第3卷 第3期 计 量 经 济 学 报 Vol. 3, No. 3
2023年7月 China Journal of Econometrics Jul., 2023

**doi:** 10.12012/CJoE2022-0067

# 高频数据环境下我国股票市场的波动率预测 —— 基于机器学习和 HAR 模型的融合研究

杨 科 [1,3], 张洲深 [1,4], 田凤平 [2]

(1. 华南理工大学经济与金融学院, 广州 510006; 2. 中山大学国际金融学院, 广州 510275; 3. 人工智能与数字经济广东省实验室 (广州), 广州 510330; 4. 中欧瑞博投资管理股份有限公司, 深圳 518048)

**摘 要** 股票市场波动率的准确预测对于投资者预判股市走势、优化资产配置和规避风险以及监管机构预警风险和稳定市场秩序等都具有重要的理论和现实意义. 本文在基于高频数据的 HAR 模型基础上, 融合机器学习中的 Lasso 和随机森林方法进行模型特征选择, 采用神经网络方法刻画变量间的非线性特征, 构建了几类崭新的已实现波动率模型, 并实证评价和比较各类模型对股票市场已实现波动率的预测性能. 实证结果表明: 跳跃成分的引入可以提高股票市场已实现波动率的样本外预测精度; 基于 Lasso 和随机森林进行特征选择的 HAR 扩展模型的样本外预测性能明显优于传统的 HAR 模型和 GARCH 类模型; 采用神经网络方法刻画波动率的非线性特征能进一步提高模型的样本外预测精度; 在所有考察的预测模型中, Lasso-NN-J 模型的样本内和样本外预测性能最佳, 并且在不同的预测滚动窗口宽度、不同的个股高频数据以及随机抽样模拟检验下, 该模型的样本外预测性能相当稳健.

**关键词** 特征选择; HAR 模型; 已实现波动率; 神经网络

**收稿日期:** 2022-08-02

**基金项目:** 国家自然科学基金 (72201284, 71673089); 国家社会科学基金重大项目 (21ZDA036); 教育部人文社会科学研究规划基金 (22YJA790077); 广东省自然科学基金面上项目 (2021A1515012643); 广东省哲学社会科学 “十三五” 规划 2020 年度项目一般项目 (GD20CYJ38)

**Supported by** National Natural Science Foundation of China (72201284, 71673089); Major Program of National Social Science Foundation of China (21ZDA036); Humanities and Social Sciences Foundation of Ministry of Education of China (22YJA790077); Natural Science Foundation of Guangdong Province (2021A1515012643); Philosophy and Social Sciences Foundation of Guangdong Province (GD20CYJ38)

**作者简介:** 杨科, 华南理工大学经济与金融学院教授, 博士生导师, 跨境金融创新研究中心副主任, 研究方向: 金融经济学、金融风险管理, E-mail: yangkdc@scut.edu.cn; 张洲深, 深圳市中欧瑞博投资管理股份有限公司量化研究员, 研究方向: 风险管理、金融市场, E-mail: zhangzhoushen@163.com; 通信作者: 田凤平, 中山大学国际金融学院副教授, 博士生导师, 研究方向: 金融工程与风险管理, E-mail: tfengp@mail.sysu.edu.cn.

# The Volatility Forecasting of Chinese Stock Market under High-frequency Data Environment: Fusion Research Based on Machine Learning and HAR Model

YANG Ke[1,3], ZHANG Zhoushen[1,4], TIAN Fengping[2]

(1. School of Economics and Finance, South China University of Technology, Guangzhou 510006, China;
2. International School of Business and Finance, Sun Yat-sen University, Guangzhou 510275, China;
3. Guangdong Key Laboratory of Urban Informatics, Guangzhou 510330, China; 4. China-Europe Rabbit Fund Management Co., Ltd., Shenzhen 518048, China)

**Abstract** The accurate forecasting of stock market volatility is of great theoretical and practical significance for investors to predict stock market trend, optimize asset allocation and avoid risks, and for regulators to warn risks and stabilize market order. In this paper, on the basis of HAR model based on high-frequency trading data, Lasso and random forest method in machine learning are combined to conduct model feature selection, and the nonlinear characteristics among variables are depicted by neural network method, so as to construct several new realized volatility models based on machine learning. Then, the performance of various models in forecasting the realized volatility of Shanghai stock index is evaluated and compared. The empirical results show that, the introduction of the jump component can improve the out-of-sample forecasting accuracy of realized volatility in the stock market. The HAR extended models based on Lasso and random forest for feature selection have significantly better out-of-sample prediction performance than the traditional HAR models and GARCH models. Using the neural network method to describe the nonlinear characteristics of volatility can further improve the out-of-sample prediction accuracy of the model. The Lasso-NN-J model has the best in-sample and out-sample prediction performance among all the investigated forecasting models, and the prediction performance of the model is quite robust under the simulation tests of different rolling window widths, different high-frequency data of individual stocks and random sampling.

