

Statistica Sinica Preprint No: SS-2025-0010	
Title	HAR-Itô Models and High-dimensional Vector Har Modeling for High-frequency Data
Manuscript ID	SS-2025-0010
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0010
Complete List of Authors	Huiling Yuan, Kexin Lu, Yifeng Guo, Lan Li and Guodong Li
Corresponding Authors	Lan Li
E-mails	ulilan@connect.hku.hk
Notice: Accepted author version.	

HAR-Itô models and high-dimensional vector HAR modeling for high-frequency data

Huiling Yuan¹, Kexin Lu², Yifeng Guo², Lan Li ^{*2}, and Guodong Li²

¹ *Donghua University*, and ² *University of Hong Kong*,

Abstract: It is an important task to model realized volatilities for high-frequency data in finance and economics, while the heterogeneous autoregressive (HAR) model has dominated the applications in this area. However, this model suffers from three drawbacks: (i.) its heterogeneous volatility components are linear combinations of daily realized volatilities with fixed weights, which limits its flexibility for different types of assets, (ii.) it is still unknown what is the high-frequency probabilistic structure for this model, as well as many other HAR-type models in the literature, and (iii.) there are no high-dimensional theoretical results for vector HAR modeling although it is common to encounter many assets in real applications. To overcome these drawbacks, this paper proposes a multilinear low-rank HAR model by using tensor techniques, where a data-driven method is adopted to automatically select the heterogeneous components. In addition, HAR-Itô models are introduced to interpret the corresponding high-frequency dynamics. Moreover, non-asymptotic properties of the high-dimensional vector HAR modeling are established, and a projected gradient descent algorithm with theoretical justifications is suggested to search for estimates. Theoretical and computational properties of the proposed method are verified by simulation experiments and real data analysis.

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

*Corresponding author.

1. Introduction

Volatility analysis is one of the most important tasks in finance and economics (Engle, 1982; Bollerslev, 1986) and, with the widespread availability of high-frequency data, more and more recent discussions have concentrated on modeling realized volatilities, which can be constructed from high-frequency intraday observations. Examples include the realized generalized autoregressive conditional heteroscedastic (GARCH) model (Hansen et al., 2012), high-frequency based volatility model (Shephard and Sheppard, 2010), heterogeneous autoregressive (HAR) model (Corsi, 2009), multiplicative error model (Engle and Gallo, 2006), and mixed data sampling model (Ghysels et al., 2006). As arguably the most popular one among these approaches, the HAR model has a simple cascade structure (Corsi, 2009). Specifically, the short-term, medium-term and long-term volatility components are first identified, and they are usually the linear combinations of daily realized volatilities. We then regress the future realized volatility on the three heterogeneous components, and this leads to an autoregressive form. Surprisingly the simple HAR model can even outperform a few powerful deep neural networks for some real applications in the area of machine learning (Bucci, 2020).

Since its appearance, the HAR model has attracted considerable attention from the literature. First, due to the importance of jump effects for high-frequency data (Andersen et al., 2012), Andersen et al. (2007) included the estimated jumps as the fourth heterogeneous component to improve the forecasting performance; see also Corsi et al. (2010) for threshold bipower variation and Patton and Sheppard (2015) for signed jumps. Corsi and Renò (2012) further added one more component of leverage effects. Secondly, some efforts

have been spent on generalizing HAR models. Bollerslev et al. (2016) suggested a HAR model with time-varying coefficients in terms of parametric forms; see also Bekierman and Manner (2018). Moreover, Chen et al. (2018) considered a similar model with coefficients having a nonparametric form, McAleer and Medeiros (2008) proposed a multiple-regime smooth transition HAR model, and Corsi et al. (2008) considered a more sophisticated HAR model with GARCH errors. The jump effect has also been discussed for these generalized HAR models (Buccheri and Corsi, 2021). Thirdly, besides the ordinary least squares, more efficient estimation methods have been discussed in the literature, including the weighted least squares (Patton and Sheppard, 2015) and robust regression (Clements and Preve, 2021) methods. Fourthly, to achieve higher forecasting accuracy, a transformation is usually employed to realized volatilities before estimation, and one usually chooses the logarithmic (Corsi, 2009) or Box-Cox transformation (Taylor, 2017). Finally, besides the stock market, the HAR model has also been applied to Bitcoin (Trucíos, 2019), energy (Jawadi et al., 2024), agricultural commodities (Luo et al., 2022), and many others.

Most of the above studies on HAR models are limited to the univariate case, while it is common in practice to forecast the realized volatilities of many assets. One popular way is to extend the HAR model to a vector version for the modeling (Hong et al., 2020). For example, Bubák et al. (2011) considered a vector HAR (VHAR) model with multivariate GARCH error terms to improve the forecasting accuracy, and similar model settings can be found in Souček and Todorova (2013). Bauer and Vorkink (2011) conducted the prediction for realized covariances, while the VHAR model plays a key role in the modeling; see also

Bollerslev et al. (2018). This paper develops a VHAR model for predicting realized volatilities across assets, focusing on realized variances rather than the full realized covariance matrix. When the number of assets N is large, the model involves $O(N^2)$ parameters, leading to a high-dimensional estimation problem. As a result, dimension-reduction methods are required to render forecasting feasible. Reinsel (1983) proposed a vector autoregressive model that imposes a low-rank structure on the row space of the coefficient matrices. This approach was extended to the VHAR framework by Cubadda et al. (2017), who introduced the vector HAR-index (VHARI) model. This reduces the number of parameters to $O(N)$, while one may want to know whether the low-rank assumption can be assumed to the column space, or even both row and column spaces, of coefficient matrices.

On the other hand, for most HAR-type models in the above, the three heterogeneous volatility components are set to the daily, weekly and monthly realized volatilities, where the latter two are the simple averages of 5- and 22-day realized volatilities, respectively (Corsi, 2009). However, this setting has been shown to be limited by increasing empirical evidence. Chen et al. (2010) questioned the role of long-term volatility components (measured by monthly realized volatility), arguing that a structural break combined with a low-order autoregressive model can outperform HAR models. Audrino and Knaus (2016) further assessed the three-component structure using LASSO (Tibshirani, 1996) for variable selection in high-order autoregressions, and can be found no supporting evidence from nine US stocks. Note that the three components correspond to the average of j -day realized volatilities with $j = 1, 5$ and 22, respectively. Hong et al. (2020) considered many other combinations of

j 's and empirically showed that the combination of $(1, 5, 22)$ performed worse. In fact, the three components in the classical HAR model can be interpreted as lag-dependent factors, which motivates us to develop a data-driven procedure for selecting these components; see model (3.5) and Remark 3 for details. The first contribution of this paper is to propose a multilinear low-rank HAR model in Section 3 for forecasting realized volatilities using tensor techniques. The low-rank structure is imposed on both the row and column spaces of the coefficient matrices, allowing heterogeneous volatility components to be selected automatically. Our framework is related to Shin et al. (2025) in spirit, but differs fundamentally in that we adopt a low-rank structural assumption rather than sparsity.

Recently some efforts have been spent on constructing high-frequency models for intraday asset prices such that the corresponding low-frequency integrated volatilities have a GARCH-type representation. Examples include the GARCH-Itô model (Kim and Wang, 2016), the factor GARCH-Itô model (Kim and Fan, 2019), the realized GARCH-Itô model (Song et al., 2021), and many others. These GARCH-Itô-type models provide a theoretical bridge to reconcile low-frequency GARCH volatility representations and high-frequency volatility processes, and hence a better interpretation and more reliable inference can be achieved by harnessing GARCH models and realized volatilities. As a natural, yet non-trivial follow-up of the GARCH-Itô literature, the second contribution of this paper is to propose the vector HAR-Itô model in Section 2, and their low-frequency integrated volatility admits a VHAR representation. To the best of our knowledge, this is the first attempt in the literature to explore the high-frequency dynamics for HAR-type models.

In the meanwhile, all currently used HAR modeling tools are for the low-dimensional setting with a fixed number of assets. However, it is common to encounter many financial assets in real applications, and it is urgent to design high-dimensional inference tools for HAR models, where the number of assets N may diverge. The third contribution of this paper is to address this gap by developing a high-dimensional VHAR framework for high-frequency data in Section 4. Specifically, the proposed model resembles a vector autoregressive model with measurement errors, and Section 4.1 establishes non-asymptotic properties of the high-dimensional estimator. Moreover, Section 4.2 proposes a projected gradient descent algorithm with theoretical guarantees.

In addition, Section 5 conducts simulation experiments to evaluate finite-sample performance of the proposed methodology, and its usefulness is further demonstrated by empirical examples in Section 6. Section 7 concludes the paper and discusses future research directions, including extensions to heavy-tailed data. All technical proofs are provided in the Supplementary Material.

Throughout the paper, tensors are denoted by calligraphic capital letters, e.g., $\mathcal{A}, \mathcal{B}$, 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 $\langle \mathbf{a}, \mathbf{b} \rangle = \sum_j a_j b_j$ and $\|\mathbf{a}\|_2 = \sqrt{\langle \mathbf{a}, \mathbf{a} \rangle}$ 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 $\mathbf{A}^\top$, $\text{rank}(\mathbf{A})$, $\sigma_j(\mathbf{A})$, $\|\mathbf{A}\|_{\text{op}} = \sigma_1(\mathbf{A})$, $\|\mathbf{A}\|_{\text{F}} = \sqrt{\sum_{i,j} \mathbf{A}_{ij}^2}$, $\|\mathbf{A}\|_* = \sum_{j=1}^{\min(p_1, p_2)} \sigma_j(\mathbf{A})$, $\|\mathbf{A}\|_\infty = \max_{1 \leq i \leq p_1} \sum_{j=1}^{p_2} |A_{ij}|$, be its transpose, rank, j -th singular value, operator norm, Frobenius, nuclear norm, l_∞ norm,

respectively. For any square matrix $\mathbf{A}$, we let $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ denote its maximum and minimum eigenvalues. Moreover, for any $p_1 \geq p_2$, the set of orthonormal matrices is denoted by $\mathcal{O}^{p_1 \times p_2} := \{\mathbf{O} \in \mathbb{R}^{p_1 \times p_2} \mid \mathbf{O}^\top \mathbf{O} = \mathbf{I}_{p_2}\}$, 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 $a_n \leq Cb_n$, $a_n \gtrsim b_n$ if $a_n \geq Cb_n$, and write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

2. Vector HAR-Itô model for high-frequency data

This section proposes the vector HAR-Itô model for high-frequency asset prices such that its low-frequency integrated volatility has a VHAR representation, as shown in Propositions 1.

