**Management Science and Engineering** 管理科学与工程, 2023, 12(5), 724-737
Published Online September 2023 in Hans. https://www.hanspub.org/journal/mse
https://doi.org/10.12677/mse.2023.125086

Hans汉斯

# 基于张量CP分解的高频数据波动率矩阵预测

**张晨琳[1]，仝　硕[2]**

[1]南京审计大学经济学院，江苏　南京
[2]南京审计大学统计与数据科学学院，江苏　南京

收稿日期：2023年8月7日；录用日期：2023年8月28日；发布日期：2023年9月14日

## 摘　要

**随着计算机技术的高速发展，高频数据的获取与存储不再是件难事，研究发现由于高频数据波动率矩阵包含了更多的信息，基于其估计的协方差会更加准确。高频数据的使用带来了微观结构噪声影响，并且资产的波动性具有较强的记忆性和持续性，投资者有异质特征，传统波动率矩阵的估计方法效果并不理想。同时，当总体维数超过样本容量时，传统的估计方法会面临维数灾难的问题。与以往预测方法不同，本文利用张量能够存储多维度信息、结构稳定等优点，与HAR模型相结合提出CP-HAR模型预测高频波动率矩阵。该模型构建思路为：首先计算T天的高频波动率矩阵Σ1，Σ2，……，Σ*T*并按天数“堆积”得到一个三阶张量$\mathcal{X} \in R^{T\times I\times J}$，随后对该三阶张量CP分解，对其中刻画时间维度方向的因子矩阵深入探究，即利用HAR模型对其中*T*个时间序列的向量进行动态自回归建模，得到预测矩阵。最后通过得到的预测矩阵与前面CP分解得到的另外两个因子矩阵合并组成新张量$\mathcal{X}_n \in R^{F\times I\times J}$，拆分看为*F*个$I\times J$的高频波动率矩阵即为预测的高频波动率矩阵。实证分析部分，选取沪深300成分股每5分钟高频数据，在资本资产定价Fama-French三因子模型的基础上，利用市值、账面市值比两个具有强解释能力的因子将所有股票分为25组，以每组为单位计算波动率矩阵并通过CP-HAR模型进行高频波动率矩阵的预测，得到65个波动率预测矩阵，并选取常见指标RMSE、MAE、MAPE以及R2评价预测效果。**

## 关键词

**高频数据波动率矩阵，张量CP分解，HAR模型，波动率矩阵，预测效果评价**

# Prediction of High-Frequency Data Volatility Matrix Based on Tensor CP Decomposition

**Chenlin Zhang[1], Shuo Tong[2]**

[1]School of Economics, Nanjing Audit University, Nanjing Jiangsu
[2]School of Statistics and Data Science, Nanjing Audit University, Nanjing Jiangsu

Received: Aug. 7th, 2023; accepted: Aug. 28th, 2023; published: Sep. 14th, 2023

文章引用: 张晨琳, 仝硕. 基于张量 CP 分解的高频数据波动率矩阵预测[J]. 管理科学与工程, 2023, 12(5): 724-737.
DOI: 10.12677/mse.2023.125086

## Abstract

**With the rapid development of computer technology, obtaining and storing high-frequency data is no longer a difficult task. Research has found that due to the volatility matrix of high-frequency data containing more information, the estimated covariance based on it will be more accurate. The use of high-frequency data brings about the impact of microstructure noise, and the volatility of assets has strong memory and persistence. Investors have heterogeneous characteristics, and the estimation methods of traditional volatility matrices are not ideal. At the same time, when the total dimension exceeds the sample size, the traditional estimation methods will face the problem of Curse of dimensionality. Different from previous prediction methods, this paper proposes a CP-HAR model to predict high-frequency volatility matrix by combining the advantages of tensors such as being able to store multi-dimensional information and having a stable structure with the HAR model. The construction idea of this model is as follows: First, calculate the high-frequency volatility matrix Σ1, Σ2, ..., Σ*T* for T days and "stack" it by days to obtain a third order tensor $\mathcal{X} \in R^{T\times I\times J}$ Then, we decompose the third order tensor CP and deeply explore the factor matrix that depicts the direction of the time dimension. That is, we use the HAR model to dynamically autoregressive model the vectors of *T* time series to obtain a prediction matrix. Finally, the prediction matrix obtained is combined with the other two factor matrices obtained from the previous CP decomposition to form a new tensor $\mathcal{X}_n \in R^{F\times I\times J}$ , The high frequency volatility matrix of *F* $I\times J$ is the predicted high frequency volatility matrix. In the empirical analysis section, high-frequency data of 300 constituent stocks in Shanghai and Shenzhen are selected every 5 minutes. Based on the capital asset fixed Fama French three factor model, all stocks are divided into 25 groups using two strong explanatory factors: market value and book to market ratio. The volatility matrix is calculated based on each group and the high-frequency volatility matrix is predicted using the CP-HAR model. 65 volatility prediction matrices are obtained, and common indicators RMSE, MAE, MAPE and R2 evaluate the prediction effect.**

## Keywords

**High Frequency Data Covariance Matrix, Tensor Decomposition, HAR Model, Volatility Matrix, Prediction Effect Evaluation**



## 1. 引言

随着现代化进程与高质量发展的不断推进，我国股票市场也在向好发展与完善，金融领域的相关研究也主要围绕股票价格的波动规律进行分析。Andersen [1] (1998)用高频数据所得的“已实现”波动率代替传统的金融资产日收益率，从而提高对于股票市场波动情况评估的精确性与准确性，Markowitz [2] (1958)将波动率矩阵运用于投资组合的评估研究中，在二维分析中起到了不可或缺的作用。这些研究成果不仅对金融市场的理论研究提供了新的思路和工具，而且对实际投资决策也具有重要的参考价值。在他们研究的基础上，人们对于波动率的探究时刻没有停止，基于高频数据的波动率建模将继续成为金融市场研究中的重要方向之一。

张量分解的应用涉及很多领域，最初 Tucker [3] (1963)将张量相关理论应用于心理测验学，Appellof 和 Davidson [4] (1981)首次将张量分解应用于计算化学中，并在几年之后，将张量分解应用在物理分离技

术中，取得了不错的效果，从这以后，张量分解在化学计量学与物理学领域开始被广泛使用。此后，张量分解涉及到的领域越来越多，比如在代数学中，Knuth [5] (1997)利用张量分解对双线性形式的分解进行研究，在信号处理中，Sidiropoulos、Bro 和 Giannakis [6] (2000)通过张量分解将传感器进行阵列处理，同时与并行因子分析相联系，提高了传感器的识别准确率。在神经学中，Beckmann [7] (2005)在真实的 FMRI 数据上证明了张量 PICA 方法能够提取合理模式，并可以帮助解释和优化组 FMRI 研究。在数据处理中，Acar E [8] (2006)使用多种张量分解方法进行数据挖掘。除此之外，张量分解也应用于图像处理、计算机视觉、金融等领域，基于张量分解的优点与可行性，本文将借助张量分解这一工具探究金融资产的波动率。