**Keywords** feature selection; HAR model; realized volatility; neural network

## 1 引言

准确度量和预测金融波动率是资产组合管理、资产定价以及风险管理等金融实务问题的前提和关键基础, 也是金融学者关注和研究的热点. 以往文献主要采用 GARCH 族模型、SV 族模型以及多分形建模方法等对金融波动率展开研究并取得了良好效果, 如 Bollerslev (1986), Harvey and Shephard (1996), Yu (2005), Mandelbrot (1999) 等. 然而, 这些传统模型都是基于低频数据, 无法挖掘出更多有价值的日内交易信息, 其预测性能有较大的改进空间. 随着信息技术的飞速发展和金融高频数据库的不断完善, 如何充分挖掘高频数据隐含的日内交易信息并对其进行波动率建模是近年来金融计量领域广泛关注的热点问题. 由 Andersen and Bollerslev (1998) 首次提出基于高频数据的已实现波动率极大地推动了金融资产波动率的研究进展, 在高频数据框架下对金融波动率进行建模和预测得到了飞速发

展, 如 Blair, Poon and Taylor (2001), Martens and Zein (2004) 构建的 GARCH-RV 模型、Andersen, Torben and Bollerslev (2003) 构建的 VAR-RV 模型以及 Koopman, Jungbacker and Hol (2005) 构建的 SV-RV 模型和 ARFIMA-RV 模型对金融波动率的预测能力明显优于传统波动率模型. 虽然这些研究都取得了较好的预测效果, 但普遍缺乏明确的经济含义, 并且在建模过程中仍然易损失市场交易信息.

Corsi (2009) 在异质市场假说和 HARCH 模型基础上, 构建了异质性自回归已实现波动率模型 (HAR-RV 模型). 该模型通过自回归结构加总日、周以及月波动率成分, 仅采用 OLS 就能估计其参数, 能成功地捕获已实现波动率的长记忆性, 并且对金融波动率的预测性能明显优于 GARCH 和 ARFIMA-RV 等模型. 在此之后, 许多学者在 HAR-RV 模型的基础上, 构建了各类新的波动率预测模型以进一步提高金融波动率的预测精度, 比较有代表性有: Andersen, Bollerslev and Diebold (2007) 将连续方差和离散跳跃方差引入 HAR-RV 模型, 构建的 HAR-RV-J 和 HAR-RV-CJ 模型; Chen and Diebold (2011) 通过将已实现波动率分解为好消息驱动的波动率成分和坏消息驱动的波动率成分并引入跳跃波动率成分, 构建的 HAR-S-RV-J 模型; Patton and Sheppard (2015) 通过将已实现波动率分解为上下行已实现半方差并引入符号跳跃方差, 构建的 HAR-RSV、HAR-RV-SJV 和 HAR-RV-SJVD 模型. 此外, 国内学者也构建了一些新的基于高频数据的波动率预测模型, 并取得了比较好的预测效果. 例如, 田凤平和杨科 (2016) 构建的同时考虑 HAR 模型参数时变性和预测因子时变性的具有时变稀疏度的 HAR 模型 (TVS-HAR), 龚旭, 文凤华和黄创霞等 (2017) 结合 EMD 方法和 HAR-RV-J 模型构建的 HAR-RV-EMD-J 模型, 罗嘉雯和陈浪南 (2018) 基于 Kalli and Griffin (2014) 的时变稀疏模型和多元 HAR 模型构建的具有时变稀疏性的多元 HAR 模型 (TVS-MHAR), 陈声利, 李一军和关涛 (2018) 基于跳跃、好坏波动率与符号跳跃构建了单级纠偏 HARQ 类模型和多级纠偏 HARQF 类模型, 龚旭, 曹杰和文凤华等 (2020) 构建的带杠杆效应和结构突变的 HAR 族模型, 瞿慧和沈微 (2020) 将百度指数作为逻辑平滑转移结构的转移变量引入到 HAR 模型, 构建的 LSTHAR 模型等.

虽然上述 HAR 类模型对金融波动率的预测性能较好, 但仍存在较大的改进空间. 大量研究表明, 金融波动率的影响因素数目非常多, 如已实现波动率的滞后值、跳跃成分等, 而任何潜在的影响因素对已实现波动率的预测能力总是随市场和预测期的不同而改变. 因此, 运用 HAR 类模型预测已实现波动率时将面临模型选择方面的问题. 为了避免模型选择风险, 通常的有效方法是将所有潜在的影响因素全部纳入 HAR 模型, 但当潜在影响因素数目很大时, 又会面临过度拟合问题, 反而降低预测模型的样本外预测精度. 另外, 经典的 HAR 模型建模时只使用了日、周以及月波动率成分, 但并不代表其他波动率成分的预测效果不显著. 因此, 如何对数目巨大的潜在影响因素进行特征选择以同时避免模型选择风险和过度拟合问题是进一步改进和扩展 HAR 类模型的重要研究方向. 此外, 金融时间序列大多的表现出非线性特征, 而 HAR 模型只能解释已实现波动率序列的线性部分, 因此, 需要在此基础上构建一些新的模型来捕获这些非线性特征. 有鉴于此, 本文通过融合新兴的机器学习方法和经典的 HAR 模型, 降低了模型的预测误差, 以期有效提高股市波动率的预测精度, 其创新和对文献的贡献主要体现为如下几个方面: 1) 将机器学习中的 Lasso 方法和随机森林方法融入传统 HAR 模型, 实现对潜在影响因素进行特征选择, 以克服模型选择风险和过度拟合

问题; 2) 将神经网络模型引入到预测模型中, 有效刻画已实现波动率序列的非线性特征; 3) 构建了八类崭新的已实现波动率预测模型, 并采用 MCS 检验和随机抽样模拟检验等方法实证评价和比较了新构建的预测模型与传统的 HAR 和 HAR-J 模型、基于低频数据的经典 GARCH 族模型以及最新发展的单级纠偏 HARQ-RV-SJ 模型、多级纠偏 HARQF-RV-CJ 模型和 TVS-HAR 模型对股票市场已实现波动率的预测性能. 本研究为股市波动率的预测提供了新的思路和方法, 拓展了机器学习在金融时间序列上的运用, 有助于投资者准确预判股市走势、实时优化资产配置、及时规避市场风险以及增强经济决策的前瞻性和审慎性, 有助于监管层更高效的监控市场动态、更前瞻和更灵活地引导股市的风险预期并强化市场监管绩效, 对于保持金融市场的长期稳定运行等具有比较重要的现实意义.