Consider N financial assets, and denote by $X_{i,t}$ and $\sigma_{i,t}$ the log price of the i -th asset and its instantaneous volatility at $t \in \mathbb{R}_+ = [0, \infty)$ for each $1 \leq i \leq N$, respectively. Let $\mathbf{X}_t = (X_{1,t}, \dots, X_{N,t})^\top$ and $\boldsymbol{\sigma}_t = (\sigma_{1,t}, \dots, \sigma_{N,t})^\top$. Define two N -dimensional Brownian motions, $\mathbf{B}_t = (B_{1,t}, \dots, B_{N,t})^\top$ and $\mathbf{W}_t = (W_{1,t}, \dots, W_{N,t})^\top$, where $dB_{i,t} \cdot dB_{j,t} = \rho_{i,j}^B dt$, $dW_{i,t} \cdot dW_{j,t} = \rho_{i,j}^W dt$, $dB_{i,t} \cdot dW_{j,t} = \rho_{i,j}^B dt$, $\rho_{i,i}^B = 1$, and $\rho_{i,i}^W = 1$ for all $1 \leq i, j \leq N$. Let $\boldsymbol{\rho}^B = (\rho_{i,j}^B) \in \mathbb{R}^{N \times N}$, $\boldsymbol{\rho}^W = (\rho_{i,j}^W) \in \mathbb{R}^{N \times N}$ and $\boldsymbol{\rho} = (\rho_{i,j}) \in \mathbb{R}^{N \times N}$, where $\boldsymbol{\rho}^B$ and $\boldsymbol{\rho}^W$ are symmetric. Moreover, let $Z_{i,t} = \int_{[t]}^t dW_{i,t}$ for $1 \leq i \leq N$, and denote $\mathbf{Z}_t = (Z_{1,t}, \dots, Z_{N,t})^\top$. For the jump component, let $\boldsymbol{\Lambda}_t = (\Lambda_{1,t}, \dots, \Lambda_{N,t})^\top$ be an N -dimensional Poisson process with constant intensity $(\lambda_1, \dots, \lambda_N)^\top$, and $\mathbf{L}_t = (L_{1,t}, \dots, L_{N,t})^\top$ be N -dimensional jump sizes. We assume that the jump components for different assets are independent, $\{\mathbf{L}_t, t \in \mathbb{R}_+\}$ are independent and identically distributed (*i.i.d.*) with finite fourth moment, $\mathbf{L}_t$ is independent of $\boldsymbol{\Lambda}_t$, $\mathbf{B}_t$, $\mathbf{W}_t$ and $\boldsymbol{\sigma}_t$, and all the above diffusion processes are adapted to the filtration $\{\mathcal{F}_t, t \in \mathbb{R}_+\}$. As a

result, the jump diffusion model can be introduced below,

$$dX_{i,t} = \mu_{i,t}dt + \sigma_{i,t}dB_{i,t} + L_{i,t}d\Lambda_{i,t}, \quad 1 \leq i \leq N. \quad (2.1)$$

Definition 1. The log prices of N assets $\{\mathbf{X}_t, t \in \mathbb{R}_+\}$ at (2.1) are said to follow a vector HAR-Itô model if, for each $1 \leq i \leq N$, the instantaneous volatility of the i -th asset satisfies

$$\begin{aligned} \sigma_{i,t}^2 = & \sigma_{i,[t]}^2 + (t - [t]) \left(\omega_i - \sigma_{i,[t]}^2 \right) + \sum_{j=1}^N \alpha_{i,j}^{(1)} \int_{[t]}^t \sigma_{j,s}^2 ds + \sum_{l=2}^P \sum_{j=1}^N \alpha_{i,j}^{(l)} \int_{[t]-l+1}^{[t]-l+2} \sigma_{j,s}^2 ds \\ & + \beta_i \int_{[t]}^t L_{i,s}^2 d\Lambda_{i,s} + v_i([t] + 1 - t) Z_{i,t}^2, \end{aligned} \quad (2.2)$$

where $[t]$ denotes the integer part of t , i.e., $[t] = n - 1$ for $n - 1 < t \leq n$, P is a positive integer, the parameters of ω_i , β_i , v_i , $\alpha_{i,j}^{(1)}$'s are all positive, and $\alpha_{i,j}^{(l)} \geq 0$ for all $2 \leq l \leq P$.

Let $\boldsymbol{\omega} = (\omega_1, \dots, \omega_N)^\top \in \mathbb{R}^N$, $\boldsymbol{\alpha}_i^{(l)} = (\alpha_{i,1}^{(l)}, \dots, \alpha_{i,N}^{(l)})^\top \in \mathbb{R}^N$, $\boldsymbol{\alpha}^{(l)} = (\boldsymbol{\alpha}_1^{(l)}, \dots, \boldsymbol{\alpha}_N^{(l)})^\top \in \mathbb{R}^{N \times N}$, and $\boldsymbol{\beta} = \text{diag}\{\beta_1, \dots, \beta_N\} \in \mathbb{R}^{N \times N}$. The instantaneous volatility $\boldsymbol{\sigma}_t$ defined at (2.2) is an N -dimensional process with continuous time and, when being restricted to integer time points, it has the form of

$$\boldsymbol{\sigma}_n^2 = \boldsymbol{\omega} + \sum_{l=1}^P \boldsymbol{\alpha}^{(l)} \mathbf{y}_{n-l+1} + \boldsymbol{\beta} \mathbf{J}_n,$$

with probability one, where the integrated volatility and jump variation for the i -th asset are defined as

$$y_{i,n} = \int_{n-1}^n \sigma_{i,t}^2 dt \quad \text{and} \quad J_{i,n} = \int_{n-1}^n L_{i,s}^2 d\Lambda_{i,s},$$

respectively. Let $\mathbf{y}_n = (y_{1,n}, \dots, y_{N,n})^\top$, and $\mathbf{J}_n = (J_{1,n}, \dots, J_{N,n})^\top$. Moreover, from (2.2), the values of $\alpha_{i,j}^{(l)}$ with $2 \leq l \leq P$ measure the contribution of the j -th stock's daily volatility to the i -th stock's spot volatility.

For $1 \leq i \leq N$, let $\omega_{i,L} = E(L_{i,t}^2)$ and $M_{i,t} = L_{i,t}^2 - \omega_{i,L}$ (Song et al., 2021), and denote $\boldsymbol{\omega}_L = \text{diag}\{\omega_{1,L}, \dots, \omega_{N,L}\} \in \mathbb{R}^{N \times N}$, $\mathbf{V} = \text{diag}\{v_1, \dots, v_N\} \in \mathbb{R}^{N \times N}$, and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_N)^\top \in \mathbb{R}^N$. Moreover, let $\boldsymbol{\varrho}_1 = \boldsymbol{\alpha}^{(1)-1}(\mathbf{e}^{\boldsymbol{\alpha}^{(1)}} - \mathbf{I}_N) \in \mathbb{R}^{N \times N}$, $\boldsymbol{\varrho}_2 = \boldsymbol{\alpha}^{(1)-2}(\mathbf{e}^{\boldsymbol{\alpha}^{(1)}} - \mathbf{I}_N - \boldsymbol{\alpha}^{(1)}) \in \mathbb{R}^{N \times N}$, and $\boldsymbol{\varrho}_3 = \boldsymbol{\alpha}^{(1)-3}(\mathbf{e}^{\boldsymbol{\alpha}^{(1)}} - \mathbf{I}_N - \boldsymbol{\alpha}^{(1)} - \frac{\boldsymbol{\alpha}^{(1)2}}{2}) \in \mathbb{R}^{N \times N}$, where $\mathbf{e}^{\mathbf{A}} = \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!} \in \mathbb{R}^{N \times N}$ for a matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$. Denote $\boldsymbol{\omega}^g = \boldsymbol{\varrho}_1 \boldsymbol{\omega} + \boldsymbol{\varrho}_2 \boldsymbol{\beta} \boldsymbol{\omega}_L \boldsymbol{\lambda} + (\boldsymbol{\varrho}_2 - 2\boldsymbol{\varrho}_3) \mathbf{V} \mathbf{1}_N \in \mathbb{R}^N$, $\boldsymbol{\beta}^g = (\boldsymbol{\varrho}_1 - \boldsymbol{\varrho}_2) \boldsymbol{\beta} \in \mathbb{R}^{N \times N}$ and $\mathbf{A}_j = (\boldsymbol{\varrho}_1 - \boldsymbol{\varrho}_2) \boldsymbol{\alpha}^{(j)} \in \mathbb{R}^{N \times N}$ with $1 \leq j \leq P$, where $\mathbf{1}_N = (1, \dots, 1)^\top \in \mathbb{R}^N$. For simplicity, we further define $d\boldsymbol{\Lambda}_t = (d\Lambda_{1,t}, \dots, d\Lambda_{N,t})$ and $\mathbf{Z}_t d\mathbf{Z}_t = (Z_{1,t} dZ_{1,t}, \dots, Z_{N,t} dZ_{N,t})^\top$.

Proposition 1. *Suppose that $n \geq P + 1$, and $\boldsymbol{\alpha}^{(1)}$ has spectral radius less than one. Then the integrated volatility for VHAR-Itô model at Definition 1 satisfies*

$$\mathbf{y}_n = \boldsymbol{\omega}^g + \sum_{j=1}^P \mathbf{A}_j \mathbf{y}_{n-j} + \boldsymbol{\beta}^g \mathbf{J}_{n-1} + \boldsymbol{\varepsilon}_n \quad (2.3)$$

with probability one, where $\boldsymbol{\varepsilon}_n = \boldsymbol{\varepsilon}_n^c + \boldsymbol{\varepsilon}_n^J \in \mathbb{R}^N$,

$$\begin{aligned} \boldsymbol{\varepsilon}_n^c &= 2\mathbf{V} \boldsymbol{\alpha}^{(1)-2} \int_{n-1}^n \left\{ ((n-t)\boldsymbol{\alpha}^{(1)} - \mathbf{I}_N) \mathbf{e}^{(n-t)\boldsymbol{\alpha}^{(1)}} + \mathbf{I}_N \right\} \mathbf{Z}_t d\mathbf{Z}_t \in \mathbb{R}^N, \\ \boldsymbol{\varepsilon}_n^J &= \boldsymbol{\beta} \boldsymbol{\alpha}^{(1)-1} \left\{ \int_{n-1}^n \left(\mathbf{e}^{(n-t)\boldsymbol{\alpha}^{(1)}} - \mathbf{I}_N \right) [\mathbf{M}_t d\boldsymbol{\Lambda}_t + \boldsymbol{\omega}_L (d\boldsymbol{\Lambda}_t - \boldsymbol{\lambda} dt)] \right\} \in \mathbb{R}^N, \end{aligned}$$

and $\boldsymbol{\varepsilon}_n^c$ and $\boldsymbol{\varepsilon}_n^J$ are innovation terms from the continuous diffusion process and the jump component, respectively. Moreover, $\{\boldsymbol{\varepsilon}_n^c\}$ and $\{\boldsymbol{\varepsilon}_n^J\}$ are i.i.d. sequence with mean zero and finite variance matrices.

From Proposition 1, the integrated volatility $\{\mathbf{y}_n\}$ has a representation of vector autoregression with exogenous variable of $\mathbf{J}_{n-1}$. In the absence of jump components, the represen-

tation at (2.3) will reduce to the following VHAR model,

$$\mathbf{y}_n = \boldsymbol{\omega}^g + \sum_{j=1}^P \mathbf{A}_j \mathbf{y}_{n-j} + \boldsymbol{\varepsilon}_n, \quad (2.4)$$

and the integrated volatility $\{\mathbf{y}_n\}$ can be consistently estimated by the realized volatility. Let $\mathbf{y}_n^{(j)} = j^{-1} \sum_{l=1}^j \mathbf{y}_{n+1-l}$ be the average of j -day integrated volatilities, and model (2.4) can be rewritten into $\mathbf{y}_n = \boldsymbol{\omega}^g + \sum_{j=1}^P \boldsymbol{\beta}_j \mathbf{y}_{n-1}^{(j)} + \boldsymbol{\varepsilon}_n$, which is used by Hong et al. (2020) to predict the realized volatility. As a result, the proposed VHAR-Itô model at Definition 1 and Proposition 1 provides a probabilistic support for this VHAR model.

