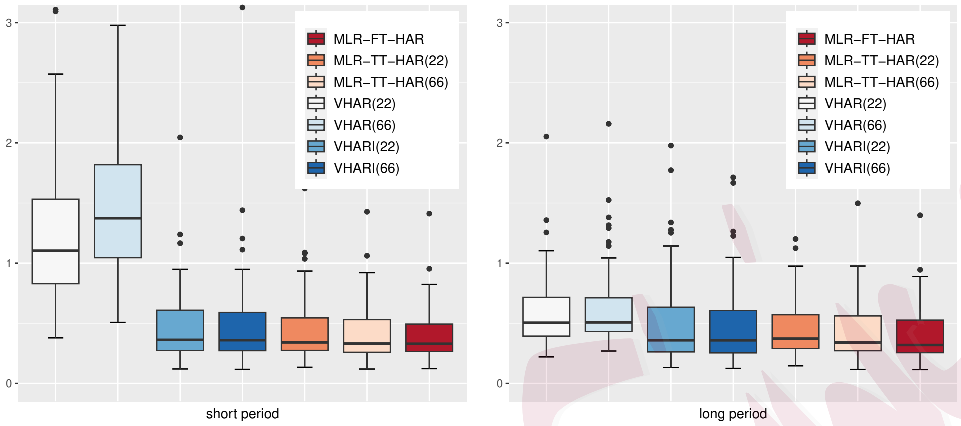

Figure 7: Boxplots of QLIKEs for the VHAR, VHARI and MLR-TT-HAR models with order $P = 22$, $P = 66$, and MLR-FT-HAR model with $Q = 22$ and $S = 3$ for $N = 90$ stocks during the short period $(2011.01 - 2013.12)$ and long period $(2009.04 - 2013.12)$. RV is used as the realized measure.

- The MLR-TT-HAR exhibits better forecasting accuracy compared to the VHAR and VHARI models. This superiority can be attributed to the MLR-TT-HAR models' ability to deduce dimensionality from the response, predictor, and temporal directions. Moreover, the MLR-FT-HAR model outperforms the MLR-TT-HAR model, indicating the importance of exploring low-rank structures from two temporal directions.

- The MLR-TT-HAR model with order $P = 22$ shows no significant difference compared to $P = 66$. This suggests that the longer-term volatility component may not contribute significantly to forecasting realized measures when the temporal direction has not been further split into two directions.

- Comparing Figures 6 and 7, it can be observed that the prediction error generally increases when there are more stocks. This could be attributed to the fact that a larger number of assets (N) implies a more complex model, and it becomes more challenging to capture the relationships among a larger set of stocks. However, among all models, the performance of the MLR-TT-HAR and MLR-FT-HAR models is relatively less influenced by N, indicating that our models can handle a large number of assets simultaneously.

6. Conclusion and discussion

This paper introduces a multilinear low-rank fourth-order tensor HAR model, referred to as the MLR-FT-HAR model, which incorporates the calendar effect in financial markets. Our proposed model decomposes the unified temporal direction into two separate directions: the short temporal direction and the calendar temporal direction. This decomposition effectively reduces the parameter space and allows for the automatic selection of heterogeneous components from both temporal directions. Moreover, when no calendar effect is present, the MLR-FT-HAR model can be simplified into the MLR-TT-HAR model. The MLR-TT-HAR model replaces the fixed heterogeneous volatility components in the HAR model with a data-driven alternative. We establish non-asymptotic properties of the OLS estimator for the MLR-FT-HAR. Additionally, a projected gradient descent algorithm is provided for parameter estimation. The theoretical properties and algorithm for the MLR-TT-HAR model are similarly deduced. An interesting statistical inference on the linear function of the estimator is also provided. A series of simulation experiments is conducted to analyze the finite-sample performance of the MLR-FT-HAR and MLR-TT-HAR models. Furthermore, the real data analysis demonstrates that both the MLR-FT-HAR and MLR-TT-HAR models outperform the VHAR and VHARI models in terms of prediction accuracy. Specifically, the MLR-FT-HAR exhibits better forecasting performance than the MLR-TT-HAR model, highlighting the importance of splitting the temporal direction into two separate directions when forecasting the realized measure.