后文的结构安排如下: 第二节阐述本文构建的已实现波动率模型和相关的研究方法; 第三节为各类已实现波动率模型预测性能的评价和比较; 第四节给出本文的主要研究结论.

## 2 模型与方法

### 2.1 已实现波动率及其跳跃成分的估计

Andersen and Bollerslev (1998) 首次提出了基于高频数据的已实现波动率度量方法, 其表述如下:

$$\mathrm{RV}_t = \sum_{j=1}^{n} \left[p_{j,t} - p_{j-1,t}\right]^2 = \sum_{j=1}^{n} r_{j,t}^2, \tag{1}$$

其中, $p_{j,t}$ 为 $t$ 日的第 $j$ 个日内对数价格, $r_{t,j} = p_{j,t} - p_{j-1,t}$ 为第 $j$ 个日内对数收益率, $n$ 为每个交易日内对数收益率的数量. 从理论上讲, 如果价格没有噪声, RV 是累积波动率的一致地估计量. Bandi and Russell (2006), Hansen and Lunde (2006) 的研究表明, 由于市场微观结构噪声引起的自相关性, 式 (1) 估计的 RV 可能不是一致估计量, 目前绝大多数文献采用 5 分钟频率的高频数据来减弱市场微观结构噪声的影响.

为了考察跳跃对已实现波动率预测的影响, Barndorff-Nielsen (2004) 基于二次幂变差理论将已实现波动率分解为连续成分和离散跳跃成分, 并将离散跳跃成分定义为跳跃波动率. 根据 Andersen, Bollerslev and Huang (2011) 的研究, 可以通过式 (2) 估计跳跃波动率:

$$J_t = \max(\mathrm{RV}_t - \mathrm{BPV}_t, 0), \tag{2}$$

其中, BPV 为已实现二次幂变差, 该估计量是积分波动率的一致估计量, 其可以表述为:

$$\mathrm{BPV}_t = \frac{\pi}{2} \sum_{j=2}^{n} \left|r_{t,j-1}\right| \left|r_{t,j}\right|. \tag{3}$$

### 2.2 HAR 模型和 HAR-J 模型

Corsi (2009) 在异质市场假说的基础上构建的 HAR-RV 模型不仅能刻画不同期限的投资者对金融波动率的贡献, 还能近似捕获波动率的长记忆性. Vortelinos (2017) 的研究也指出该模型能很好地探究波动率的运行规律. 标准的 HAR-RV 模型可表述为:

$$\mathrm{RV}_t = \beta_0 + \beta_D \mathrm{RV}_{t-1} + \beta_W \mathrm{RV}_{t-1}^W + \beta_M \mathrm{RV}_{t-1}^M + \varepsilon_t, \tag{4}$$

其中, $\mathrm{RV}_t^W = \frac{1}{5}\sum_{i=1}^{5}\mathrm{RV}_{t-i+1}$ 和 $\mathrm{RV}_t^M = \frac{1}{22}\sum_{i=1}^{22}\mathrm{RV}_{t-i+1}$ 分别表示周度和月度 RV. HAR-RV 模型相当于有约束条件的 AR(22) 模型, 其系数 $\beta_0$, $\beta_D$, $\beta_W$ 和 $\beta_M$ 可通过简单的 OLS 估计得到. 在 HAR-RV 模型中加入式 (2) 估计的跳跃波动率成分, 可以得到如下的 HAR-J 模型:

$$\mathrm{RV}_t = \beta_0 + \beta_D\mathrm{RV}_{t-1} + \beta_W\mathrm{RV}_{t-1}^W + \beta_M\mathrm{RV}_{t-1}^M + \beta_D^J J_{t-1} + \beta_W^J J_{t-1}^W + \beta_M^J J_{t-1}^M + \varepsilon_t. \quad (5)$$

### 2.3 基于特征选择的 HAR 类模型

为了同时避免模型选择风险和过度拟合问题, 本文根据 HAR 模型和 HAR-J 模型的设定形式, 运用 Lasso (全称 least absolute shrinkage and selection operator) 和随机森林 (random forest, RF) 对模型的解释变量进行特征选择, 构建了四类新的基于特征选的 HAR 扩展模型: Lasso-HAR 模型、Lasso-HAR-J 模型、RF-HAR 模型和 RF-HAR-J 模型, 以期通过特征选择来进一步提高已实现波动率模型的预测精度.