## 2. CP-HAR 模型构建

### 2.1. HAR-RV 模型

Muller [9]等(1993)提出了 HAR 模型的理论基础——异质性市场假说，其认为投资者在投资决策中存在着异质性，即不同的投资者在资产的估值、投资时间、持有期限和风险偏好等方面存在差异。在异质性市场假说中，由于投资者存在差异，因此同一资产的价格在不同的投资者之间会存在差异。这种差异可能是由于不同的信息、分析方法或心理因素造成的，也会导致市场价格的不稳定性，同时也为一些投资者创造了利润获得的机会。随着该领域研究的推进，Corsi [10] (2009)基于异质性市场假说提出了异质性自回归模型并研究了其在实现波动率预测中的应用。该模型包括了三个不同频率的收益率序列：高频率、中频率、低频率。高频率收益率是指每日收益率，反映了短期市场波动情况，具有很强的噪声和随机性；中频率收益率指每周或每月的收益率，反映了市场在较短时间内的变化趋势，具有一定的预测能力；低频率收益率是指每季度或每年的收益率，反映市场的长期趋势及结构性变化，具有较强的预测能力。

HAR-RV 模型是一种应用于金融市场波动率预测的模型，其理论基础可以从四个方面来概括：

1) 已实现波动率

HAR-RV 模型基于已实现波动率(Realized Volatility, RV)，即将一段时间内的高频率收益率序列平方求和后除去时间长度得到的波动率指标。相对于传统的波动率指标(如对数收益率的标准差)，RV 更加精确地反映了市场波动率的真实情况，因此被广泛应用于金融市场波动率预测。

2) 异质市场假说理论

HAR-RV 模型还基于异质市场假说理论，即不同投资者在市场上的风险偏好和交易策略不同，导致市场上存在异质性特征。在 HAR-RV 模型中，将不同投资者对市场波动率的贡献看作异质性因素，从而更好地描述了市场波动率的动态特性。

3) 非真实长记忆性模型

HAR-RV 模型还基于非真实长记忆性模型，即利用高阶自回归模型(如 ARFIMA 模型)来描述时间序列中的长程相关性，以更好地预测市场波动率。HAR-RV 模型中的异质性自回归模型是一种具有非真实长记忆性的模型，它考虑了不同频率收益率序列的异质性特征，并可以更好地捕捉市场波动率的长期记忆性。

4) 最后，三种成份的随机可加连串波动的 HAR-RV 模型

基于异质市场的理论，日度已实现波动率通常会受到滞后一期的日度已实现波动率和同期的周度已实现波动率的影响。同样的，周度已实现波动率会受到滞后一期的周度已实现波动率及同期的月度已实现波动率的影响。而月度已实现波动率只受自身滞后一期的月度已实现波动率影响，表达式如下：

$$RV_t^m = \beta^m + \beta_1^m RV_{t-1}^m + \varepsilon_t^m \tag{1}$$

$$RV_t^w = \beta^w + \beta_1^w RV_{t-1}^w + \beta_2^w E_{t-1}\left[RV_t^m\right] + \varepsilon_t^w \tag{2}$$

$$RV_t^d = \beta^d + \beta_1^d RV_{t-1}^d + \beta_2^d E_{t-1}\left[RV_t^w\right] + \varepsilon_t^d \tag{3}$$

式(1)、(2)和(3)相加，同时将同类项合并，即可推出 HAR-RV 的一般形式：

$$RV_{t+1}^d = \beta_0 + \beta_d RV_t^d + \beta_w RV_t^w + \beta_m RV_t^m + \varepsilon_{t+1} \tag{4}$$

其中 $\beta_0$ 为常数项，$RV_t^d$ 是日度已实现波动率，$RV_t^w$ 是周度已实现波动率，$RV_t^m$ 是月度已实现波动率。通常一周交易天数为 5 天，一月交易天数为 22 天，周度已实现波动率和月度已实现波动率计算方式如下：

$$RV_t^w = \frac{1}{5}\left(RV_t^d + \cdots + RV_{t-4}^d\right) \tag{5}$$

$$RV_t^m = \frac{1}{22}\left(RV_t^d + \ldots + RV_{t-21}^d\right) \tag{6}$$

## 2.2. CP-HAR 模型构建流程

研究问题常用的是三阶张量，本文也在三阶张量的基础上构建 CP-HAR 模型。即对于指定的三阶张量 $\tau \in R^{I\times J\times K}$，进行 CP 分解将其分解为有限个秩一张量的和，达到数据降维的效果。对其中我们所需要探究的维度构建矩阵并利用 HAR 模型动态建模去学习观察具体结构。

从本文的角度出发，即用 CP-HAR 模型预测高频波动率矩阵，CP-HAR 模型构建思路如图 1：


**Figure 1.** CP-HAR construction process
**图 1.** CP-HAR 构建流程

具体构建方法如下：

首先，计算每天的高频波动率矩阵 $\Sigma_t$，直到第 $T$ 天的高频波动率矩阵 $\Sigma_T$，将得到的高频波动率矩阵 $\Sigma_1$，$\Sigma_2$，……，$\Sigma_T$ 按天数“堆积”，得到一个三阶张量 $\mathcal{X} \in R^{T\times I\times J}$。

接着，通过交替最小二乘法对得到的三阶张量 CP 分解，可以得到 $\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}$ 三个因子矩阵，若 CP 分解秩为 N，则 $\boldsymbol{A} \in R^{T\times N}$，$\boldsymbol{B} \in R^{I\times N}$，$\boldsymbol{C} \in R^{J\times N}$。若定义 $a_t$ 为矩阵 $\boldsymbol{A}$ 的列向量，则 $a_t \in R^T$ 是 $\boldsymbol{A} \in R^{T\times N}$ 的第 $t$ 列。对刻画时间维度方向的因子矩阵 $\boldsymbol{A}$ 中 $T$ 个时间序列方向向量 $a_t$ 预测研究，即利用 HAR 模型对这 T 个时间序列的向量 $a_t$ 进行动态自回归建模，得到 $F$ 个预测向量 $f_r$，这里的 $F$ 也可看作预测的天数，预测向量 $f_r$ 代表第 $f$ 列的时间序列方向，即是下文式(4.7)中的 $a_{t+1}$。将这 $F$ 个预测向量 $f_r$ 按顺序排列就得到矩阵 $\boldsymbol{D}$，其中 $D \in R^{F\times N}$。

最后，由于张量中任意元素都是可以被表示出来，利用这一性质用预测得到的 $\boldsymbol{D}$ 矩阵与 CP 分解得到的 $\boldsymbol{B}$、$\boldsymbol{C}$ 矩阵作为三个因子矩阵可以通过定义函数合并组成一个新的张量 $\mathcal{X}_n \in R^{F\times I\times J}$，其中矩阵 $\boldsymbol{D}$ 代表新张量的时间序列方向。对于新的张量 $\mathcal{X}_n \in R^{F\times I\times J}$，拆分来看就是 $F$ 个 $I\times J$ 的高频波动率矩阵即为预测的高频波动率矩阵。