Moreover, the term of $Z_{i,t} = \int_{[t]}^t dW_{i,t}$ plays a key role in designing the instantaneous volatility $\sigma_{i,t}$ at (2.2). We may consider a more general form of $Z_{i,t} = \int_{[t]}^t z_{i,t} dW_{i,t}$ with $z_{i,t}$ being a nonrandom continuous function, and a result similar to Proposition 1 can be established. However, when $z_{i,t}$ is a random function, say $z_{it} = \sigma_{i,t}$ in Kim and Wang (2016), the derived innovations $\{\varepsilon_n^c\}$ will be a martingale difference sequence only although we can still have an autoregressive form similar to (2.3). In fact, by choosing suitable random functions for $z_{i,t}$, the derived innovation ε_n^c can even be heavy-tailed (Shin et al., 2025).

Remark 1. When the number of assets N is large, there may exist a low-dimensional structure in coefficient matrices $\mathbf{A}_j$'s or $\boldsymbol{\alpha}^{(j)}$'s since a market of many assets is usually driven by a few summarized forces (Bai and Ng, 2008; Lam and Yao, 2012). To this end, by using the idea of autoregressive index models (Reinsel, 1983), Cubadda et al. (2017) imposed a low-rank assumption to the following matrix

$$(\mathbf{A}_1^\top, \dots, \mathbf{A}_P^\top) = (\boldsymbol{\alpha}^{(1)\top}, \dots, \boldsymbol{\alpha}^{(P)\top})(\boldsymbol{\varrho}_1^\top - \boldsymbol{\varrho}_2^\top) \in \mathbb{R}^{N \times NP},$$

corresponding to the row space of $\mathbf{A}_j$'s or $\boldsymbol{\alpha}^{(j)}$'s, and a significant improvement of forecasting accuracy can be observed for the realized volatility. Consider another matrix $(\mathbf{A}_1, \dots, \mathbf{A}_P) = (\boldsymbol{\varrho}_1 - \boldsymbol{\varrho}_2)(\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(P)}) \in \mathbb{R}^{N \times NP}$, corresponding to the column space of $\mathbf{A}_j$'s or $\boldsymbol{\alpha}^{(j)}$'s, and the ranks of these two matrices are not equal in general. This motivates us to consider the low-rankness of the column space of $\mathbf{A}_j$'s or $\boldsymbol{\alpha}^{(j)}$'s or even both row and column spaces simultaneously.

3. Multilinear low-rank HAR model

Consider N financial assets, and their log prices $\mathbf{X}_t = (X_{1,t}, \dots, X_{N,t})^\top$ with $t \in \mathbb{R}_+$ are generated by the vector HAR-Itô model at (2.1) and (2.2) with no jump component. Let $\mathbf{y}_n = (y_{1,n}, \dots, y_{N,n})^\top \in \mathbb{R}^N$ be the centered N -dimensional integrated volatility with $\mathbb{E}(\mathbf{y}_n) = 0$, and we next construct a realized measure for it.

For asset i , let $t_{n,k} \in [n-1, n]$ with $0 \leq k \leq m$ be time points with intraday observations, where m is the number of observations within day n , and $n-1 = t_{n,0} < t_{n,1} < \dots < t_{n,m} = n$. The realized volatility can be defined as $RV_{i,n} = \sum_{k=1}^m (X_{i,t_{n,k}} - X_{i,t_{n,k-1}})^2$, and we focus on the centered version, $\tilde{y}_{i,n} = RV_{i,n} - T^{-1} \sum_{l=1}^T RV_{i,l}$, where T is the total number of days. Denote $\tilde{\mathbf{y}}_n = (\tilde{y}_{1,n}, \dots, \tilde{y}_{N,n})^\top \in \mathbb{R}^N$, and we then have

$$\tilde{\mathbf{y}}_n = \mathbf{y}_n + \boldsymbol{\eta}_n, \quad 1 \leq n \leq T, \quad (3.1)$$

where $\boldsymbol{\eta}_n = (\eta_{1,n}, \dots, \eta_{N,n})^\top \in \mathbb{R}^N$ is the estimation error, and the dependence of notations $\tilde{\mathbf{y}}_n$ and $\boldsymbol{\eta}_n$ on m is suppressed for simplicity. Suppose that the number of observed prices m is the same for all $1 \leq n \leq T$ and $1 \leq i \leq N$, and $\{t_{n,k}\}$ are synchronized and equally spaced,

i.e. $t_{n,k} - t_{n,k-1} = m^{-1}$. By a method similar to those in Kim and Wang (2016) and Song et al. (2021), we can show that

$$\max_{1 \leq n \leq T} \max_{1 \leq i \leq N} E(|\eta_{i,n}|^2) \leq O(m^{-1/2}). \quad (3.2)$$

Remark 2. The assumption of synchronization and equal space is for simplicity only, and it can be relaxed to the generalized sampling time (Aït-Sahalia et al., 2010), refresh time (Barndorff-Nielsen et al., 2011) or previous tick (Zhang, 2011). Moreover, besides the realized volatility, we may consider other realized measures, such as the multi-scale realized volatility (Zhang, 2006), kernel realized volatility (Barndorff-Nielsen et al., 2008), and pre-averaging realized volatility (Jacod et al., 2009).

From Proposition 1, the low-frequency integrated volatility satisfies

$$\mathbf{y}_n = \sum_{j=1}^P \mathbf{A}_j \mathbf{y}_{n-j} + \boldsymbol{\varepsilon}_n, \quad P+1 \leq n \leq T, \quad (3.3)$$

where the intercept $\boldsymbol{\omega}^g$ is centered out since $\mathbb{E}(\mathbf{y}_n) = 0$, $\mathbf{A}_j$'s are all $N \times N$ coefficient matrices, and innovations $\{\boldsymbol{\varepsilon}_n\}$ are *i.i.d* with $\boldsymbol{\varepsilon}_n = (\varepsilon_{1,n}, \dots, \varepsilon_{N,n})^\top \in \mathbb{R}^N$, $\mathbb{E}(\boldsymbol{\varepsilon}_n) = 0$ and $\text{var}(\boldsymbol{\varepsilon}_n) < \infty$.

The matrix polynomial for model (3.3) is defined as $\boldsymbol{\mathcal{A}}(z) = \mathbf{I}_N - \mathbf{A}_1 z - \dots - \mathbf{A}_P z^P$, where $z \in \mathbb{C}$ with $\mathbb{C}$ being the complex space.

Assumption 1. The determinant of $\boldsymbol{\mathcal{A}}(z)$ is not equal to zero for all $|z| \leq 1$.

The above assumption is a necessary and sufficient condition for the strict stationarity of a vector autoregression, and it hence ensures that $\{\mathbf{y}_n\}$ is strictly stationary. In the meanwhile, the integrated volatility $\{\mathbf{y}_n\}$ are generated by the Itô process at (2.1) and (2.2) rather than the equation at (3.3), i.e., the sequence $\{\mathbf{y}_n\}$ is not a vector autoregressive process. As a result,

Assumption 1 is not necessary, while the stationarity of $\{\mathbf{y}_n\}$ can simplify the presentation, as well as technical details, dramatically. When Assumption 1 does not hold, we cannot center the integrated volatility since it is no longer stationary, and this problem can be solved by using the original integrated volatility at (3.1) and then including an intercept at (3.3). The corresponding theoretical discussions can be referred to Zheng and Cheng (2021).

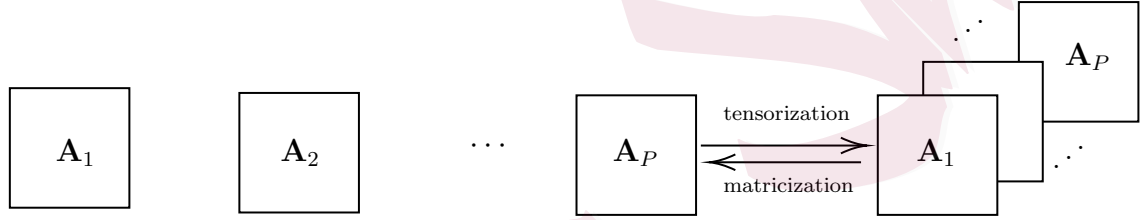


Figure 1: Rearranging $\mathbf{A}_j$'s into a third-order tensor $\mathcal{A}$.

For model (3.3), the number of parameters is N^2P , which can be very large since both N and P may diverge to infinity. This paper uses tensor techniques to conduct dimension reduction for the parameter space. Specifically, the coefficient matrices are first rearranged into a third-order tensor $\mathcal{A} \in \mathbb{R}^{N \times N \times P}$ such that $\mathcal{A}_{(1)} = (\mathbf{A}_1, \dots, \mathbf{A}_P)$; see Figure 1 for an illustration. We then assume multilinear low ranks (r_1, r_2, r_3) to the coefficient tensor $\mathcal{A}$, and accordingly there exists a Tucker decomposition (De Lathauwer et al., 2000),

$$\mathcal{A} = \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3, \quad (3.4)$$

where $\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is 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 factor matrices. We call formulas (3.1) and (3.3), together with the low-rank structure at (3.4), the *multilinear low-rank HAR* (MLR-HAR) model for simplicity.

Note that $\mathcal{A}_{(2)} = (\mathbf{A}_1^\top, \dots, \mathbf{A}_P^\top)$ and $\mathcal{A}_{(3)} = (\text{vec}(\mathbf{A}_1), \dots, \text{vec}(\mathbf{A}_P))^\top$, and the spaces spanned by $\mathcal{A}_{(1)}$, $\mathcal{A}_{(2)}$ and $\mathcal{A}_{(3)}$ are the column, row and temporal spaces of coefficient matrices, respectively. As a result, the low-rank assumption at (3.4) restricts the parameter space from three directions simultaneously, and the number of parameters is reduced to $r_1 r_2 r_3 + N r_1 + N r_2 + P r_3$.