#### 2.3.1 基于 Lasso 的 HAR 模型

Lasso 是由 Tibshirani (1996) 提出的一种降维变量选择方法, 该方法通过将惩罚函数引入目标函数来压缩最优变量系数解中所含的变量个数. 自 Kock (2012) 的研究以来, 越来越多学者在计量经济学领域开始使用 Lasso 作为模型特征选择工具. 假设 $x_t = (x_{t1}, x_{t2}, \cdots, x_{tp})'$ 为预测变量, 在本文中为已实现波动率的影响因素, 包括 $\mathrm{RV}_{t-i}$ $(i = 1, 2, \cdots, 22)$ 和 $J_{t-i}$ $(i = 1, 2, \cdots, 22)$, 考虑如下回归方程:

$$\mathrm{RV}_t = \alpha + \phi' \times x_t + \varepsilon_t. \quad (6)$$

该回归模型的 Lasso 模型估计如下:

$$\hat{\phi}^{\mathrm{lasso}} = \arg\min_{\phi}\left\{\sum_{t=1}^{n}\left(\mathrm{RV}_i - \sum_{j=1}^{p}\phi_j x_{tj}\right)^2 + \lambda\sum_{j=1}^{p}|\phi_j|\right\}. \quad (7)$$

在式 (7) 中, 由 $L^1$-范数作为正则项, 所以 Lasso 的解是稀疏的, 即部分 $\phi_j$ 将会等于 0, 这就达到了特征选择的结果. 由于 $\phi_j$ 对调整参数或惩罚系数 $\lambda$ 比较敏感, 调整参数 $\lambda$ 对于特征选择和预测结果影响较大, 因此 Lasso 模型的关键在于调整参数 $\lambda$ 的选取. 文献中经常采用 Bootstrap 交叉验证、K 折交叉验证和蒙特卡罗交叉验证等方法来选择调整参数 $\lambda$. 基于数据驱动的准则, 本文采用 K 折交叉验证和均方误差 (MSE) 结合的方法来选择最优的调整参数 $\lambda$, 然后在此基础上分析预测模型的样本内和样本外预测结果.

本文首先确定调整参数 $\lambda$ 的最优取值, 然后通过使用 Lasso 模型对解释变量 $\mathrm{RV}_{t-i}$ $(i = 1, 2, \cdots, 22)$ 进行特征选择, 并将选择得到的变量运用 OLS 回归得到 Lasso-HAR 模型:

$$\mathrm{RV}_t = \beta_0 + \beta_{\mathrm{lasso}}\mathrm{RV}_{t-1}^{\mathrm{lasso}} + \varepsilon_t, \quad (8)$$

其中, $\mathrm{RV}_{t-1}^{\mathrm{lasso}}$ 是由最优惩罚系数 $\lambda$ 选择出来的特征, $\beta_{\mathrm{lasso}}$ 为相应的回归系数. 同样, 通过使用 Lasso 模型对解释变量 $\mathrm{RV}_{t-i}$ $(i = 1, 2, \cdots, 22)$ 和 $J_{t-i}$ $(i = 1, 2, \cdots, 22)$ 进行选择, 并将

选择得到的变量运用 OLS 回归得到 Lasso-HAR-J 模型:

$$\mathrm{RV}_t = \beta_0 + \beta_{\mathrm{lasso}}\mathrm{RV}_{t-1}^{\mathrm{lasso}} + \beta_{\mathrm{lasso}}^{J} J_{t-1}^{\mathrm{lasso}} + \varepsilon_t, \tag{9}$$

其中, $\mathrm{RV}_{t-1}^{\mathrm{lasso}}$ 和 $J_{t-1}^{\mathrm{lasso}}$ 为在给定惩罚系数 $\lambda$ 选择出来的特征, $\beta_{\mathrm{lasso}}$ 和 $\beta_{\mathrm{lasso}}^{J}$ 为相应的回归系数.

### 2.3.2 基于 RF 的 HAR 模型

Breiman (2001) 提出的 RF 是一种集成算法, 其基分类器是决策树, 在进行决策时, 根据某一规则对决策树的叶子结点进行划分, 直到满足终止条件为止. 常用的规则包括信息增益和基尼系数. 本文采用基尼系数作为评价标准, 通过控制 RF 选择变量的个数和 Lasso 选择变量的个数保持一致, 以便更好地研究模型的特征选择能力. 具体而言, 通过使用 RF 对解释变量 $\mathrm{RV}_{t-i}$ $(i = 1, 2, \cdots, 22)$ 进行选择, 选择与式 (8) $\mathrm{RV}_{t-1}^{\mathrm{lasso}}$ 相同的维度作为选择特征个数, 得到解释变量 $\mathrm{RV}_{t-1}^{rf}$, 并运用 OLS 回归得到 RF-HAR 模型:

$$\mathrm{RV}_t = \beta_0 + \beta_{rf}\mathrm{RV}_{t-1}^{rf} + \varepsilon_t. \tag{10}$$

同样, 通过使用 RF 对解释变量 $\mathrm{RV}_{t-i}$ $(i = 1, 2, \cdots, 22)$ 和 $J_{t-i}$ $(i = 1, 2, \cdots, 22)$ 进行选择, 并将选择得到的变量运用 OLS 回归得到 RF-HAR-J 模型:

$$\mathrm{RV}_t = \beta_0 + \beta_{rf}\mathrm{RV}_{t-1}^{rf} + \beta_{rf}^{J} J_{t-1}^{rf} + \varepsilon_t, \tag{11}$$

其中, $\mathrm{RV}_{t-1}^{rf}$ 和 $J_{t-1}^{rf}$ 是由最优惩罚系数 $\lambda$ 选择出来的特征, $\beta_{rf}$ 和 $\beta_{rf}^{J}$ 为相应的回归系数.