There are three potential directions for extending the proposed MLR-FT-HAR and MLR-TT-HAR models. Firstly, financial and economic data often exhibit heavy-tailed behaviors (Shin et al., 2023), which deviates from the assumption of sub-Gaussian errors commonly used to derive non-asymptotic properties in the literature. Investigating theoretical properties of high-dimensional modeling under a heavier-tailed assumption is important to make the models more robust and applicable to real-world data. Secondly, the multivariate HAR model has been extended for modeling the realized covariance (Bauer and Vorkink, 2011; Bollerslev et al., 2018), which measures the covariance of high-frequency returns. Extending the proposed methodology to high-dimensional statistical inference for the realized covariance version of the HAR model is an interesting direction. This extension would allow for the modeling and forecasting of high-dimensional covariance matrices, which are crucial in various financial applications. Finally, options-implied information plays a significant role in volatility forecasting. Combining low-frequency, high-frequency, and options data (Yuan et al., 2025) has been a research focus in constructing volatility models. Discussing high-dimensional modeling for the HAR models with options-implied information would be of practical importance.

Supplementary Material

The online Supplementary Material contains the tensor notations and Tucker decomposition, the proofs of the two theorems, Corollary 1, simulation results for the MLR-TT-HAR model, and one Table for the selected ranks of the MLR-FT-HAR, MLR-TT-HAR and VHARI models in Real data analysis.

Acknowledgments

We are deeply grateful to the editor, an associate editor and two anonymous referees for their valuable comments that led to the substantial improvement of the manuscript. Yuan's research was partially supported by National Natural Science Foundation of China (No. 72403089), State Key Program of National Natural Science Foundation of China (No. 72531003), 2025 Shanghai Bai Yulan Talent Young Program, Shanghai Pujiang Program (No. 23PJ1402400), and China Postdoctoral Science Foundation (No. 2023M741190). Li's research was partially supported by GRF grants 17313722 and 17309625 from the Hong Kong Research Grant Council. Wan's research was partially supported by National Natural Science Foundation of China (No. 72273120). Zhou's research was partially supported by State Key Program of National Natural Science Foundation of China (No. 72531003). Yuan and Lu are co-first authors, and Li and Wan are co-corresponding authors.

References

Agterberg, J. and A. R. Zhang (2024). Statistical inference for low-rank tensors: Heteroskedasticity, subgaussianity, and applications. pp. arXiv:2410.06381v1.

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

Auddy, A. and M. Yuan (2023). Large dimensional independent component analysis: Statistical optimality and computational tractability. pp. arXiv:2303.18156.

Audrino, F. and S. D. Knaus (2016). Lassoing the HAR model: A model selection perspective on realized volatility dynamics. *Econometric Reviews 35*, 1485–1521.

Basu, S. and G. Michailidis (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics 43*, 1535–1567.

Bauer, G. H. and K. Vorkink (2011). Forecasting multivariate realized stock market volatility. *Journal of Econometrics 160*, 93–101.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*, 307–327.

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

Bubák, V., E. Kočenda, and F. Žikeš (2011). Volatility transmission in emerging European foreign exchange markets. *Journal of Banking & Finance 35*, 2829–2841.

Chen, H., G. Raskutti, and M. Yuan (2019). Non-convex projected gradient descent for generalized low-rank tensor regression. *The Journal of Machine Learning Research 20*, 172–208.

Chen, Y., W. K. Härdle, and U. Pigorsch (2010). Localized realized volatility modeling. *Journal of the American Statistical Association 105*, 1376–1393.

Clements, A. and D. P. A. Preve (2021). A practical guide to harnessing the HAR volatility model. *Journal of Banking & Finance 133*, 106285.