The proposed MLR-HAR model has a form similar to that of vector autoregressive models with measurement errors (Staudenmayer and Buonaccorsi, 2005) and, by plugging (3.3) into (3.1), we have

$$\tilde{\mathbf{y}}_n = \sum_{j=1}^P \mathbf{A}_j \tilde{\mathbf{y}}_{n-j} + \boldsymbol{\epsilon}_n \quad \text{with} \quad \boldsymbol{\epsilon}_n = \underbrace{\boldsymbol{\eta}_n - \sum_{j=1}^P \mathbf{A}_j \boldsymbol{\eta}_{n-j}}_{\text{estimation error}} + \underbrace{\boldsymbol{\varepsilon}_n}_{\text{model error}}, \quad (3.5)$$

where $P+1 \leq n \leq T$. Suppose that the coefficient tensor admits the higher-order singular value decomposition (HOSVD) in Kolda and Bader (2009), $\mathcal{A} = \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$, i.e. $\mathcal{G}$ is all-orthogonal, and $\mathbf{U}_j$'s are orthonormal. Let $\mathcal{H} = \mathcal{G} \times_3 \mathbf{U}_3$, and $\mathbf{H}_j \in \mathbb{R}^{r_1 \times r_2}$ be its j -th frontal slice for $1 \leq j \leq P$, i.e. $\mathcal{H}_{(1)} = (\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_P)$. Thus, $\mathcal{A} = \mathcal{H} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2$, and we can rewrite model (3.5) into

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

where $\mathbf{U}_1^\top \tilde{\mathbf{y}}_n$ and $\mathbf{U}_2^\top \tilde{\mathbf{y}}_{n-j}$ are the summarized factors of responses and predictors, respectively. The HAR model in Cubadda et al. (2017) corresponds to the case with $r_1 = N$ and $\mathbf{U}_1$ being an identity matrix, and it hence has more parameters than the proposed MLR-HAR model; see Remark 1 for more discussions.

Remark 3. Let $\mathcal{S} = \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \in \mathbb{R}^{N \times N \times r_3}$, 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. $\mathcal{S}_{(1)} = (\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_{r_3})$. Denote $\mathbf{U}_3 = (\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(r_3)}) \in \mathbb{R}^{P \times r_3}$ and $\mathbf{u}^{(k)} = (u_1^{(k)}, \dots, u_P^{(k)})^\top \in \mathbb{R}^P$, where $1 \leq k \leq r_3$. As a result, $\mathcal{A} = \mathcal{S} \times_3 \mathbf{U}_3$, and model (3.5) can be reformulated into

$$\tilde{\mathbf{y}}_n = \mathbf{S}_1 \tilde{\mathbf{x}}_n^{(1)} + \dots + \mathbf{S}_{r_3} \tilde{\mathbf{x}}_n^{(r_3)} + \boldsymbol{\epsilon}_n \quad \text{with} \quad \tilde{\mathbf{x}}_n^{(k)} = \sum_{j=1}^P u_j^{(k)} \tilde{\mathbf{y}}_{n-j}. \quad (3.6)$$

Just like the daily, weekly and monthly realized volatilities, $\tilde{\mathbf{x}}_n^{(k)}$'s are the summarized factors along the temporal direction, and they can be treated as r_3 heterogeneous volatility components, which are automatically selected by the estimation method. Moreover, $\mathbf{U}_3$ is the corresponding loading matrix, and model (3.6) will reduce to the VHAR model (Souček and Todorova, 2013) when $\mathbf{U}_3 = \mathbf{U}_C$ at (??).

Remark 4 (Shared dynamic parameters). The proposed MLR-HAR model can be further simplified by assuming that all assets share the same dynamic parameters. Let $\mathbf{a} = (a_1, \dots, a_P)^\top$ and $\mathbf{u} = \mathbf{1}_N / \sqrt{N}$, where $\mathbf{1}_N$ is the N -dimensional vector of ones. We consider the following restriction,

$$\mathbf{A}_j = a_j \mathbf{u} \mathbf{u}^\top = a_j \frac{\mathbf{1}_N \mathbf{1}_N^\top}{N}, \quad j = 1, \dots, P.$$

For $\mathbf{a} \neq \mathbf{0}$, the corresponding coefficient tensor admits a Tucker decomposition of the form (3.4),

$$\mathcal{A} = \mathcal{G} \times_1 \mathbf{u} \times_2 \mathbf{u} \times_3 \frac{\mathbf{a}}{\|\mathbf{a}\|_2},$$

where $\mathcal{G} \in \mathbb{R}^{1 \times 1 \times 1}$ is the core tensor with $\mathcal{G}_{111} = \|\mathbf{a}\|_2$. Thus, $\mathcal{A}$ has multilinear ranks $(1, 1, 1)$, with the two spatial factor matrices fixed as $\mathbf{U}_1 = \mathbf{U}_2 = \mathbf{u}$.

Let $\bar{\tilde{y}}_n = N^{-1} \mathbf{1}_N^\top \tilde{\mathbf{y}}_n$ denote the cross-sectional mean at time n . Under this restriction, model (3.5) can be written as

$$\tilde{y}_{i,n} = \sum_{j=1}^P a_j \bar{\tilde{y}}_{n-j} + \epsilon_{i,n}, \quad i = 1, \dots, N.$$

The multivariate least-squares objective equals N times that for the cross-sectional mean series $\{\bar{\tilde{y}}_n\}$ plus a term independent of $\mathbf{a}$. Hence, the restricted least-squares estimator is given by

$$\hat{\mathbf{a}} = \arg \min_{\mathbf{a} \in \mathbb{R}^P} \sum_{n=1}^T \left(\bar{\tilde{y}}_n - \sum_{j=1}^P a_j \bar{\tilde{y}}_{n-j} \right)^2.$$

As a result, fixing the response and predictor factor matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ at $\mathbf{u}$ reduces the number of dynamic parameters to P , independently of N .

4. High-dimensional vector HAR modeling

4.1 High-dimensional vector HAR modeling

It is common to encounter many assets in real applications, and the number of assets N can be very large, say growing with sample size T at an arbitrary rate. In our context, this corresponds to high-dimensional VHAR modeling, and this section focuses on establishing the non-asymptotic properties of the estimator for MLR-HAR models.

Suppose that we have prior knowledge of multilinear ranks (r_1, r_2, r_3) of the coefficient tensor $\mathcal{A}$. Consider a latent vector time series $\{\mathbf{y}_{-P+1}, \dots, \mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_T\}$, generated by model (3.5), with the low-rank constraint (3.4). We denote the true coefficient matrices by $\mathbf{A}_i^*$ s and the folded true coefficient tensor by $\mathcal{A}^*$. The corresponding parameter space can be defined

as

$$\Theta(r_1, r_2, r_3) = \{\mathcal{A} \in \mathbb{R}^{N \times N \times P} : \text{rank}(\mathcal{A}_{(i)}) \leq r_i \text{ for } 1 \leq i \leq 3\}.$$

This paper considers the ordinary least squares (OLS) estimator for the parameters of MLR-HAR models as follows:

$$\hat{\mathcal{A}}_{\text{MLR}} = \arg \min_{\mathcal{A} \in \Theta(r_1, r_2, r_3)} \frac{1}{T} \sum_{n=1}^T \|\tilde{\mathbf{y}}_n - \mathcal{A}_{(1)} \tilde{\mathbf{x}}_n\|_2^2,$$

where $\tilde{\mathbf{x}}_n = (\tilde{\mathbf{y}}_{n-1}^\top, \dots, \tilde{\mathbf{y}}_{n-P}^\top)^\top$. The estimated coefficient tensor $\hat{\mathcal{A}}_{\text{MLR}}$ can be decomposed into $\hat{\mathcal{A}}_{\text{MLR}} = \hat{\mathcal{G}} \times_1 \hat{\mathbf{U}}_1 \times_2 \hat{\mathbf{U}}_2 \times_3 \hat{\mathbf{U}}_3$ with multilinear ranks (r_1, r_2, r_3) by Tucker decomposition.

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

Assumption 2. Model error $\boldsymbol{\varepsilon}_n = \boldsymbol{\Sigma}_\varepsilon^{1/2} \boldsymbol{\xi}_n$, where $\{\boldsymbol{\xi}_n\}$ are *i.i.d.* random vectors with $\mathbb{E}(\boldsymbol{\xi}_n) = \mathbf{0}$, $\text{var}(\boldsymbol{\xi}_n) = \mathbf{I}_N$, and $\boldsymbol{\Sigma}_\varepsilon = \text{var}(\boldsymbol{\varepsilon}_n)$ is a positive definite matrix. In addition, the entries $(\boldsymbol{\xi}_{i,n})_{1 \leq i \leq N}$ of $\boldsymbol{\xi}_n$ are mutually independent and λ^2 -sub-Gaussian, and model errors $\{\boldsymbol{\varepsilon}_n\}$ are independent of estimation errors $\{\boldsymbol{\eta}_n\}$.

Remark 5. The sub-Gaussianity condition in Assumption 2 is commonly imposed in high-dimensional analysis (Wainwright, 2019). The theoretical results in this section are established under this condition. An extension to heavy-tailed innovations that retains the Tucker structure is discussed in Section 7 as a direction for future research. The assumption of independence between model errors and estimation errors is mainly used to simplify the proofs and can be relaxed at the cost of more involved technical arguments.

We next derive the non-asymptotic error bounds that account for the temporal and cross-sectional dependence of $\{\tilde{\mathbf{y}}_n\}$ (Basu and Michailidis, 2015). We define the following quantities,

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

where the matrix polynomial $\mathcal{A}(z) = \mathbf{I}_N - \mathbf{A}_1 z - \dots - \mathbf{A}_P z^P$, $\bar{\mathcal{A}}(z)$ is the conjugate transpose of $\mathcal{A}(z)$. Moreover, the AR(P) in (3.5) can be alternatively represented as a NP -dimensional AR(1) model with its transition matrix $\mathbf{B}$ defined in the Supplementary Material. We denote the corresponding matrix polynomial by $\mathcal{B}(z) = \mathbf{I}_{NP} - \mathbf{B}z$. We then define

$$\mu_{\min}(\mathcal{B}) := \min_{|z|=1} \lambda_{\min}(\bar{\mathcal{B}}(z)\mathcal{B}(z)) \quad \text{and} \quad \mu_{\max}(\mathcal{B}) := \max_{|z|=1} \lambda_{\max}(\bar{\mathcal{B}}(z)\mathcal{B}(z)),$$

which quantify the smallest and largest eigenvalues that the Hermitian matrix $\bar{\mathcal{B}}(z)\mathcal{B}(z)$ can take on the unit circle. Note that $\mu_{\min}(\mathcal{B})$, $\mu_{\max}(\mathcal{B})$ are not necessarily the same as $\mu_{\min}(\mathcal{A})$ and $\mu_{\max}(\mathcal{A})$ (Basu and Michailidis, 2015), and we can show $\mu_{\min}(\mathcal{B}^*) \leq \mu_{\max}(\mathcal{A}^*)$.