## 2.4 基于 NN 的 HAR 类模型

金融时间序列往往表现出非线性的特征, 而传统的 HAR 模型等价于带约束的 AR (22) 模型, 属于线性模型范畴, 无法捕获金融已实现波动率的非线性特征, 因此本文将神经网络 (neural network, 后简称 NN) 引入到 HAR 模型, 构建了基于 NN 的 HAR 模型. 相比如其他机器学习模型, NN 模型的参数较多, 容易出现过拟合问题. 因此, 本文对 NN 的深度和每一层的节点数进行了严格的限制, 通过构建相对简单的 NN 来防止过拟合风险. 本文将 $\mathrm{RV}_{t-1}$, $\mathrm{RV}_{t-5}$ 和 $\mathrm{RV}_{t-22}$ 作为输入, $\mathrm{RV}_t$ 作为输出, 构建的 HAR-NN 模型为:

$$\mathrm{RV}_t = f\left(\phi_{co} + \sum_{h=1}^{2} \phi_{ho} g\left(\phi_{ch} + \phi_{1h}\mathrm{RV}_{t-1} + \phi_{2h}\mathrm{RV}_{t-5} + \phi_{3h}\mathrm{RV}_{t-22}\right)\right) + \varepsilon_t, \tag{12}$$

其中, $f(\cdot)$ 和 $g(\cdot)$ 分别是隐含层和输出层的激活函数. 由于本文侧重于探讨采用 NN 模型刻画已实现波动率之间非线性的关系, 对于不同激活函数的选取以及隐含层数量的确定均沿用文献中普遍采用的方法. 在本文中, $g$ 采用 tanh 激活函数 $\left(y = \frac{\mathrm{e}^{x} - \mathrm{e}^{-x}}{\mathrm{e}^{x} + \mathrm{e}^{-x}}\right)$, $f$ 采用线性激活函数 $(y = x)$, 输入层的维数为 3, 隐藏层的维度为 2.

此外, 本文还将 $\mathrm{RV}_{t-1}$, $\mathrm{RV}_{t-5}$, $\mathrm{RV}_{t-22}$, $J_{t-1}$, $J_{t-5}$ 和 $J_{t-22}$ 作为输入, 将 $\mathrm{RV}_t$ 作为输

出, 构建的 HAR-NN-J 模型表述如下:

$$\begin{aligned}\mathrm{RV}_t =& f\left(\phi_{co}+\sum_{k=1}^{3}\phi_{ko}g^2\left(\phi_{ck}+\sum_{h=1}^{6}\phi_{ho}g^1(\phi_{ch}+\phi_{1h}\mathrm{RV}_{t-1}+\phi_{2h}\mathrm{RV}_{t-5}+\phi_{3h}\mathrm{RV}_{t-22}+\right.\right.\\ &\left.\left.\phi_{4h}J_{t-1}+\phi_{5h}J_{t-5}+\phi_{6h}J_{t-22}\right)+\varepsilon_t,\end{aligned} \tag{13}$$

其中, 输出层维度为 6, 隐含层选择 2 层, 各隐含层神经元的个数分别为 6 和 3. $g^1$ 为第一个隐含层的激活函数, 采用 sigmoid 激活函数 ($y=\frac{1}{1+\mathrm{e}^{-x}}$), $g^2$ 为第一个隐含层的激活函数, 采用 tanh 激活函数, $f$ 为输出层的 linear 激活函数.

### 2.5 基于神经网络和特征选择的 HAR 模型

结合 NN 类模型以及 Lasso 和 RF, 本文进一步构建了另外四类新的已实现波动率模型: Lasso-NN 模型、Lasso-NN-J、RF-NN 模型和 RF-NN-J, 以期同时考虑特征选择和非线性特征来进一步提高已实现波动率模型的预测精度. 具体而言, 通过使用 Lasso 模型对解释变量 $\mathrm{RV}_{t-i}$ ($i=1,2,\cdots,22$) 进行选择, 并将选择得到变量构建神经网络模型得到 Lasso-NN 模型:

$$\mathrm{RV}_t=\begin{cases}f\left(\phi_{co}+\sum_{h=1}^{2}\phi_{ho}g\left(\phi_{ch}+\sum_{i=1}^{p}\phi_{ih}\mathrm{RV}_{t-1,i}^{\mathrm{lasso}}\right)\right)+\varepsilon_t, & p<5,\\ f\left(\phi_{co}+\sum_{k=1}^{3}\phi_{ko}g^2\left(\phi_{ck}+\sum_{h=1}^{6}\phi_{ho}g\left(\phi_{ch}+\sum_{i=1}^{p}\phi_{ih}\mathrm{RV}_{t-1,i}^{\mathrm{lasso}}\right)\right)\right)+\varepsilon_t, & p\geq 5,\end{cases} \tag{14}$$

其中, $\mathrm{RV}_{t-1,i}^{\mathrm{lasso}}$ 为在给定惩罚系数 $\lambda$ 选择出来的特征, 当使用 Lasso 选择出来的特征维度小于 5 时, 使用 1 层隐含层, 采用 tanh 激活函数; 当使用 tanh 激活函数. 同样, 使用 Lasso 模型对解释变量 $\mathrm{RV}_{t-i}$ ($i=1,2,\cdots,22$) 和 $J_{t-i}$ ($i=1,2,\cdots,22$) 进行选择, 并将选择得到的变量使用 NN 模型训练得到 Lasso-NN-J 模型:

$$\mathrm{RV}_t=\begin{cases}f\left(\phi_{co}+\sum_{h=1}^{2}\phi_{ho}g\left(\phi_{ch}+\sum_{i=1}^{p}\phi_{ih}\mathrm{RV}_{t-1,i}^{\mathrm{lasso}}+\sum_{i=1}^{q}\phi_{ih}J_{t-1,i}^{\mathrm{lasso}}\right)\right)+\varepsilon_t, & p+q<5,\\ f\left(\phi_{co}+\sum_{k=1}^{3}\phi_{ko}g^2\left(\phi_{ck}+\sum_{h=1}^{6}\phi_{ho}g\left(\phi_{ch}+\sum_{i=1}^{p}\phi_{ih}\mathrm{RV}_{t-1,i}^{\mathrm{lasso}}+\sum_{i=1}^{q}\phi_{ih}J_{t-1,i}^{\mathrm{lasso}}\right)\right)\right)+\varepsilon_t, & p+q\geq 5,\end{cases} \tag{15}$$