以上为 CP-HAR 模型的构建去预测高频波动率矩阵，本文选择每周的交易天数为 5 天，每月的交易天数为 22 天，即可写出 CP-HAR 模型的具体形式：

$$a_{t+1} = \alpha + Q_1 a_t^D + Q_2 a_t^W + Q_3 a_t^M + \varepsilon_{t+1} \tag{7}$$

其中，$a_t$ 为 $\mathcal{X} \in R^{T\times I\times J}$ 通过 CP 分解得到的因子矩阵 A 中第 t 个时间序列方向向量，那么 $a_{t+1}$ 是第 $t+1$ 天

的列向量。$\alpha$ 为列常数项，$Q_1$ 为 1 天的滞后系数矩阵，$Q_2$ 为 5 天的滞后系数矩阵，$Q_3$ 为 22 天的滞后系数矩阵，$\varepsilon_{t+1}$ 为残差。

$a_t^D$ 是第 $t$ 天的列向量，$a_r^W = \frac{1}{5}\left(a_{r-1}^D + a_{r-2}^D + a_{r-3}^D + a_{r-4}^D + a_{r-5}^D\right)$ 为过去 5 天平均列向量值，即代表一周的；$a_r^M = \frac{1}{22}\left(a_{r-1}^D + a_{r-2}^D + \cdots\cdots + a_{r-21}^D + a_{r-22}^D\right)$ 为过去 22 天的平均列向量值，即代表一个月的。

## 3. CP-HAR 模型的计算

通过观察式(7)，可知 CP-HAR 模型的计算重点在于对 $T$ 个时间序列方向向量 $a_t$ 以及对系数 $Q_1$，$Q_2$，$Q_3$ 和常数项的计算。

首先，对时间序列方向向量 $a_t$ 的计算可以转化为对张量 $\mathcal{X} \in R^{T\times I\times J}$ 进行 CP 分解后，得到的其中刻画时间维度方向的因子矩阵的计算，因此需要通过具体求解张量 $\mathcal{X} \in R^{T\times I\times J}$ 的 CP 分解过程来得到我们想要的因子矩阵，本文用矩阵 $\boldsymbol{A}$ 代表时间维度方向的因子矩阵。

CP 分解的求解首先要确定分解的秩 1 张量的个数，而张量的秩 Rank-n 近似无法渐进地得到。所以通过迭代的方法对 $R$ 从 1 开始遍历直到找到一个合适的解。这需要定义一个损失函数 loss，并在循环里套一个比较，即计算第二个 $R$ 的时候的损失函数与第一个结果比较，一直重复下去，寻找一个损失函数值最小时的秩。

确定出分解的秩 1 张量的个数后，本文选取最常用的交替最小二乘法(ALS)求解，具体算法步骤如下(以 $N$ 阶张量为例)：

算法 1：CP 交替最小二乘法(CP-ALS)

1) 输入：$N$ 阶原始张量 $\mathcal{X}$

2) 初始化：$A^{(N)} \in R^{I_n\times R}$，其中 $n=1$，……，$N$

3) 循环 $n=1$，……，$N$

$$V \leftarrow V = A^{(1)T}\left(A^{(1)} \odot \cdots A^{(n-1)} \odot A^{(n+1)} \cdots \odot A^{(N)T}A^{(N)}\right)$$

$$A^{(n)} = X^n\left(A^{(N)} \odot \cdots A^{(n+1)} \odot A^{(n-1)} \cdots \odot A^1\right)V^?$$

得到规范列 $A^{(N)}$ 和 $\lambda$

迭代结束

4) 最终输出：$\lambda$，$A^{(1)}$，$A^{(2)}$，……，$A^{(N)}$

程序结束。

交替优化方法的核心思想是用非负矩阵里面的代价函数来衡量 CP 分解得到结果的相似度，针对本文的具体做法为：上文构建的三阶张量 $\mathcal{X} \in R^{T\times I\times J}$，其分解之后会得到三个因子矩阵，为了找到这些因子矩阵，现找一个秩为 $R$ 的张量 $\widehat{\mathcal{X}}$ 使其尽可能地和 $\mathcal{X}$ 接近，相当于求解无约束问题的式子：

$$\min_{\widehat{\mathcal{X}}} \left\|\mathcal{X} - \widehat{\mathcal{X}}\right\|$$

其中 $\widehat{\mathcal{X}}$ 为：

$$\widehat{\mathcal{X}} = \sum_{r=1}^{R} \lambda_r a_r \circ b_r \circ c_r = \left[\lambda; A, B, C\right] \tag{8}$$

固定其中两个矩阵，相当于回到了常见的线性最小二乘问题，假设矩阵 $\boldsymbol{B}$，$\boldsymbol{C}$ 已经固定，对张量 $\mathcal{X}, \widehat{\mathcal{X}}$ 做 mode-1 展开，即张量在第一个维度展开得到的矩阵，即求：

$$\min_{\widehat{\mathcal{X}}} \left\| \mathcal{X}_{(1)} - \widehat{\mathcal{X}_{(1)}} \right\|$$

其中 $\widehat{\mathcal{X}_{(1)}}$ 的表达式为：

$$\widehat{\mathcal{X}_{(1)}} = A(C \odot B)^T$$

因此原问题变为：

$$\min_{\hat{A}} \left\| \mathcal{X}_{(1)} - \hat{A}(C \odot B)^T \right\|_F$$

计算的是 F-范数，其中 $\hat{A} = A \cdot \mathrm{diag}(\lambda)$ 的最优解为：

$$\hat{A} = \mathcal{X}_{(1)} \left[ (C \odot B)^T \right]^{\dagger} \tag{9}$$

其中符号“†”为 Moore-Penrose 广义逆。

概括来说实现思路为：固定矩阵 $\boldsymbol{B}$、$\boldsymbol{C}$，求解最优解 $\boldsymbol{A}$；接着固定矩阵 $\boldsymbol{A}$、$\boldsymbol{C}$，求解最优解 $\boldsymbol{B}$；最后固定矩阵 $\boldsymbol{A}$、$\boldsymbol{B}$，求解最优解 $\boldsymbol{C}$；并一直重复步骤 1，2，3，直到满足了某个收敛条件算法为止，收敛条件可以是：① 目标函数损失值不再下降或者下降很少。② 因子矩阵不再变化或者变化很小。不过这正是交替最小二乘法相对来说较大的缺点，我们无法保证收敛到全局最小点，也无法保证是否为驻点，只能保证收敛到目标函数不再下降为止。

选取目标损失函数的值不再下降或者下降很少作为收敛条件算法，结合以上算法可以得到时间维度方向的因子矩阵 $\boldsymbol{A} = \begin{bmatrix} a_1 & a_2 \cdots & \cdots & a_T \end{bmatrix}$，若 CP 分解的秩为 $N$，则 $a_t$ 为 $N$ 维的列向量。