Theorem 1. *Suppose that (3.2) and Assumptions 1 and 2 hold. If the sample size $T \gtrsim \max(\lambda^2, \lambda^4) (\kappa_{U,B}/\kappa_{L,A})^2 d_{\mathcal{M}}$, $m^{1/4} T^{2\delta} \gtrsim N^2 P \exp(d_{\mathcal{M}})$ and $m^{1/4} \gtrsim T^{1+2\delta}$ for some $\delta > 0$, then*

$$\|\hat{\mathcal{A}}_{\text{MLR}} - \mathcal{A}^*\|_{\text{F}} \leq \frac{C}{\kappa_{L,A}} \left[(\lambda^2 \sqrt{\lambda_{\max}(\Sigma_{\varepsilon}) \kappa_{U,A}} + \lambda \sqrt{\kappa_{U,B}}) \sqrt{\frac{d_{\mathcal{M}}}{T}} + \frac{T^{1+2\delta}}{m^{1/4}} \right] \quad \text{and}$$

$$\frac{1}{T} \sum_{n=1}^T \|(\hat{\mathcal{A}}_{\text{MLR}})_{(1)} \tilde{\mathbf{x}}_n - \mathcal{A}_{(1)}^* \tilde{\mathbf{x}}_n\|_2^2 \leq \frac{C}{\kappa_{L,A}} \left[(\lambda^2 \sqrt{\lambda_{\max}(\Sigma_{\varepsilon}) \kappa_{U,A}} + \lambda \sqrt{\kappa_{U,B}}) \sqrt{\frac{d_{\mathcal{M}}}{T}} + \frac{T^{1+2\delta}}{m^{1/4}} \right]^2,$$

with probability at least $1 - \exp(-Cd_{\mathcal{M}}) - 2 \exp(-CT(\kappa_{L,A}/\kappa_{U,B})^2 \min\{\lambda^{-2}, \lambda^{-4}\})$, where $\kappa_{U,A} = \lambda_{\max}(\Sigma_{\varepsilon})/\mu_{\min}(\mathcal{A}^*)$, $\kappa_{L,A} = \lambda_{\min}(\Sigma_{\varepsilon})/\mu_{\max}(\mathcal{A}^*)$, $\kappa_{U,B} = \lambda_{\max}(\Sigma_{\varepsilon})/\mu_{\min}(\mathcal{B}^*)$, $d_{\mathcal{M}} =$

$r_1 r_2 r_3 + Nr_1 + Nr_2 + Pr_3$ is the number of parameters for the MLR-HAR model, λ is sub-Gaussian parameter given in Assumption 2, and C is a positive constant given in the proof.

The above theorem provides the upper bounds of both estimation and prediction errors, and they consist of two terms: the first one is due to the OLS estimation error for model (3.3), and the second is caused by using realized volatilities to estimate integrated volatilities at (3.1). The quantities $\kappa_{L,A}$, $\kappa_{U,A}$ and $\kappa_{U,B}$ in the error bounds characterize the spectral density of $\{\mathbf{y}_n\}$, which describes how the covariance of $\{\mathbf{y}_n\}$ is distributed across frequencies. Specifically, the spectral density of the proposed MLR-HAR model is given by $f(\theta) = (2\pi)^{-1} \mathbf{A}^{-1}(e^{-i\theta}) \boldsymbol{\Sigma}_\varepsilon \bar{\mathbf{A}}^{-1}(e^{-i\theta})$ for $\theta \in [-\pi, \pi]$, and its maximum and minimum eigenvalues satisfy $\lambda_{\max}(f(\theta)) \leq (2\pi)^{-1} \lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon) / \mu_{\min}(\mathbf{A}^*) = \kappa_{U,A} / (2\pi)$ and $\lambda_{\min}(f(\theta)) \geq (2\pi)^{-1} \lambda_{\min}(\boldsymbol{\Sigma}_\varepsilon) / \mu_{\max}(\mathbf{A}^*) = \kappa_{L,A} / (2\pi)$, respectively. When $\kappa_{L,A}$, $\kappa_{U,A}$, $\kappa_{U,B}$ are bounded away from zero and infinity, the estimation error $\|\hat{\mathbf{A}}_{\text{MLR}} - \mathbf{A}^*\|_F = O_P(\sqrt{d_{\mathcal{M}}/T} + T^{1+2\delta}/m^{1/4})$, where $d_{\mathcal{M}}$ measures the complexity of MLR-HAR models, and the term of $T^{1+2\delta}/m^{1/4}$ converges to zero when m diverges with a rate faster than T^4 . Similarly, the prediction error $T^{-1} \sum_{n=1}^T \|(\hat{\mathbf{A}}_{\text{MLR}})_{(1)} \tilde{\mathbf{x}}_n - \mathbf{A}_{(1)}^* \tilde{\mathbf{x}}_n\|_2^2 = O_P(d_{\mathcal{M}}/T + T^{2+4\delta}/m^{1/2})$, and their consistency can be achieved if $m \rightarrow \infty$, $T \rightarrow \infty$, $d_{\mathcal{M}}/T \rightarrow 0$ and $T^{4+8\delta}m^{-1} \rightarrow 0$. In addition, when the order P diverges to infinite, the effective dimension $d_{\mathcal{M}} = r_1 r_2 r_3 + Nr_1 + Nr_2 + Pr_3$ grows accordingly, while the estimation and prediction error bounds remain of the same order.

4.2 Projected gradient descent algorithm

From Subsection 4.1, it is a nonconvex problem to search for the OLS estimator $\hat{\mathcal{A}}_{\text{MLR}}$, and this makes the parameter estimation challenging numerically and theoretically. This subsection introduces a projected gradient descent (PGD) algorithm by adopting the method in Chen et al. (2019), and its theoretical guarantee is also provided.

We first introduce a projection of any tensor $\mathcal{M} \in \mathbb{R}^{N \times N \times P}$ onto the parameter space $\Theta(r_1, r_2, r_3)$. For $1 \leq i \leq 3$, let $\mathcal{M}_i$ be the matricization operator, which maps a tensor to its mode- i matricization, and $\mathcal{M}_i^{-1}$ be the inverse operator, i.e. $\mathcal{M}_i(\mathcal{M}) = \mathcal{M}_{(i)}$ and $\mathcal{M}_i^{-1}(\mathcal{M}_{(i)}) = \mathcal{M}$. Moreover, denote by P_r a projection operator, which maps a matrix to its best rank r approximation. Specifically, for a matrix, P_r first conducts the SVD to it, and then the r largest singular values are kept while the others are suppressed to zero. As a result, for a tensor $\mathcal{M} \in \mathbb{R}^{N \times N \times P}$, we can define its projection onto $\Theta(r_1, r_2, r_3)$ below,

$$\hat{P}_{\Theta(r_1, r_2, r_3)}(\mathcal{M}) := (\mathcal{M}_3^{-1} \circ P_{r_3} \circ \mathcal{M}_3) \circ (\mathcal{M}_2^{-1} \circ P_{r_2} \circ \mathcal{M}_2) \circ (\mathcal{M}_1^{-1} \circ P_{r_1} \circ \mathcal{M}_1)(\mathcal{M}).$$

Specifically, we first calculate mode-1 matricization of $\mathcal{M}$, then find out the best rank r_1 approximation by the SVD, and finally fold it back to a third-order tensor. The same action is further applied to the second and third modes sequentially; see Algorithm 1 for details. The order of which matricization is performed is nonessential, and it will not affect the forthcoming convergence analysis. Moreover, $\hat{P}_{\Theta(r_1, r_2, r_3)}(\cdot)$ is an approximate projection onto $\Theta(r_1, r_2, r_3)$ only, while the exact projection is well known to be an NP-hard problem (Hillar and Lim, 2013).

It is ready to introduce the PGD method to search for the OLS estimator $\hat{\mathcal{A}}_{\text{MLR}}$; see

Algorithm 1. Specifically, we first update the estimate by the commonly used gradient descent method, and the updated tensor is then projected onto $\Theta(r'_1, r'_2, r'_3)$ since it may not have low Tucker ranks. Let (r'_1, r'_2, r'_3) be the running Tucker ranks used in Algorithm 1, and denote $d'_{\mathcal{M}} = (r_1 + r'_1)(r_2 + r'_2)(r_3 + r'_3) + N(r_1 + r'_1) + N(r_2 + r'_2) + P(r_3 + r'_3)$.

Algorithm 1 Projected gradient descent algorithm for HAR modeling

Input : data $\{\tilde{\mathbf{y}}_n\}$, parameter space $\Theta = \Theta(r'_1, r'_2, r'_3)$, iterations K , step size η

Initialize : $k = 0$ and $\hat{\mathcal{A}}_0 \in \Theta$.

for $k = 1, 2, \dots, K$ **do**

$\tilde{\mathcal{A}}_k = \hat{\mathcal{A}}_{k-1} - \eta \nabla L(\hat{\mathcal{A}}_{k-1})$ (gradient descent)

$\hat{\mathcal{A}}_k \leftarrow \tilde{\mathcal{A}}_k$

for $j = 1, 2, 3$ **do**

$\mathcal{M}_{(j)} = \mathcal{M}_j(\hat{\mathcal{A}}_k)$ (mode- j matricization)

$\widehat{\mathcal{M}}_{(j)} = P_{r'_j}(\mathcal{M}_{(j)})$ (best rank r'_j approximation by the SVD)

$\hat{\mathcal{A}}_k = \mathcal{M}_j^{-1}(\widehat{\mathcal{M}}_{(j)})$ (fold into tensor by reversing the mode- j matricization)

end for

end for

Output : $\hat{\mathcal{A}}_K$

Theorem 2. Suppose that (3.2) and Assumptions 1 and 2 hold, step size $\eta = 2/(3\kappa_{U,A})$, and the running Tucker ranks $r'_i \geq \left(\sqrt[3]{1 + \kappa_{L,A}/(24\kappa_{U,A})} - 1\right)^{-2} r_i$ with $1 \leq i \leq 3$. If $T \gtrsim \max(\lambda^2, \lambda^4) (\kappa_{U,B}/\kappa_{L,A})^2 d'_{\mathcal{M}}$, $m^{1/4} T^{2\delta} \gtrsim N^2 P \exp(d'_{\mathcal{M}})$ and $m^{1/4} \gtrsim T^{1+2\delta}$ for some $\delta > 0$,

then

$$\|\widehat{\mathcal{A}}_K - \mathcal{A}^*\|_F \leq \left(1 - \frac{\kappa_{L,A}}{24\kappa_{U,A}}\right)^K \|\widehat{\mathcal{A}}_0 - \mathcal{A}^*\|_F + \frac{C}{\kappa_{L,A}} \left[(\lambda^2 \sqrt{\lambda_{\max}(\Sigma_\varepsilon) \kappa_{U,A}} + \lambda \sqrt{\kappa_{U,B}}) \sqrt{\frac{d'_M}{T}} + \frac{T^{1+2\delta}}{m^{1/4}} \right],$$

with probability at least $1 - \exp(-Cd'_M) - 2 \exp(-CT(\kappa_{L,A}/\kappa_{U,B})^2 \min\{\lambda^{-2}, \lambda^{-4}\})$, where $\kappa_{U,A} = \lambda_{\max}(\Sigma_\varepsilon)/\mu_{\min}(\mathcal{A}^*)$, $\kappa_{L,A} = \lambda_{\min}(\Sigma_\varepsilon)/\mu_{\max}(\mathcal{A}^*)$, $\kappa_{U,B} = \lambda_{\max}(\Sigma_\varepsilon)/\mu_{\min}(\mathcal{B}^*)$, λ is sub-Gaussian parameter given in Assumption 2, and C is a positive constant given in the proof.