其中, $\mathrm{RV}_{t-1}^{\mathrm{lasso}}$ 和 $J_{t-1}^{\mathrm{lasso}}$ 为在给定惩罚系数 $\lambda$ 选择出来的特征, 当输入特征维度小于 5 时, 使用 1 层隐含层, 隐含层的神经元数量为 2 个, 采用 tanh 激活函数; 当输入特征维度大于等于 5 时, 使用 2 层隐含层, 分别采用 sigmoid 激活函数和 tanh 激活函数, 隐含层的神经元数量分别为 6 个和 3 个.

通过使用 RF 对解释变量 $\mathrm{RV}_{t-i}$ $(i=1,2,\cdots,22)$ 进行选择, 并将选择得到的变量构建 NN 模型, 得到如下的 RF-NN 模型:

$$\mathrm{RV}_t=\begin{cases} f\left(\phi_{co}+\sum\limits_{h=1}^{2}\phi_{ho}g\left(\phi_{ch}+\sum\limits_{i=1}^{p}\phi_{ih}\mathrm{RV}_{t-1,i}^{\mathrm{rf}}\right)\right)+\varepsilon_t, & p<5,\\ f\left(\phi_{co}+\sum\limits_{k=1}^{3}\phi_{ko}g^2\left(\phi_{ck}+\sum\limits_{h=1}^{6}\phi_{ho}g\left(\phi_{ch}+\sum\limits_{i=1}^{p}\phi_{ih}\mathrm{RV}_{t-1,i}^{\mathrm{rf}}\right)\right)\right)+\varepsilon_t, & p\geq 5, \end{cases} \tag{16}$$

其中, $\mathrm{RV}_{t-1,i}^{rf}$ 为在给定惩罚系数 $\lambda$ 选择出来的特征, 当使用 RF 选择出来的特征维度小于 5 时, 使用 1 层隐含层, 采用 tanh 激活函数; 当使用随机森林选择出来的特征维度大于等于 5 时, 使用 2 层隐含层, 分别采用 sigmoid 激活函数和 tanh 激活函数. 同样, 使用随机森林模型对解释变量 $\mathrm{RV}_{t-i}$ $(i=1,2,\cdots,22)$ 和 $J_{t-i}$ $(i=1,2,\cdots,22)$ 进行选择, 并将选择得到的变量使用 NN 模型训练得到 RF-NN-J 模型:

$$\mathrm{RV}_t=\begin{cases} f\left(\phi_{co}+\sum\limits_{h=1}^{2}\phi_{ho}g\left(\phi_{ch}+\sum\limits_{i=1}^{p}\phi_{ih}\mathrm{RV}_{t-1,i}^{\mathrm{rf}}+\sum\limits_{i=1}^{q}\phi_{ih}J_{t-1,i}^{\mathrm{rf}}\right)\right)+\varepsilon_t, & p+q<5,\\ f\left(\phi_{co}+\sum\limits_{k=1}^{3}\phi_{ko}g^2\left(\phi_{ck}+\sum\limits_{h=1}^{6}\phi_{ho}g\left(\phi_{ch}+\sum\limits_{i=1}^{p}\phi_{ih}\mathrm{RV}_{t-1,i}^{\mathrm{rf}}+\sum\limits_{i=1}^{q}\phi_{ih}J_{t-1,i}^{\mathrm{rf}}\right)\right)\right)+\varepsilon_t, & p+q\geq 5, \end{cases} \tag{17}$$

其中, $\mathrm{RV}_{t-1}^{rf}$ 和 $J_{t-1}^{rf}$ 为在给定惩罚系数 $\lambda$ 选择出来的特征, 当输入特征维度小于 5 时, 使用 1 层隐含层, 采用 tanh 激活函数, 隐含层的神经元数量为 2 个; 当输入特征维度大于等于 5 时, 使用 2 隐含层, 分别采用 sigmoid 激活函数和 tanh 激活函数, 隐含层的神经元数量分别为 6 个和 3 个.

## 3　预测性能评价和比较

### 3.1　研究数据

鉴于文献中大多采用 5 分钟频率的高频数据来减弱市场微观结构噪声的影响, 本文采用上证指数的 5 分钟高频数据作为研究样本来计算已实现波动率及其跳跃波动成分. 数据来源于 CSMAR 数据库, 其时间跨度为 2006 年 1 月 1 日至 2018 年 3 月 31 日. 表 1 给出了上证指数已实现波动率及其跳跃成分的描述性统计量. 从表 1 可以看出, 跳跃波动成分的均值相对于已实现波动率的均值而言比较小, 在总共 2990 个交易日中发生了 768 次跳跃, 跳跃占比 25.69%.

### 3.2　预测结果与分析

本文通过比较样本内和样本外预测结果来评估上述模型的预测性能. 样本内预测方面, 选用全样本对上述 HAR 模型进行估计, 并在此基础上得到样本内预测值; 样本外预测方面, 采用滚动的固定时间窗口方法得到样本外的向前一步预测值, 具体步骤参见魏宇 (2010), 龚旭, 曹杰和文凤华等 (2020), 其中, 滚动的固定窗口长度为全样本长度的 1/3, 这里取 990 天, 预测区间覆盖了异常波动的牛熊市和窄幅波动的低迷市等.