具体 CP 分解过程如图 2：

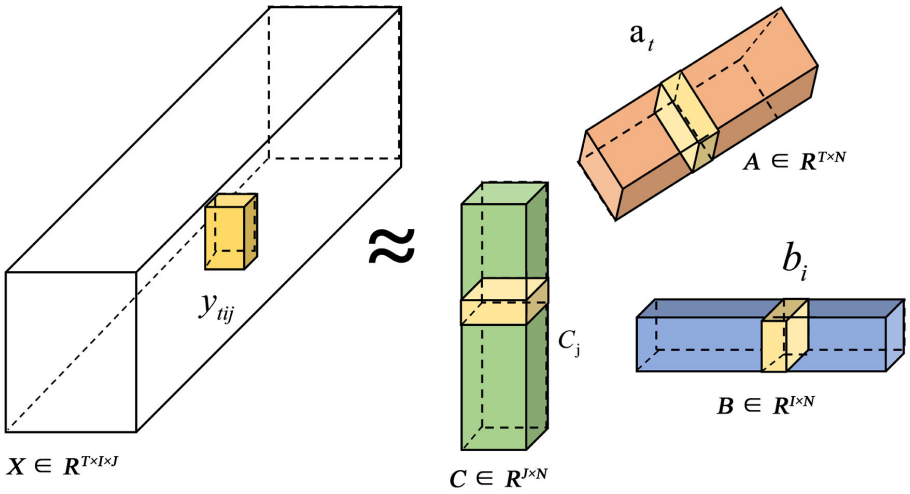


**Figure 2.** CP decomposition process
**图 2.** CP 分解过程

通过上图可以清晰地看出张量 $\mathcal{X} \in R^{T \times I \times J}$ 通过 CP 分解得到了 $\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}$ 三个因子矩阵，若这些因子矩阵的任意元素分别记为 $a_{tr}$，$b_{ir}$，$c_{jr}$，观察上图的黄色填充区域，则任意的元素 $y_{tij}$ 可以写成如下形式：

$$y_{tij} \approx \sum_{r=1}^{N} a_{tr} b_{ir} c_{jr} \tag{10}$$

求出 $T$ 个时间序列方向向量 $a_t$ 之后，接下来重点对式(7)中 $Q_1$，$Q_2$，$Q_3$ 计算。设：

$$Q_1=\begin{bmatrix} q_{1.1} & q_{1.2} & q_{1.3} \\ q_{1.4} & q_{1.5} & q_{1.6} \\ q_{1.7} & q_{1.8} & q_{1.9} \end{bmatrix} Q_2=\begin{bmatrix} q_{2.1} & q_{2.2} & q_{2.3} \\ q_{2.4} & q_{2.5} & q_{2.6} \\ q_{2.7} & q_{2.8} & q_{2.9} \end{bmatrix}$$

$$Q_3=\begin{bmatrix} q_{3.1} & q_{3.2} & q_{3.3} \\ q_{3.4} & q_{3.5} & q_{3.6} \\ q_{3.7} & q_{3.8} & q_{3.9} \end{bmatrix} \boldsymbol{\alpha}=\begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \boldsymbol{\alpha}_3 \end{bmatrix}$$

由于 HAR 类模型需要半年至一年的时间才能描述一个稳定的情况，如用第 1 天到第 40 天的数据预测第 41 天就会不稳定，由于存在 22 个冗余数据，所以本文选择 222 天的列向量 $a_t$ 即 0.8 年作为移动窗口做样本内预测，即用 1 到 222 天的列向量去预测第 223 天的，2 到 223 天的列向量去预测第 224 天的，至到最后一个列向量。这里采用最常用的最小二乘法(OLS)进行估计，求出相对应的参数 $Q$ 以及常数项，一共可以得到 T-222 个预测模型，即得到 $F$ 个预测向量 $f_r$，这里 $F$ 的数值就为 T-222。算法步骤为：

算法 2：时间因子矩阵的 HAR 预测

1) 输入：$T$ 个时间序列方向向量 $a_t$
2) 利用式(5)，(6)计算周列向量 $a_t^W$ (第 5 天始)和月列向量 $a_t^W$ (第 22 天始)，并按日期与日向量对齐排列
3) 分别设 $Q_1$，$Q_2$，$Q_3$ 中各元素分别为 beta1.n，beta2.n，beta3.n；n = 1，……，9，常数项各元素 $\alpha_m$，m = 1, 2, 3
4) 设置移动窗口 0.8 年进行样本内预测，使用最小二乘法对应拟合，计算出各个元素
5) 将计算得到的每天 $Q_1$，$Q_2$，$Q_3$ 以及常数项 $\alpha_m$ 代入公式(4.7)计算，共得到 T-222 个模型
6) 输出：T-222 个 $a_{t+1}$

预测出的 T-222 个 $a_{t+1}$ 即为 $F$ 个预测向量 $f_r$，将这 $F$ 个预测向量 $f_r$ 按天数顺序排列得到矩阵 $\boldsymbol{D}$，其中 $\boldsymbol{D}\in R^{F\times N}$。利用合并函数将 $\boldsymbol{D}$ 矩阵与 $\boldsymbol{B}$、$\boldsymbol{C}$ 矩阵组成一个新的张量 $\mathcal{X}_n\in R^{F\times I\times J}$，具体算法如下：

算法 3：新张量 $\mathcal{X}_n\in R^{F\times I\times J}$ 的组合

1) 输入：矩阵 $\boldsymbol{B}$、矩阵 $\boldsymbol{C}$、矩阵 $\boldsymbol{D}$
2) 定义合并函数 cp_combine，输入为 mat1，mat2，mat3
3) 调用 np 中的 einsum 函数，'ir, jr, tr - > ijt'对输入的 mat1，mat2，mat3 计算
4) 在函数 cp_combine 中 mat1，mat2，mat3 对应位置输入矩阵 $\boldsymbol{D}$，矩阵 $\boldsymbol{B}$，矩阵 $\boldsymbol{C}$
5) 输出：新张量 $\mathcal{X}_n\in R^{F\times I\times J}$

## 4. CP-HAR 模型实际数据分析

基于上文 CP-HAR 模型的构建与计算，本小节进行实际数据的分析。选取沪深 300 指数的成分股每 5 分钟高频数据，来源于 JoinQuant，时间为 2020 年 11 月 2 日至 2021 年 12 月 31 日共 287 个交易日，由于我国股票(证券)市场正常交易日的营业时间为 9:30~11:30、13:00~15:00，共计 4 小时的交易时间，按照 5 分钟的最小交易时间为单位进行划分，每天会得到 48 个有效数据，数据指标主要包括交易的每五分钟时间、交易日、每五分钟的开盘价格、每五分钟的收盘价格、交易数量等等。