The two terms of the upper bound in the above theorem correspond to the optimization and statistical errors, respectively, and the statistical error has a form similar to that in Theorem 1. Note that $\kappa_{L,A} < \kappa_{U,A}$, and hence the linear convergence rate is implied for the optimization error. Specifically, for any $\epsilon > 0$, we can choose the number of iterations $K = \lceil [\log(\epsilon) - \log \|\widehat{\mathcal{A}}_0 - \mathcal{A}^*\|_F] / \log[1 - \kappa_{L,A}/(24\kappa_{U,A})] \rceil$ such that the optimization error is smaller than ϵ . For the initial value $\widehat{\mathcal{A}}_0$, we may simply set it to zero in practice. Finally, the Tucker ranks of $\mathcal{A}$ are typically unknown in practice and can be selected empirically or via a high-dimensional Bayesian information criterion (BIC),

$$\text{BIC}(\mathbf{r}) = \log \left\{ \frac{1}{T} \sum_{n=1}^T \left\| \tilde{\mathbf{y}}_n - (\widehat{\mathcal{A}}(\mathbf{r}))_{(1)} \tilde{\mathbf{x}}_n \right\|_2^2 \right\} + \frac{\bar{\lambda} d_{\mathcal{M}}(\mathbf{r}) \log(T)}{T}, \quad (4.1)$$

where $\widehat{\mathcal{A}}(\mathbf{r}) = \widehat{\mathcal{A}}_{\text{MLR}}$ with Tucker ranks $\mathbf{r} = (r_1, r_2, r_3)$, $d_{\mathcal{M}}(\mathbf{r}) = r_1 r_2 r_3 + N r_1 + N r_2 + P r_3$ denotes the number of free parameters, and $\bar{\lambda}$ is a tuning parameter.

Note that the integrated volatility is positive, however, the vector HAR model at (3.5) does not guarantee its positiveness in forecasting the future values. This paper suggests to model $\log(\mathbf{y}_n)$ (after centering) and then exponentiate the forecasts. Alternatively, in the

empirical analysis, we may add back the mean and truncating on the original scale at a very small value $\varphi > 0$, i.e. $\hat{z}_{i,n} = \max\{\hat{y}_{i,n} + \bar{y}_i, \varphi\}$.

5. Simulations studies

Two simulation experiments are conducted in this section. The first evaluates the finite-sample performance of OLS estimation for the proposed MLR-HAR model under high-dimensional settings in Sections 4.1, and the second examines the convergence of the proposed algorithm in Section 4.2.

The first experiment verifies the non-asymptotic error bound in Theorem 1, which consists of two components: $O_p(\sqrt{d_{\mathcal{M}}/T})$ from the low-frequency model error and $O_p(T^{1+2\delta}/m^{1/4})$ from high-frequency estimation error. Since these two parts are difficult to separate, we design two data-generating processes to evaluate them individually.

For the first process, low-frequency data $\{y_{i,n}\}$ are generated directly from model (3.3) without high-frequency noise, i.e. $m = \infty$, so the error bound reduces to $O_p(\sqrt{d_{\mathcal{M}}/T})$. The coefficient tensor has the form of $\mathcal{A} = \mathbf{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 \in \mathbb{R}^{N \times N \times 22}$, where the core tensor $\mathbf{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ has i.i.d. standard normal entries rescaled to $\|\mathbf{G}\|_F = 0.5$, and the factor matrices $\mathbf{U}_i$'s are formed by the first r_i left singular vectors of Gaussian random matrices, ensuring stationarity (Assumption 1). The error terms $\{\varepsilon_n\}$ follow a multivariate standard normal distribution, with Tucker rank $(r_1, r_2, r_3) = (2, 2, 2)$ or $(3, 3, 3)$. We consider $N = 10, 20, 25$ and $T = 50(4 + 3j + j^2)$ for $1 \leq j \leq 5$, and apply Algorithm 1 to obtain $\hat{\mathcal{A}}_{\text{MLR}}$. Figure 2 reports the average Frobenius error over 500 replications, showing clear linearity in

$\sqrt{d_{\mathcal{M}}/T}$. The error decreases to zero as T increases, confirming the first part of the bound. Differences in slopes arise from dimension-dependent constants in Theorem 1, such as λ , $\kappa_{L,A}$, $\kappa_{U,A}$ and $\kappa_{U,B}$.

For the second data-generating process, we introduce an i.i.d. high-frequency error sequence $\{\boldsymbol{\eta}_n\}$ in (3.1), where $\boldsymbol{\eta}_n \sim \mathcal{N}(0, m^{-1/2} \mathbf{I}_N)$, while $\{\mathbf{y}_n\}$ are generated in the first process. We set $T = 50(1 + j)$ for $1 \leq j \leq 5$, keep other settings unchanged. The realized volatility is computed using $m = T^4$ intraday observations. Note that, from Theorem 1, $\sqrt{T/d_{\mathcal{M}}} \|\hat{\mathbf{A}}_{\text{MLR}} - \mathbf{A}^*\|_{\text{F}} = O_p(1) + O_p(T^{3/2+2\delta}/m^{1/4})$, and hence we plot the adjusted estimation error $\sqrt{T/d_{\mathcal{M}}} \|\hat{\mathbf{A}}_{\text{MLR}} - \mathbf{A}^*\|_{\text{F}}$ against $T^{3/2}/m^{1/4}$ in Figure 2. The clear linear relationship confirms the second part of Theorem 1, while the different intercepts are due to the $O_p(1)$ term.

The second experiment is to evaluate the convergence performance of Algorithm 1. The high-frequency data are generated from the VHAR-Itô model at (2.1) and (2.2) with the absence of jump components and drift terms, i.e. $L_{i,t} = 0$ and $\mu_{i,t} = 0$ with $1 \leq i \leq N$. The dimension is fixed at $N = 30$, and the sample size is $T = 1000$. We set the time interval to $\Delta = 1/78, 1/390, 1/780$, respectively. For the two Brownian motions $\mathbf{B}_t$ and $\mathbf{W}_t$, the increments $\mathbf{B}_{t+\Delta} - \mathbf{B}_t$ and $\mathbf{W}_{t+\Delta} - \mathbf{W}_t$ follow multivariate normal distributions with mean zero, variance matrix $\Delta \cdot \mathbf{I}_N$, and their correlation coefficient matrix being $-0.6 \cdot \mathbf{I}_N$. We fix the initial log prices to $X_{i,0} = \log(50)$ and the initial instantaneous volatility to $\sigma_{i,0} = 0.1$ for all $1 \leq i \leq N$. The model parameters are set to $(\omega_i, v_i) = (0.2, 0.4)$ with $1 \leq i \leq N$, and $(\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(P)}) = \boldsymbol{\alpha}^*(\mathbf{U}_{\text{C}}^{\top} \otimes \mathbf{I}_N) \in \mathbb{R}^{N \times N \times 22}$, where $\mathbf{U}_{\text{C}} \in \mathbb{R}^{22 \times 3}$ is defined in (??),

and $\boldsymbol{\alpha}^* \in \mathbb{R}^{N \times 3N}$ is a randomly generated matrix with rank two and Assumption 1 being satisfied. From Remark 1, we can calculate the coefficient tensor, $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N \times N \times 22}$, of its low-frequency representation, it can be verified to have the Tucker rank of $(r_1, r_2, r_3) = (2, 2, 3)$, and the algorithm considers three different running ranks: $(r'_1, r'_2, r'_3) = (2, 2, 3)$, $(5, 5, 5)$ and $(10, 10, 10)$. The realized volatility is calculated with $m = 78, 390$ or 780 . Figure 3 gives the standardized root mean square errors $\|\hat{\boldsymbol{\mathcal{A}}}_K - \boldsymbol{\mathcal{A}}^*\|_F / \|\boldsymbol{\mathcal{A}}^*\|_F$ for the first 20 iterations over 500 replications. The plot reveals that all cases exhibit a similar decay pattern. Lower estimation errors can be implied by more accurate pre-specified ranks and/or larger numbers of intraday observations.

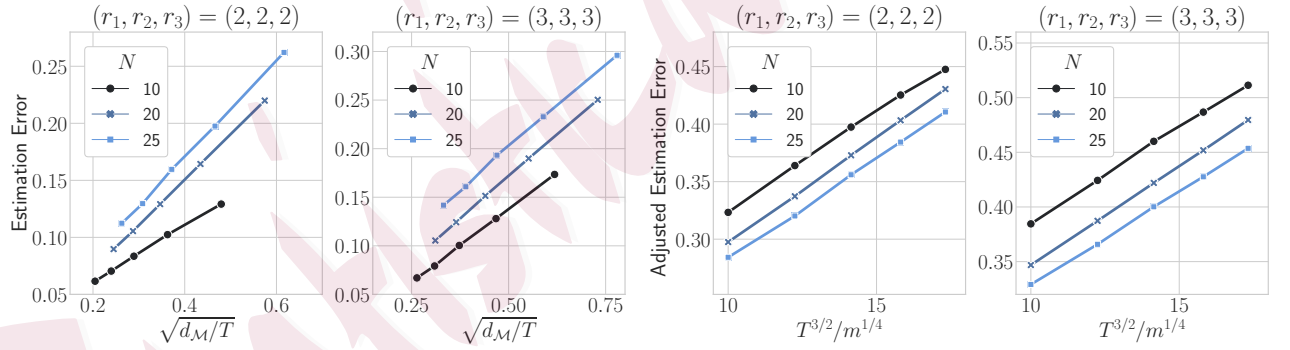


Figure 2: Estimation errors $\|\hat{\boldsymbol{\mathcal{A}}}_{\text{MLR}} - \boldsymbol{\mathcal{A}}^*\|_F$ for the first data generating process with respect to $\sqrt{d_M/T}$ (two left panels), and adjusted estimation errors $\sqrt{T/d_M}\|\hat{\boldsymbol{\mathcal{A}}}_{\text{MLR}} - \boldsymbol{\mathcal{A}}^*\|_F$ from the second data generating process with respect to $T^{3/2}/m^{1/4}$ (two right panels).

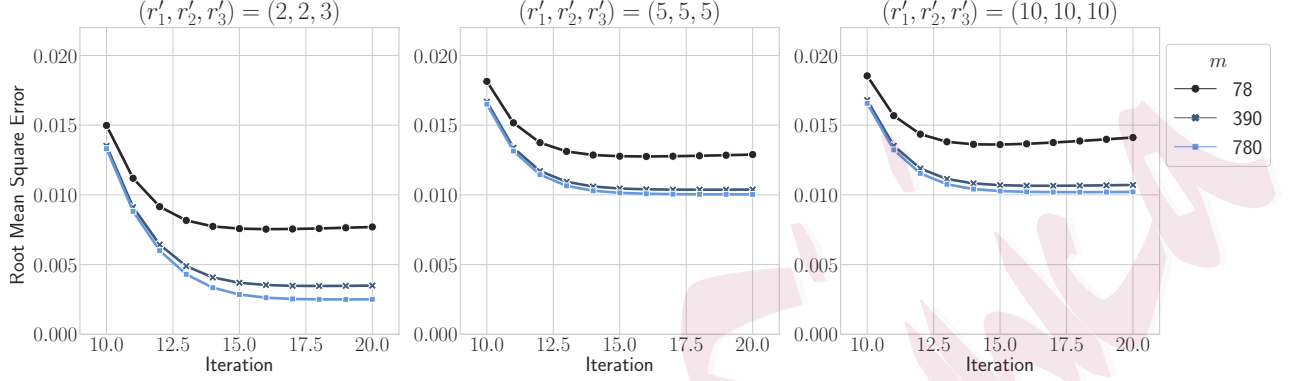


Figure 3: Standardized root mean square errors $\|\hat{\mathcal{A}}_{\text{MLR}} - \mathcal{A}^*\|_{\text{F}} / \|\mathcal{A}^*\|_{\text{F}}$ for the first 20 iterations with running ranks $(r'_1, r'_2, r'_3) = (2, 2, 3)$, $(5, 5, 5)$ and $(10, 10, 10)$.