**表 1 上证指数 RV 及其 Jump 成分的描述性统计量**

| | 均值 | 标准差 | 最小值 | 最大值 | 偏度 | 峰度 |
|---|---|---|---|---|---|---|
| RV | 0.8325 | 0.4574 | 0.2317 | 4.7522 | 2.4918 | 10.9068 |
| Jump | 0.0063 | 0.0246 | 0.0000 | 0.7225 | 16.3752 | 396.9947 |

注: RV 表示当期已实现波动率; Jump 表示当期跳跃, 0 表示不存在跳跃.

### 3.2.1 样本内预测结果评价和比较

首先使用 Lasso 进行特征选择, 对于调整参数 $\lambda$ 选择的特征个数, 在运用随机森林进行特征选择时, 控制特征的维数选择与 Lasso 相同的维度. 本文基于前 990 天的样本数据采用 $K$ 折交叉验证法选择最优的调整参数 $\lambda$, 并同时将其作为样本内和样本外预测的调整参数[1]. 表 2 给出了基于最优调整参数的样本内预测结果, 评价指标采用 MSE, MAE, $R^2$ 和可解释方差 EV (Explained Variance). 从表 2 的样本内预测评价结果可知: 1) 调整参数 $\lambda$ 取最优值时, 不带跳跃和带跳跃特征的模型选择中都选择了 6 个特征, 根据 Lasso 模型选择的结果, 选择的特征均是 $\mathrm{RV}_{t-1}$, $\mathrm{RV}_{t-2}$, $\mathrm{RV}_{t-3}$, $\mathrm{RV}_{t-4}$, $\mathrm{RV}_{t-17}$, $\mathrm{RV}_{t-20}$, 都没有选择跳跃特征, 这说明跳跃成分在固定的样本期内对上证指数 RV 的解释性不强[2]; 2) 基于 Lasso 特征选择的 HAR 族模型在所有评价标准下对上证指数 RV 的样本内预测效果都明显优于经典的 HAR 模型和 HAR-J 模型, 而基于 RF 特征选择的 HAR 族模型仅在 $R^2$ 和 EV 评价标准下对上证指数 RV 的样本内预测效果优于经典的 HAR 模型和 HAR-J 模型, 但在 MSE 和 MAE 评价标准下其样本内预测效果表现不佳; 3) 除了 EV 评价标准, 在其他所有评价标准下, 基于 Lasso 特征选择的 Lasso-NN-J 模型对上证指数 RV 的样本内预测效果最优. 因此, 总体而言, 基于 Lasso 特征选择的 HAR 族模型对上证指数 RV 的样本内预测效果都明显优于 HAR 模型和 HAR-J 模型, 而基于 RF 特征选择的 HAR 族模型的样本内预测效果相对较差; 在大多数的情况下, 基于 Lasso 特征选择的 Lasso-NN-J 模型对上证指数 RV 的样本内预测效果最优.

### 3.2.2 样本外预测结果评价和比较

本节采用 HAR、HAR-J、Lasso-HAR、Lasso-HAR-J、Lasso-NN、Lasso-NN-J、RF-HAR、RF-HAR-J、RF-NN 和 RF-NN-J 模型, 运用基于滚动固定时间窗的样本外预测方法对上证指数 RV 进行预测. 为了更好地比较上述各模型的样本外预测性能, 本文还同时考虑了陈声利, 李一军和关涛 (2018) 构建的单级纠偏 HARQ 类模型和多级纠偏 HARQF 类模型以及田凤平和杨科 (2016) 构建的同时考虑 HAR 模型参数时变性和预测因子时变性的 TVS-

[1]由于篇幅限制, 正文中没有报告 Lasso 模型进行交叉验证的结果, 有需要的读者可以向作者索取调整参数 $\lambda$ 与预测误差的关系图.

[2]大量研究表明, 跳跃成分包含了股票市场或经济市场的重要信息, 其对已实现波动率预测具有显著的重要影响, 而本文的样本内拟合结果显示, Lasso 模型选择都没有选择跳跃特征, 其原因是在固定的样本内期间, 跳跃成分对上证指数 RV 的影响程度不如 RV 滞后值强, 当我们进一步考虑不同的样本内区间时, 我们发现在大部分所考虑样本内区间, Lasso 模型选择会选择跳跃特征. 因此, 就平均而言, 跳跃成分对我国股市已实现波动率的样本内预测确实具有显著的重要影响, 这与现有研究结论是一致的.

**表 2　样本内预测精度评价结果**

| 模型 | 特征数 | MSE | MAE | $R^2$ | EV |
|---|---|---|---|---|---|
| HAR | 3 | 0.0634 | 0.1503 | 0.5651 | 0.5651 |
| HAR-J | 6 | 0.0620 | 0.1498 | 0.5786 | 0.5786 |
| Lasso-HAR | 6 | 0.0619 | 0.1498 | 0.5601 | 0.5601 |
| Lasso-HAR-J | 6 | 0.0619 | 0.1498 | 0.5601 | 0.5601 |
| Lasso-NN | 6 | 0.0583 | 0.1467 | 0.4057 | 0.4062 |
| Lasso-NN-J | 6 | **0.0582** | **0.1460** | **0.3918** | 0.3941 |
| RF-HAR | 6 | 0.0638 | 0.1516 | 0.5609 | 0.5609 |
| RF-HAR-J | 6 | 0.0639 | 0.1518 | 0.5603 | 0.5603 |
| RF-NN | 6 | 0.0683 | 0.1561 | 0.4095 | 0.4099 |
| RF-NN-J | 6 | 0.0684 | 0.1523 | 0.3923 | **0.3939** |

注: 加粗数据表示所有预测模型中评价指标值最小者.