首先，对沪深 300 的所有成分股进行筛选，将 2020 年 11 月 2 日至 2021 年 12 月 31 日期间停牌，或是上新等情况的股票，即有数据缺失的股票筛选掉，发现还剩 272 只成分股。把筛选后的 272 只成分股作为研究对象，在资本资产定价 Fama-French 三因子模型的基础上，取市值和账面市值比两个重要的解释因子，等比各分成为五份将所有的股票分为 25 组，以组为单位进行研究，通过市值加权的收益率来计

算每组的已实现波动率，以及每组间的高频协方差波动率。这样分组之后的加权计算可以有效地改进组合风险被低估的影响，起到更加精细的作用。

通过 JoinQuant 获取每只股票的市值和账面市值比从而进行分组，由于具体分组的股票数据众多不便于观察，计算已实现波动率，得到市值加权后的每组 2020 年 11 月 2 日至 2021 年 12 月 31 日共 287 个交易日的 5 分钟高频波动率如下图 3。

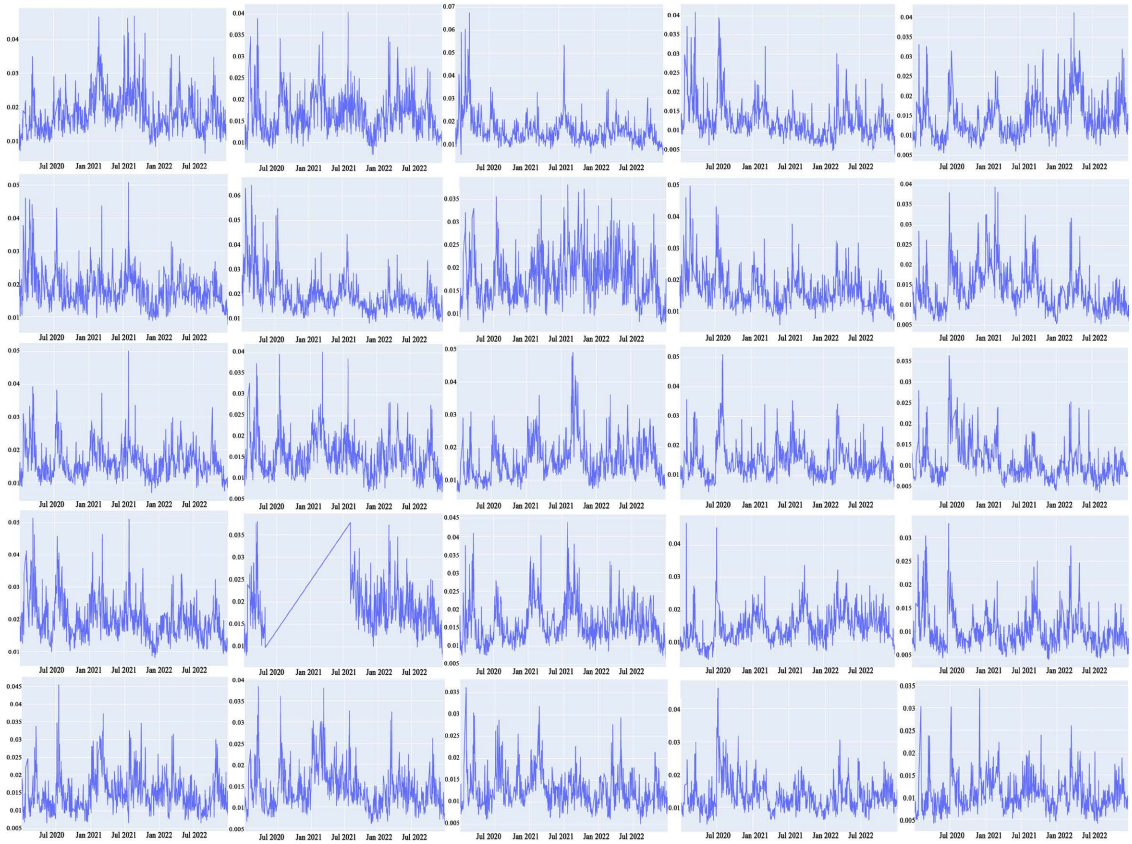

**Figure 3.** Realized volatility of each group
**图 3.** 各组的已实现波动率

其中，$\beta$ 代表的是市值，Low-$\beta$ 到 High-$\beta$ 即从左往右市值越来越高，Small-ME 到 High-ME 即从上到下账面市值比越来越高，顺序是从上往下及从左向右，即第一列为 group1 到 group5 日已实现波动率，最后一列为 group21 到 group25 日已实现波动率，通过图表可以清楚地看出通过市值因子及账面市值比因子将所有股票细分为的 25 组，如 group1 包括代号 002607、300033、300529、300595、600763、603486 共六只股票，以此类推得到所有组中的成分股。

描述性统计如下表 1：

**Table 1.** Descriptive statistics
**表 1.** 描述性统计

| 组别 | 数量(只) | 占比率 | 市值总和(亿元) | 总市值中占比(%) |
|---|---|---|---|---|
| group-1 | 6 | 2.21% | 2097.98 | 0.62% |
| group-2 | 11 | 4.04% | 3479.95 | 1.03% |
| group-3 | 14 | 5.15% | 5023.56 | 1.48% |

**Continued**

| | | | | |
|---|---|---|---|---|
| group-4 | 16 | 5.88% | 4412.04 | 1.30% |
| group-5 | 8 | 2.94% | 2989.7 | 0.88% |
| group-6 | 9 | 3.31% | 5037.86 | 1.49% |
| group-7 | 9 | 3.31% | 5809.59 | 1.71% |
| group-8 | 11 | 4.04% | 6066.3 | 1.79% |
| group-9 | 12 | 4.41% | 6669.1 | 1.97% |
| group-10 | 6 | 2.21% | 3692.61 | 1.09% |
| group-11 | 9 | 3.31% | 7457.32 | 2.20% |
| group-12 | 14 | 5.15% | 11869.68 | 3.50% |
| group-13 | 8 | 2.94% | 6950.22 | 2.05% |
| group-14 | 12 | 4.41% | 8949.49 | 2.64% |
| group-15 | 11 | 4.04% | 9670.62 | 2.85% |
| group-16 | 12 | 4.41% | 12708.59 | 3.75% |
| group-17 | 10 | 3.68% | 10549.37 | 3.11% |
| group-18 | 10 | 3.68% | 9640.94 | 2.84% |
| group-19 | 13 | 4.78% | 17248.89 | 5.09% |
| group-20 | 13 | 4.78% | 15863.39 | 4.68% |
| group-21 | 13 | 4.78% | 40161.05 | 11.84% |
| group-22 | 11 | 4.04% | 31864.08 | 9.40% |
| group-23 | 10 | 3.68% | 29647.68 | 8.74% |
| group-24 | 6 | 2.21% | 25816.6 | 7.61% |
| group-25 | 18 | 6.62% | 55473.56 | 16.36% |
| 总计 | 272 | 1 | 339150.17 | 1 |