6. Real data analysis

This section evaluates the empirical performance of the proposed MLR-HAR model using high-frequency data on constituent stocks of the S&P 500 index. The analysis focuses on the $N = 150$ or 229 stocks with the largest trading volumes during 2017–2024. We consider three rolling-window forecasting settings. The first uses 2017–2018 as the initial estimation period and evaluates forecasts over 2019, with a window length of $T = 502$ trading days. The other two settings evaluate forecasts over 2024, using 2020–2023 and 2017–2023 as the initial estimation periods, with window lengths of $T = 1006$ and $T = 1760$, respectively. At each forecast origin, the estimation window advances by one trading day while retaining its fixed length. Therefore, the two 2024 settings compare forecasts based on different estimation histories over the same evaluation period.

The daily trading data, obtained from Wharton Research Data Services, span 9:30 am to

4:00 pm. Following Liu et al. (2015), we use five-minute returns, yielding $m = 78$ intraday observations, and overnight returns are excluded due to their susceptibility to jumps driven by external factors. Integrated volatility is estimated by realized variance (RV) and the jump-robust median realized volatility (medRV) of Andersen et al. (2012). Both measures are log-transformed and centered before estimation. In this section, $\tilde{y}_{i,n}$ denotes the centered log-realized variance, and $\hat{y}_{i,n}$ denotes its one-step-ahead forecast.

The proposed MLR-HAR model with order $P = 22$ is compared with two benchmarks: the VHAR and VHARI models. While the competitors fix volatility components to daily, weekly, and monthly measures, our model selects them automatically. We also consider $P = 66$ to examine potential longer-term components. The VHAR model uses fixed lag aggregation but leaves response and predictor ranks unrestricted, whereas VHARI assumes a low rank $r < N$ in the row space of the coefficient matrices. Accordingly, the numbers of parameters are $3N^2$, $4Nr - r^2$ and $r_1r_2r_3 + Nr_1 + Nr_2 + Pr_3$ for the VHAR, VHARI and MLR-HAR models, respectively, indicating that our model requires substantially fewer parameters than VHAR. To determine the Tucker ranks of MLR-HAR model, we use BIC at (4.1) with a tuning parameter $\lambda = 0.5 \times 10^{-4}$. The ranks are selected within the ranges $1 \leq r_1, r_2 \leq 10$ and $1 \leq r_3 \leq 3$. The modified BIC is also employed to select the ranks of the VHARI models, with the results reported in the Supplementary Material.

Out-of-sample performance is evaluated using a rolling-window forecasting procedure. The estimation window moves sequentially over the evaluation period with fixed window sizes $T = 502$ for the 2019 evaluation period, and $T = 1006$ or 1760 for the 2024 evaluation

period. At each step, a one-step-ahead forecast is generated using the rolling estimation window described above. The selected ranks in the MLR-HAR and VHARI models are fixed throughout the forecasting exercise. Forecast accuracy is assessed using the quasi-likelihood (QLIKE) loss on the variance scale,

$$\text{QLIKE}_i = \frac{1}{M} \sum_{n=1}^M [\exp(\tilde{y}_{i,n} - \hat{y}_{i,n}) - (\tilde{y}_{i,n} - \hat{y}_{i,n}) - 1], \text{ with } 1 \leq i \leq N,$$

where $\tilde{y}_{i,n}$ and $\hat{y}_{i,n}$ denote the observed and predicted centered log-realized variances, respectively, and $M = 252$ is the number of evaluation days. The centering constant for each asset cancels in the ratio of variance levels.

Figures 4 and 5 present the QLIKE losses for VHAR, VHARI, MLR-HAR(22), and MLR-HAR(66). The main findings are as follows. (i.) The MLR-HAR generally yields lower median QLIKE losses than VHAR and VHARI across the three forecasting settings, although the improvements over VHARI are smaller than those over VHAR. (ii.) The two MLR-HAR variants perform similarly. The specification with $P = 66$ does not exhibit systematic gains relative to $P = 22$, suggesting that volatility components beyond the one-month horizon provide limited additional predictive information. (iii.) The relative model ranking remains largely unchanged between the samples with $N = 150$ and $N = 229$. MLR-HAR continues to perform favorably relative to VHAR and VHARI in the larger stock sample. (iv.) The results based on medRV (Panel B), a jump-robust measure of integrated volatility, are qualitatively similar to those based on RV (Panel A). The model ranking is broadly consistent across the two realized measures.

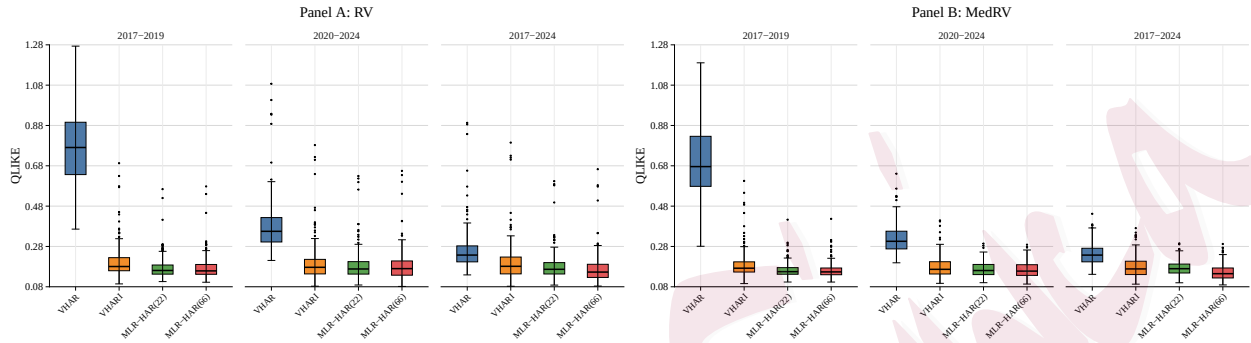


Figure 4: Boxplots of out-of-sample QLIKE losses for $N = 150$ stocks, comparing VHar, VHarI, MLR-HAR(22), and MLR-HAR(66). From left to right, the groups use data from 2017–2019, 2020–2024, and 2017–2024, with the final year of each period used for evaluation. Panel A uses RV and Panel B uses medRV.

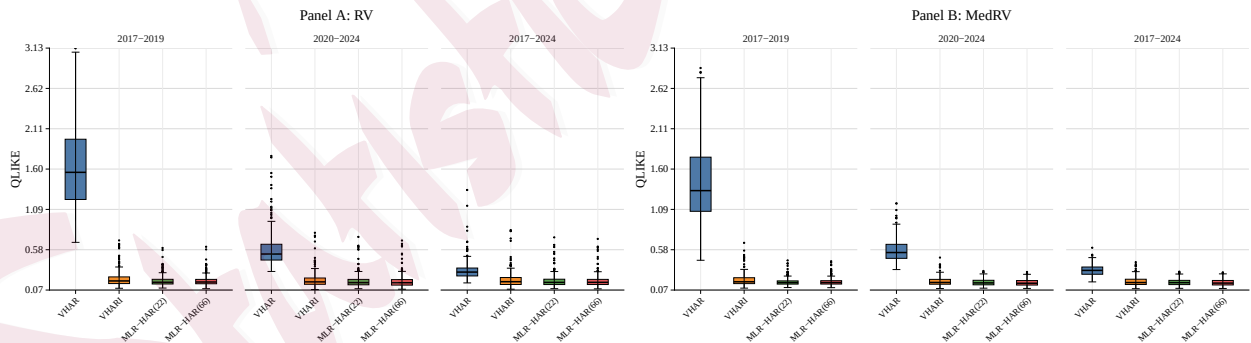


Figure 5: Boxplots of out-of-sample QLIKE losses for $N = 229$ stocks, comparing VHar, VHarI, MLR-HAR(22), and MLR-HAR(66). From left to right, the groups use data from 2017–2019, 2020–2024, and 2017–2024, with the final year of each period used for evaluation. Panel A uses RV and Panel B uses medRV.

The four competing models provide forecasts of individual asset variances. To assess their

implications for portfolio allocation, we use diagonal minimum-variance portfolio weights and an out-of-sample diagonal-risk criterion. This common diagonal covariance approximation avoids introducing an additional model for cross-asset return covariances. Since the models generate forecasts of centered log realized variances, we first reconstruct the corresponding variance levels before forming portfolios. Specifically, the forecasted variance is obtained as $\hat{z}_{i,n} = \exp(\hat{y}_{i,n} + \bar{y}_i)$, and the conditional covariance matrix is approximated by $\hat{\Sigma}_n = \text{diag}(\hat{z}_{1,n}, \dots, \hat{z}_{N,n})$. The diagonal minimum-variance portfolio (DMVP) weights are defined as

$$\hat{\omega}_n = \arg \min_{\omega: \mathbf{1}^\top \omega = 1} \omega^\top \hat{\Sigma}_n \omega = \left(\frac{\hat{z}_{1,n}^{-1}}{\sum_{j=1}^N \hat{z}_{j,n}^{-1}}, \dots, \frac{\hat{z}_{N,n}^{-1}}{\sum_{j=1}^N \hat{z}_{j,n}^{-1}} \right)^\top.$$

Annualized risk is evaluated under the diagonal covariance approximation as follows,

$$\text{AVol} = \sqrt{\frac{252}{M} \sum_{n=1}^M \hat{\omega}_n^\top \Sigma_n \hat{\omega}_n}, \quad \Sigma_n = \text{diag}(\tilde{z}_{1,n}, \dots, \tilde{z}_{N,n}),$$

where $\tilde{z}_{i,n} = \exp(\tilde{y}_{i,n} + \bar{y}_i)$, and M denotes the number of out-of-sample trading days in the evaluation period. Lower values of AVol indicate better performance under the diagonal-risk criterion. Because cross-asset return covariances are excluded from the evaluation, AVol should not be interpreted as the annualized volatility of realized portfolio returns.

Table 1 shows that MLR-HAR achieves larger reductions in the diagonal-risk criterion relative to VHAR, while its gains over VHARI are generally modest. For RV with $N = 229$, the 2019 evaluation gives values of 1.79%, 1.00%, and 0.99% for VHAR, VHARI, and MLR-HAR(22), respectively. Based on the rounded entries, the reduction from VHARI to MLR-HAR(22) is approximately 1.0%. For the 2024 evaluation with $T = 1760$, MLR-HAR(66) attains 1.16%, compared with 1.21% for VHARI, corresponding to a relative reduction of

approximately 4.1%. Several comparisons are tied at the reported precision, and increasing the lag order does not consistently improve performance.

Table 1: Out-of-sample annualized diagonal-risk criterion (%) for VHAR, VHARI, and MLR-HAR. Settings I, II, and III use data from 2017–2019, 2020–2024, and 2017–2024, respectively, with the final year of each period used for evaluation. Panel A uses RV and Panel B uses medRV.