Cleveland, W. S. and S. J. Devlin (1980). Calendar effects in monthly time series: Detection by spectrum analysis and graphical methods. *Journal of the American Statistical Association 75*, 487–496.

Cleveland, W. S. and S. J. Devlin (1982). Calendar effects in monthly time series: Modeling and adjustment. *Journal of the American Statistical Association 77*, 520–528.

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

Cubadda, G., B. Guardabascio, and A. Hecq (2017). A vector heterogeneous autoregressive index model for realized volatility measures. *International Journal of Forecasting 33*, 337–344.

De Lathauwer, L., B. De Moor, and J. Vandewalle (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications 21*, 1253–1278.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica 50*, 987–1007.

Engle, R. F. and G. M. Gallo (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics 131*, 3–27.

Ghysels, E., P. Santa-Clara, and R. Valkanov (2006). Predicting volatility: Getting the most out of return data sampled at different frequencies. *Journal of Econometrics 131*, 59–95.

Han, R., R. Willett, and A. R. Zhang (2022). An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics 50*, 1–29.

Han, Y., R. Chen, D. Yang, and C.-H. Zhang (2020). Tensor factor model estimation by iterative projection. pp. arXiv:2006.02611.

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

Hillar, C. J. and L.-H. Lim (2013). Most tensor problems are NP-hard. *Journal of the ACM (JACM) 60*(6), 1–39.

Hong, W. T., J. Lee, and E. Hwang (2020). A note on the asymptotic normality theory of the least squares estimates in multivariate HAR-RV models. *Mathematics 8*, 2083.

Levy, T. and Y. Joseph (2012). The week-of-the-year effect: Evidence from around the globe. *Journal of Banking & Finance 36*, 1963–1974.

Liu, L. Y., A. J. Patton, and K. Sheppard (2015). Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. *Journal of Econometrics 187*, 293–311.

Patton, A. J. and K. Sheppard (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility. *The Reviews of Economics and Statistics 97*, 683–697.

Proietti, T. and D. J. Pedregal (2023). Sensonality in high frequency time series. *Econometrics and Statistics 27*, 62–82.

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

Shin, M., D. Kim, Y. Wang, and J. Fan (2023). Factor and idiosyncratic VAR-Itô volatility models for heavy-tailed high-frequency financial data. *arXiv preprint arXiv:2109.05227*.

Souček, M. and N. Todorova (2013). Realized volatility transmission between crude oil and equity futures markets:

A multivariate HAR approach. *Energy Economics 40*, 586–597.

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

Sullivan, R., A. Timmermann, and H. White (2001). Dangers of data mining: The case of calendar effects in stock returns. *Journal of Econometrics 105*, 249–286.

Taylor, N. (2017). Realised variance forecasting under Box-Cox transformations. *International Journal of Forecasting 33*, 770–785.

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

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge: Cambridge University Press.

Xia, D., A. R. Zhang, and Y. Zhou (2022). Inference for low-rank tensors-no need to debias. *Annuals of Statistics 50*(2), 1220–1245.

Xu, K., E. Chen, and Y. Han (2025). Statistical inference for low-rank tensor models. pp. arXiv:2501.16223v1.

Yuan, H., K. Lu, and G. Li (2025). Volatility analysis with high-frequency and low-frequency historical data, and options-implied information. *Statistica Sinica 35*, 2305–2323.

School of Statistics and Academy of Statistics and Interdisciplinary Sciences, and KLATASDS-MOE, East China Normal University. E-mail: hlyuan@sfs.ecnu.edu.cn; yzhou@fem.ecnu.edu.cn

Department of Statistics & Actuarial Science, The University of Hong Kong. E-mail: neithen@connect.hku.hk; gdli@hku.hk

Department of Decision Analytics and Operations, City University of Hong Kong. E-mail: Alan.Wan@cityu.edu.hk