通过上述表格，可以看出每组的股票个数差距不大，相对比较平均，多数占总数比在 4%上下，其中第 1 组和第 10 组以及第 24 组的股票数量最少为 6 只，第 25 组股票数量最多为 18 只，极差为 12；每组在总市值中的占比也呈现每五组为单位的增大趋势，符合所提到的账面市值比以及市值的分组顺序。

对分组后的股票，以每组为单位利用式(11)计算各组之间的高频协方差波动率，从而构建出高频波动率矩阵。一天得到一个 25 × 25 的高频波动率矩阵，287 天共计 287 个高频波动率矩阵。

各组间的协方差波动率公式为：

$$RV_t^{XY} = \sum_{i=1}^{n} r_{ti}^{X} r_{ti}^{Y} \tag{11}$$

其中，$X$，$Y$泛指代表的每两个组，$t$ 代表天数的时间，即 $t = 1$，2，……，287；$i$ 代表一天中高频数据的时刻，每五分钟数据的情况下，$i$ 取 1 到 48；$r_{ti}$ 是每组中所有成分股用市值加权的收益率。

第 1 天到第 287 天的高频波动率矩阵为：

$$\begin{bmatrix} 7.03\times10^{-6} & 5.03\times10^{-6} & 4.82\times10^{-6} & \cdots & \cdots & 1.46\times10^{-6} \\ 5.03\times10^{-6} & 5.46\times10^{-6} & 3.17\times10^{-6} & \cdots & \cdots & 9.87\times10^{-7} \\ 4.82\times10^{-6} & 3.17\times10^{-6} & 1.61\times10^{-5} & \cdots & \cdots & 1.86\times10^{-6} \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ 1.10\times10^{-6} & 7.89\times10^{-7} & 7.87\times10^{-7} & \cdots & \cdots & 1.24\times10^{-6} \end{bmatrix}$$

$$\begin{bmatrix} \vdots \\ \vdots \\ 2.59\times10^{-6} & 8.10\times10^{-7} & 6.71\times10^{-7} & \cdots & \cdots & 5.10\times10^{-7} \\ 8.10\times10^{-7} & 1.69\times10^{-6} & 1.16\times10^{-6} & \cdots & \cdots & 5.66\times10^{-7} \\ 6.71\times10^{-7} & 1.16\times10^{-6} & 5.44\times10^{-6} & \cdots & \cdots & 1.34\times10^{-6} \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ 5.01\times10^{-7} & 5.66\times10^{-7} & 1.34\times10^{-6} & \cdots & \cdots & 8.51\times10^{-7} \end{bmatrix}$$

将算出的 287 个高频波动率矩阵以天数为顺序“堆”在一起构成一个三阶张量 $\mathcal{X}_0 \in R^{287\times25\times25}$，如图 4。

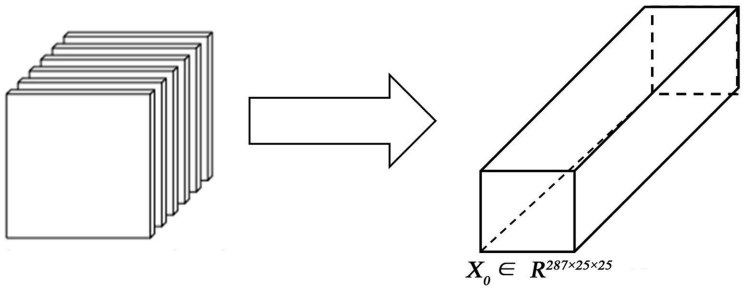


**Figure 4.** Construction of tensors
**图 4.** 张量的构建

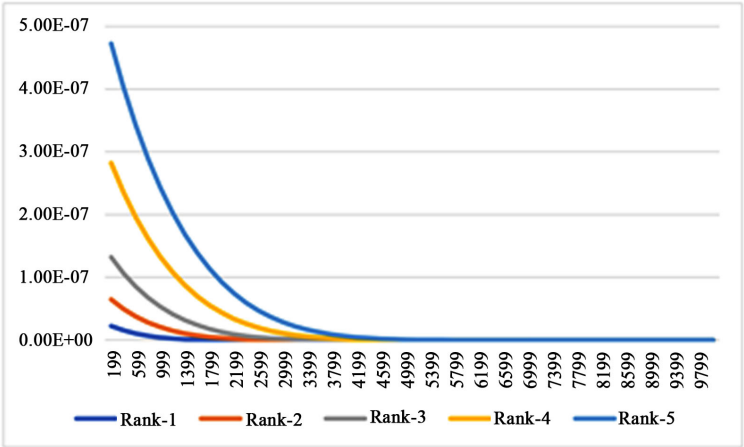


**Figure 5.** Rank-n loss function
**图 5.** Rank-n 损失函数

张量的秩 Rank-n 近似无法渐进地得到，在 $\mathcal{X}_0 \in R^{287\times25\times25}$ 代入 CP-HAR 模型求解之前，先确定其 CP 分解秩，考虑到后续计算量和难度，本文从 Rank-1 到 Rank-5 中做比较并从中寻找最合适的 Rank 值，设迭代次数 10000 次，梯度 lambda_1 = le − 4，学习率 alpha = le − 4，每 200 次迭代显示一次代价函数，观察并比较各个 Rank 值不同迭代次数的代价函数的数值及其变化趋势，见图 5。

观察上图，可以发现无论选择多少 Rank 的值，其损失函数值一开始都是随着迭代的次数而减少，减少的速度由 Rank1 到 Rank5 呈递增趋势；在迭代的前 3000 至 4000 次左右，Rank1 到 Rank5 的损失值始终从大到小，在不断减少的过程中到达某一驻点，但无法保证具体的驻点数值，只能保证在收敛到损失函数不

再下降为止。由于此图纵坐标单位的局限性，无法看到每个 Rank 值对应的驻点，只能模糊的看到损失函数在迭代大概五六千次到达驻点并稍有上升趋势，列出表格将具体数值记录下来。

表 2 统计各个 Rank 值的最小损失函数值和对应的迭代次数：

**Table 2.** Loss minimum function value
**表 2.** loss 最小函数值

| Rank 值 | Rank1 | Rank2 | Rank3 | Rank4 | Rank5 |
|---|---|---|---|---|---|
| 最小损失函数值 | 7.98E−12 | 7.46E−12 | 6.90E−12 | 6.95E−12 | 7.00E−12 |
| 对应迭代次数 | 2799 | 4399 | 5599 | 6799 | 7199 |

通过表 2，可以发现 Rank3 的最小损失函数值是最低的为 6.90E-12，其次是 Rank4 为 6.95E-12，与 Rank3 相差不大，Rank1 与 Rank2 虽计算成本不高但是精确度不够好，再考虑后续建模的计算成本，所以本文选取的秩为 3。