		Panel A: RV				Panel B: medRV			
<i>N</i>	Setting	VHAR	VHARI	MLR-HAR		VHAR	VHARI	MLR-HAR	
		<i>P</i> = 22	<i>P</i> = 22	<i>P</i> = 22	<i>P</i> = 66	<i>P</i> = 22	<i>P</i> = 22	<i>P</i> = 22	<i>P</i> = 66
150	I	1.70	1.21	1.19	1.20	1.45	1.13	1.11	1.11
	II	1.74	1.60	1.59	1.60	1.62	1.50	1.49	1.50
	III	1.50	1.47	1.46	1.47	1.41	1.38	1.37	1.37
229	I	1.79	1.00	0.99	0.99	1.50	0.93	0.92	0.92
	II	1.64	1.31	1.31	1.31	1.49	1.24	1.23	1.24
	III	1.30	1.21	1.21	1.16	1.20	1.13	1.13	1.12

VHARI already reduces the predictor dimension, whereas MLR-HAR also restricts the response space and estimates the lag components from the data. The additional response restriction may improve estimation accuracy when the predictable dynamics across assets can be represented by a small number of common components. Data-driven lag components may be useful when the relevant lag patterns are not adequately represented by fixed daily,

weekly, and monthly averages. However, an overly restrictive response rank may introduce approximation bias, and further reductions in forecast loss may be limited when VHARI already provides an adequate approximation.

Lower QLIKE losses do not necessarily imply proportional reductions in the diagonal-risk criterion. Under the diagonal covariance approximation, portfolio weights depend on relative variance forecasts across assets. In particular, multiplying all variance forecasts at a given date by the same positive constant leaves the weights unchanged. Thus, correcting a common multiplicative scale error can reduce the QLIKE loss without changing the portfolio weights.

As a supplementary descriptive analysis, we examine cross-asset connectedness for $N = 150$ and $N = 229$ using an index constructed from the estimated lag coefficient matrices. This index differs from measures based on forecast error variance decomposition. Its definition and rolling estimates are reported in the Supplementary Material.

7. Conclusion and discussion

This paper extends the HAR framework and contributes to both high-and low-frequency volatility modeling. In the high-frequency setting, we develop vector HAR-Itô models to bridge HAR-type specifications with their corresponding Itô diffusion processes. To the best of our knowledge, this is among the first studies to investigate the high-frequency dynamics of HAR-type models. In the low-frequency setting, we propose a vector low-rank HAR model that imposes low-rank structures simultaneously on the response, predictor, and lag dimensions. This structure substantially reduces the parameter space, allowing the model

to accommodate many assets, and replaces the fixed heterogeneous volatility components in the HAR model with data-driven ones. Remark 4 further illustrates how the framework accommodates a shared-parameter specification with only P dynamic coefficients. We establish non-asymptotic error bounds for high-dimensional VHAR estimation and develop a projected gradient descent algorithm with linear convergence up to statistical error. The empirical analysis evaluates the MLR-HAR model in volatility forecasting and portfolio allocation.

The reductions in the diagonal-risk criterion relative to VHAR are generally modest. The response-space restriction and data-driven lag components may be useful when the predictable dynamics across assets admit a low-dimensional representation and fixed HAR aggregation does not adequately capture the relevant lag patterns. However, an overly restrictive response rank may introduce approximation bias, and further reductions in forecast loss may be limited when VHAR already provides an adequate approximation.

The non-asymptotic results in Section 4 are established under sub-Gaussian innovations. Extending the MLR-HAR framework to heavy-tailed settings is of practical interest (Fan and Kim, 2018; Shin et al., 2023, 2025). One possible approach is to combine a Huber loss with suitably truncated predictors while retaining the Tucker rank constraints. Nuclear norm regularization of a single matricization does not, in general, impose these multilinear constraints. Statistical guarantees would require addressing temporal dependence and high-frequency measurement errors in the presence of heavy-tailed innovations. The associated nonconvex optimization problem also requires further study. We leave this Tucker-constrained extension for future research.

Other possible extensions are as follows. Firstly, since the HAR model (Corsi, 2009) is motivated by grouping different time horizons into volatility components, one may rearrange $\tilde{\mathbf{y}}_{n-1}, \dots, \tilde{\mathbf{y}}_{n-P}$ at (3.3) into a third-order tensor with modes corresponding to assets, weeks, and months. Tensor methods could then be applied to automatically select medium- and long-term components, leading to a fourth-order coefficient tensor $\mathcal{A}$ and an extension of our high-dimensional framework. Secondly, realized covariances (Bauer and Vorkink, 2011), viewed as matrix-valued time series (Chen et al., 2021), have gained increasing attention. One may similarly connect HAR-covariance representations with Itô diffusions and develop a low-rank HAR-covariance model for forecasting. Finally, neural network approaches such as recurrent neural networks (Hopfield, 1982) and autoencoders (Gu et al., 2021) could be explored as alternatives to the low-rank specification considered here for modeling high-dimensional realized volatility and covariance.

Supplementary Material

The online Supplementary Material includes tensor notation and Tucker decomposition; proofs of Proposition 1 and Theorems 1 and 2, together with the auxiliary lemmas; a description of the univariate HAR-Itô model; a table of selected ranks for the MLR-HAR and VHARI models; and the definition and rolling estimates of the coefficient-based total spillover index for the VHAR, VHARI, and MLR-HAR models.

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. Li's research was partially supported by GRF grants 17313722 and 17309625 from the Hong Kong Research Grant Council.

References

- Aït-Sahalia, Y., J. Fan, and D. Xiu (2010). High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association* 105, 1504–1517.
- Andersen, T. G., T. Bollerslev, and F. Diebold (2007). Roughing it up: Including jump components in the measurement, modeling and forecasting of return volatility. *The Review of Economics and Statistics* 89, 701–720.
- Andersen, T. G., D. Dobrev, and E. Schaumburg (2012). Jump-robust volatility estimation using nearest neighbor truncation. *Journal of Econometrics* 169, 75–93.
- Audrino, F. and S. D. Knaus (2016). Lassoing the HAR model: A model selection perspective on realized volatility dynamics. *Econometric Reviews* 35, 1485–1521.
- Bai, J. and S. Ng (2008). Large dimensional factor analysis. *Foundations and Trends in Econometrics* 3, 89–163.
- Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica* 76, 1481–1536.
- Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2011). Multivariate realised kernels: Consistent

- positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics* 162, 149–169.
- 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.
- Bekierman, J. and H. Manner (2018). Forecasting realized variance measures using time-varying coefficient models. *International Journal of Forecasting* 34, 276–287.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307–327.
- Bollerslev, T., A. J. Patton, and R. Quaedvlieg (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics* 192, 1–18.
- 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.
- Buccheri, G. and F. Corsi (2021). HARK and SHARK: Realized volatility modeling with measurement errors and nonlinear dependencies. *Journal of Financial Econometrics* 19, 614–649.
- Bucci, A. (2020). Realized volatility forecasting with neural networks. *Journal of Financial Econometrics* 18, 502–531.
- 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, R., H. Xiao, and D. Yang (2021). Autoregressive models for matrix-valued time series. *Journal of Economet-*

- rics* 222, 539–560.
- Chen, X. B., J. Gao, D. Li, and P. Silvapulle (2018). Nonparametric estimation and forecasting for time-varying coefficient realized volatility models. *Journal of Business & Economic Statistics* 36, 88–100.
- 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.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7, 174–196.
- Corsi, F., S. Mittnik, C. Pigorsch, and U. Pigorsch (2008). The volatility of realized volatility. *Econometric Reviews* 27, 46–78.
- Corsi, F., D. Pirino, and R. Renò (2010). Threshold bipower variation and the impact of jumps on volatility forecasting. *Journal of Econometrics* 159, 276–288.
- Corsi, F. and R. Renò (2012). Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling. *Journal of Business & Economic Statistics* 30, 368–380.
- 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.
- Fan, J. and D. Kim (2018). Robust high-dimensional volatility matrix estimation for high-frequency factor model. *Journal of the American Statistical Association* 113, 1268–1283.
- 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.
- Gu, S., B. Kelly, and D. Xiu (2021). Autoencoder asset pricing models. *Journal of Econometrics* 222, 429–450.
- 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.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences* 79, 2554–2558.
- Jacod, J., Y. Li, P. A. Mykland, M. Podolskij, and M. Vetter (2009). Microstructure noise in the continuous case: The pre-averaging approach. *Stochastic processes and their applications* 119, 2249–2276.
- Jawadi, F., D. Bourghelle, P. Rozin, I. A. Cheffou, and G. S. Uddin (2024). Sentiment and energy price volatility: A nonlinear high frequency analysis. *Energy Economics* 133, 107465.
- Kim, D. and J. Fan (2019). Factor GARCH-Itô models for high-frequency data with application to large volatility matrix prediction. *Journal of Econometrics* 208, 395–417.
- Kim, D. and Y. Wang (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.
- Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM Review* 51, 455–500.
- Lam, C. and Q. Yao (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics* 40, 694–726.
- 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.
- Luo, J., T. Klein, Q. Ji, and C. Hou (2022). Forecasting realized volatility of agricultural commodity futures with infinite Hidden Markov HAR models. *International Journal of Forecasting* 38, 51–73.
- McAleer, M. and M. C. Medeiros (2008). A multiple regime smooth transition heterogeneous autoregressive model for long memory and asymmetries. *Journal of Econometrics* 147, 104–119.
- 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.
- Reinsel, G. (1983). Some results on multivariate autoregressive index models. *Biometrika* 70, 145–156.
- 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, and J. Fan (2023). Adaptive robust large volatility matrix estimation based on high-frequency financial data. *Journal of Econometrics* 237, 105514.
- Shin, M., D. Kim, Y. Wang, and J. Fan (2025). Factor and idiosyncratic VAR volatility matrix models for heavy-tailed high-frequency financial observations. *Journal of Econometrics* 252.
- Song, X., D. Kim, H. Yuan, X. Cui, Z. Lu, Y. Zhou, and Y. Wang (2021). Volatility analysis with realized GARCH-Itô models. *Journal of Econometrics* 222, 393–410.

-
- 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.
- Staudenmayer, J. and J. P. Buonaccorsi (2005). Measurement error in linear autoregressive models. *Journal of the American Statistical Association* 100, 841–852.
- 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.
- Trucíos, C. (2019). Forecasting Bitcoin risk measures: A robust approach. *International Journal of Forecasting* 35, 836–847.
- Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge: Cambridge University Press.
- Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli* 12, 1019–1043.
- Zhang, L. (2011). Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics* 160, 33–47.
- Zheng, Y. and G. Cheng (2021). Finite-time analysis of vector autoregressive models under linear restrictions. *Biometrika* 108, 469–489.
- School of Mathematics and Statistics, Donghua University. E-mail: hlyuan@dhu.edu.cn (ORCID ID: <https://orcid.org/0000-0003-3557-295X>).
- Department of Statistics & Actuarial Science, The University of Hong Kong. E-mail: neithen@connect.hku.hk ; gyf9712@connect.hku.hk; ulilan@connect.hku.hk; gdli@hku.hk (ORCID ID: <https://orcid.org/0000-0003-3137-8471>).