将构建的张量 $\mathcal{X}_0 \in R^{287\times25\times25}$ 代入 CP-HAR 模型求解，这里交替最小二乘法求解的 $R$ 为 3，设置的迭代次数为 5800 次、梯度 lambda_1 = le − 4、学习率 alpha = le − 4 得出三个因子矩阵 $\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}$ (保留五位小数)以及 65 个预测模型，即 $Q_1$，$Q_2$，$Q_3$ (表 3 展示前 15 天的 $Q_1$)等参数，具体结果如下：

$$A=\begin{bmatrix} 0.05217 & 0.02877 & 0.05226 & \cdots & 0.02728 \\ 0.03959 & 0.01942 & 0.01738 & \cdots & 0.02945 \\ -0.00124 & 0.04297 & 0.01274 & \cdots & 0.05714 \end{bmatrix}$$

$$B=\begin{bmatrix} 0.00553 & 0.00395 & 0.00369 & \cdots & 0.00195 \\ 0.00573 & 0.00415 & 0.00491 & \cdots & 0.00126 \\ 0.00591 & 0.00421 & 0.00374 & \cdots & 0.00163 \end{bmatrix}$$

$$C=\begin{bmatrix} 0.00555 & 0.00395 & 0.00369 & \cdots & 0.00195 \\ 0.00574 & 0.00415 & 0.00491 & \cdots & 0.00126 \\ 0.00592 & 0.00421 & 0.00374 & \cdots & 0.00164 \end{bmatrix}$$

**Table 3.** $Q_1$ value from the first 15 days (rounded to four decimal places in the table)
**表 3.** 前 15 天的值 $Q_1$ (表中保留四位小数)

| day | $q_{1.1}$ | $q_{1.2}$ | $q_{1.3}$ | $q_{1.4}$ | $q_{1.5}$ | $q_{1.6}$ | $q_{1.7}$ | $q_{1.8}$ | $q_{1.9}$ |
|---|---|---|---|---|---|---|---|---|---|
| 2021/09/27 | −0.0393 | 0.1996 | −0.9817 | −0.0096 | 0.1726 | −0.3925 | −0.0416 | 0.3016 | 0.0043 |
| 2021/09/28 | −0.0383 | 0.1912 | −1.0166 | −0.0066 | 0.1923 | −0.4023 | −0.0415 | 0.3120 | 0.0292 |
| 2021/09/29 | −0.0283 | 0.1889 | −1.0165 | −0.0048 | 0.1871 | −0.4397 | −0.0403 | 0.2792 | −0.0417 |
| 2021/09/30 | −0.0014 | 0.1633 | −1.0397 | −0.0235 | 0.2621 | −0.4143 | −0.0367 | 0.3133 | 0.0016 |
| 2021/10/08 | 0.0031 | 0.1827 | −1.0299 | −0.0262 | 0.2563 | −0.3976 | −0.0292 | 0.3096 | −0.0330 |
| 2021/10/11 | 0.0027 | 0.1837 | −1.0352 | −0.0283 | 0.2615 | −0.3930 | −0.0324 | 0.3154 | −0.0232 |
| 2021/10/12 | −0.0007 | 0.1982 | −1.0371 | −0.0332 | 0.2534 | −0.3741 | −0.0306 | 0.3068 | −0.0025 |
| 2021/10/13 | −0.0029 | 0.2124 | −0.9864 | −0.0359 | 0.2408 | −0.3946 | −0.0218 | 0.3034 | −0.0844 |
| 2021/10/14 | −0.0243 | 0.2364 | −0.9312 | −0.0464 | 0.2734 | −0.4991 | −0.0430 | 0.3522 | −0.2464 |
| 2021/10/15 | −0.0370 | 0.2313 | −0.8770 | −0.0478 | 0.2851 | −0.5617 | −0.0401 | 0.3811 | −0.3620 |
| 2021/10/18 | −0.0323 | 0.2485 | −0.9038 | −0.0471 | 0.2762 | −0.4984 | −0.0322 | 0.3690 | −0.3077 |
| 2021/10/19 | −0.0346 | 0.2505 | −0.8771 | −0.0473 | 0.2745 | −0.5007 | −0.0329 | 0.3662 | −0.3320 |
| 2021/10/20 | −0.0325 | 0.2260 | −0.7933 | −0.0411 | 0.2775 | −0.5702 | −0.0364 | 0.3440 | −0.4243 |
| 2021/10/21 | −0.0441 | 0.1855 | −0.6618 | −0.0615 | 0.3157 | −0.7367 | −0.0406 | 0.2953 | −0.5954 |
| 2021/10/22 | −0.0473 | 0.1725 | −0.6167 | −0.0533 | 0.3230 | −0.7946 | −0.0269 | 0.2687 | −0.6667 |

并得到通过 CP-HAR 模型预测的矩阵 $\boldsymbol{D}$ (保留五位小数)：

$$\boldsymbol{D}=\begin{bmatrix}0.03779 & 0.05390 & 0.05251 & \cdots & 0.02690\\ 0.04700 & 0.05124 & 0.05186 & \cdots & 0.02815\\ 0.03667 & 0.05219 & 0.04093 & \cdots & 0.03698\end{bmatrix}$$

用 $\boldsymbol{B}$、$\boldsymbol{C}$ 矩阵与预测得到的 $\boldsymbol{D}$ 矩阵作为三个因子矩阵通过定义一个合并函数组成一个新的三阶张量 $\mathcal{X}_0^n \in R^{65\times25\times25}$，拆分来看就是 65 个 25 × 25 的高频波动率矩阵，如图 6，即为 CP-HAR 模型预测的高频波动率矩阵。

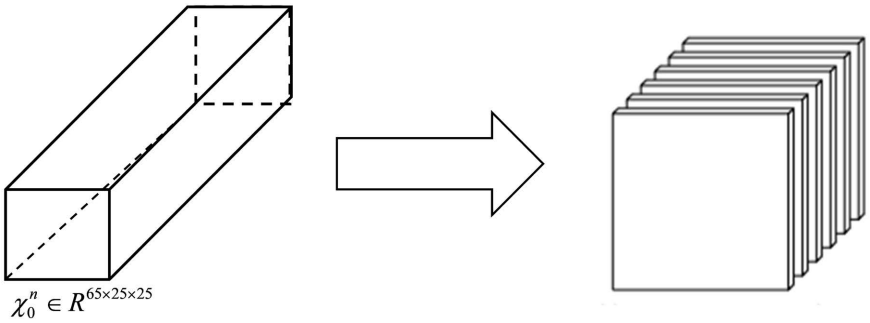


**Figure 6.** Tensor split volatility matrix
**图 6.** 张量拆分后波动率矩阵

预测后的 65 个高频波动率矩阵分别为：

$$\begin{bmatrix}3.98\times10^{-6} & 2.86\times10^{-6} & 2.91\times10^{-6} & \cdots & \cdots & 1.10\times10^{-6}\\ 2.86\times10^{-6} & 2.05\times10^{-6} & 2.09\times10^{-6} & \cdots & \cdots & 7.89\times10^{-7}\\ 2.91\times10^{-6} & 2.09\times10^{-6} & 2.16\times10^{-6} & \cdots & \cdots & 7.87\times10^{-7}\\ \vdots & \vdots & \vdots & \ddots & & \vdots\\ \vdots & \vdots & \vdots & & \ddots & \vdots\\ 1.10\times10^{-6} & 7.89\times10^{-7} & 7.87\times10^{-7} & \cdots & \cdots & 3.16\times10^{-7}\end{bmatrix}$$

$$\vdots$$

$$\begin{bmatrix}2.08\times10^{-6} & 6.86\times10^{-7} & 6.48\times10^{-7} & \cdots & \cdots & 6.67\times10^{-7}\\ 6.35\times10^{-7} & 1.36\times10^{-6} & 1.44\times10^{-6} & \cdots & \cdots & 6.42\times10^{-7}\\ 5.66\times10^{-7} & 1.53\times10^{-6} & 6.36\times10^{-6} & \cdots & \cdots & 9.69\times10^{-7}\\ \vdots & \vdots & \vdots & \ddots & & \vdots\\ \vdots & \vdots & \vdots & & \ddots & \vdots\\ 5.09\times10^{-7} & 7.87\times10^{-7} & 1.66\times10^{-6} & \cdots & \cdots & 9.73\times10^{-7}\end{bmatrix}$$

## 5. CP-HAR 模型的预测与总结

### 5.1. CP-HAR 模型的预测

本文选取常见的评价指标对 CP-HAR 模型的预测和拟合效果进行评价分析，评价的指标包括：均方根误差(Root Mean Squared Error, RMSE)，平均绝对误差(Mean Absolute Percentage Error, MAE)、平

均绝对百分比误差(Mean Absolute Percentage Error, MAPE)以及可决系数 R2，RMSE、MAE、MAPE 越小，说明该模型的预测值与真实值越接近，预测效果越好。R2 越接近 1，则说明模型与样本观测值拟合的程度就越高。

CP-HAR 模型得到的预测值与真实值都是 25 × 25 的波动率矩阵，对比 65 天的预测值与真实值误差，将这些矩阵按行拉直排列可以得到相对应共 625 行的行向量，将这些行向量的对应元素进行交叉误差验证。作为对比，通过式(4)对真实波动率矩阵的各个元素做自回归预测，同样选取 0.8 年作为滚动时间窗口做样本内预测，得到 65 个通过 HAR-RV 模型预测的高频波动率矩阵，同样地将 HAR-RV 模型预测的 65 个波动率矩阵中的对应元素与真实值进行误差验证。得到两个模型的评价指标值，具体如表 4：

**Table 4.** Various error indicators
**表 4.** 各误差指标

| 模型 | RMSE | MAE | MAPE | $R^2$ |
|---|---|---|---|---|
| CP-HAR | 3.4786E−07 | 2.3893E−07 | 21.29128 | 0.561426 |
| HAR-RV | 4.0575E−07 | 2.6275E−07 | 22.47512 | 0.507934 |

RMSE、MAE 分别表示了数据的绝对误差情况，从绝对误差角度看 CP-HAR 模型相比于 HAR-RV 模型有着更小的绝对值预测误差；从相对误差的角度看，MAPE 误差能够评估不同模型对于相同数据的拟合效果，较大的 MAPE 反映模型预测的结果相比于真实值有着更大的差距，通过上表可得出 CP-HAR 模型比 HAR-RV 模型减少了约 5.4%的 MAPE 误差。从模型的拟合程度角度看，CP-HAR 模型的可决系数 $R^2$ 明显好于 HAR-RV 模型，说明 CP-HAR 模型更好地拟合了样本值。综合比较四个评价指标，以及结合 CP-HAR 模型的可决系数，可以得出 CP-HAR 模型在预测高频波动矩阵上取得了不错的效果，相比于 HAR-RV 模型有着更精确的波动预测能力。

### 5.2. 本文总结

本文利用张量能够存储多维度信息且结构稳定等优点，并基于张量 CP 分解法在实际应用中的优势与可行性，使其与 HAR-RV 模型相结合提出 CP-HAR 模型预测高频波动率矩阵，即从一个不同的角度与方向预测高频波动率矩阵。实证研究表明：CP-HAR 模型预测得到的高频波动率矩阵在预测效果的评价指标 RMSE、MAE、MAPE 以及 $R^2$ 上表现良好，即 CP-HAR 模型能够很好地预测高频波动率矩阵，相比于 HAR-RV 模型有着更精确的波动预测能力。

## 参考文献

[1] Andersen, T.G. and Bollerslev, T. (2003) Modeling and Forecasting Realized Volatility. *Econometrica*, **71**, 579-625. https://doi.org/10.1111/1468-0262.00418

[2] Markowitz, H. (1952) Portfolio Selection. *Journal of Finance*, **7**, 77-91. https://doi.org/10.1111/j.1540-6261.1952.tb01525.x

[3] Tucker, L.R. (1963) Implications of Factor Analysis of Three-Way Matrices for Measurement of Change. In: Harris C.W., ed., *Problems in Measuring Change*, 122-137, University of Wisconsin Press, Madison.

[4] Appellof, C.J. and Davidson, E.R. (1981) Strategies for Analyzing Data from Video Fluorometric Monitoring of Liquid Chromatographic Effluents. *Analytical Chemistry*, **53**, 2053-2056. https://doi.org/10.1021/ac00236a025

[5] Knuth, D.E. (2014) Art of Computer Programming. Volume 2: Seminumerical Algorithms. Addison Wesley Professional, Boston.

[6] Sidiropoulos, N.D., Bro, R. and Giannakis, G.B. (2000) Parallel Factor Analysis in Sensor Array Processing. *IEEE Transactions on Signal Processing*, **48**, 2377-2388. https://doi.org/10.1109/78.852018

[7] Beckmann, C.F. and Smith, S.M. (2005) Tensorial Extensions of Independent Component Analysis for Multisubject FMRI Analysis. *Neuroimage*, **25**, 294-311. https://doi.org/10.1016/j.neuroimage.2004.10.043

[8] Acar, E., Çamtepe, S.A., Krishnamoorthy, M.S. and Yener, B. (2005) Modeling and Multiway Analysis of Chatroom Tensors. In: Kantor, P., Muresan, G., Roberts, F., Zeng, D.D., Wang, F.-Y., Chen, H.C. and Merkle, R.C., eds., *Intelligence and Security Informatics*, Springer, Berlin, Heidelberg.

[9] Müller, U., Dacorogna, M., Dave, R., *et al*. (1993) Fractals and Intrinsic Time—A Challenge to Econometricians. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5370

[10] Corsi, F. (2009) A Simple Long Memory Model of Realized Volatility. *Journal of Financial Econometrics*, **7**, 174-196. https://doi.org/10.1093/jjfinec/nbp001