HAR 模型作为对照模型. 陈声利, 李一军和关涛 (2018) 的研究表明, 单级纠偏 HARQ 类模型的样本外预测性能普遍优于多级纠偏 HARQF 类模型, 并且 HARQ-RV-SJ 模型表现最佳, 为了避免实证中比较模型过多, 造成预测评价比较结果混乱, 本文选取了 HARQ-RV-SJ 模型以及多级纠偏 HARQF 类模型中样本外向前一步预测性能表现最优 (MSE 评价标准下) 的 HARQF-RV-CJ 模型作为对照模型, 模型的设定形式参见陈声利, 李一军和关涛 (2018). TVS-HAR 模型的样本外预测值基于所有参数的后验均值计算得到, 具体模型设定形式和预测步骤可参见田凤平和杨科 (2016) 以及罗嘉雯和陈浪南 (2018). 此外, 本文还考虑了基于低频数据的经典 GARCH 族模型: GARCH, TARCH, GJR 和 EGARCH 作为对照模型, 借此考察运用高频数据是否具备优越性.

表 3 给出了在调整参数取最优值情况下, 评价指标 MSE, MAE, $R^2$ 和 EV 对各个预测模型的样本外预测精度评价结果. 尽管这些评价指标给出了各类 HAR 模型的预测精度排名, 但无法提供模型预测性能在统计学意义上的显著差异, 只能判断在一个特定的数据样本和某一特定的评价指标下, 某些模型比其他模型的预测精度高, 而这一判断无法推广到其他数据样本或其他评价标准. 为了增强结论的稳健性, 本文还进一步采用 Hansen, Lunde and Nason (2011) 的 “模型置信集” (model confidence set, MCS) 检验来评价和比较上述各种预测模型的预测性能. MCS 检验是在一组候选预测模型集合 $M_0$ 中进行持续的显著性检验, 不断剔除集合 $M_0$ 中预测能力较差的候选预测模型, 直到没有模型被剔除为止, 剩余幸存模型即为模型置信集. 该检验每次检验的零假设都是候选预测模型集合 $M_0$ 中某两个模型具有相同的预测能力, 文献中大多采用范围统计量和二次方统计量作为实证检验准则来检验这一零假设. 由于范围统计量和二次方统计量的真实分布非常复杂, 实证研究中这两个统计量及相应的 $p$ 值可通过 “自助法” (bootstrap) 模拟获得. 若某些预测模型的 $p$ 值大于临界值, 则这些模型通过 MCS 检验, 为模型置信集中的幸存模型, 具有比 MCS 检验剔除的模型更好的预测性能. MCS 检验 $p$ 值越大, 表明对应预测模型的预测性能越优. 该检验的具体流程和原理可参考 Hansen, Lunde and Nason (2011).

表 4 报告了, 在显著水平为 0.1 和 0.25 时, 所有预测模型在 MSE 和 MAE 评价标准和

**表 3 样本外预测精度评价结果**

| 模型 | MSE | MAE | $R^2$ | EV |
|---|---|---|---|---|
| HAR | 0.1031 | 0.1781 | 0.5928 | 0.5929 |
| HAR-J | 0.1032 | 0.1802 | 0.5988 | 0.5993 |
| Lasso-HAR | 0.1012 | 0.1766 | 0.6018 | 0.6019 |
| Lasso-HAR-J | 0.1012 | 0.1766 | 0.6018 | 0.6019 |
| Lasso-NN | 0.0950 | 0.1698 | 0.4444 | 0.4448 |
| Lasso-NN-J | **0.0935** | 0.1616 | **0.4398** | **0.4403** |
| RF-HAR | 0.1016 | 0.1774 | 0.5927 | 0.5923 |
| RF-HAR-J | 0.1022 | 0.1777 | 0.5905 | 0.5907 |
| RF-NN | 0.0934 | 0.1650 | 0.4575 | 0.4575 |
| RF-NN-J | 0.0936 | **0.1609** | 0.4404 | 0.4420 |
| HARQ-RV-SJ | 0.0956 | 0.1724 | 0.4572 | 0.4896 |
| HARQF-RV-CJ | 0.0960 | 0.1736 | 0.4605 | 0.4899 |
| TVS-HAR | 0.0938 | 0.1619 | 0.4401 | 0.4439 |
| GARCH | 0.1048 | 0.1885 | 0.6132 | 0.6005 |
| TARCH | 0.1033 | 0.1814 | 0.6105 | 0.6004 |
| GJR | 0.1032 | 0.1811 | 0.6106 | 0.6000 |
| EGARCH | 0.1033 | 0.1817 | 0.6108 | 0.6007 |

注: 加粗数据表示所有预测模型中评价指标值最小者.

**表 4 MCS 检验结果**

| 模型 | MSE | MAE |
|---|---|---|
| HAR | 0.003 | 0.001 |
| HAR-J | 0.002 | 0.000 |
| Lasso-HAR | 0.003 | 0.001 |
| Lasso-HAR-J | 0.000 | 0.000 |
| Lasso-NN | 0.002 | 0.001 |
| Lasso-NN-J | **1.000**** | **1.000**** |
| RF-HAR | 0.002 | 0.000 |
| RF-HAR-J | 0.002 | 0.000 |
| RF-NN | **0.265**** | **0.240*** |
| RF-NN-J | **0.860**** | **0.851**** |
| HARQ-RV-SJ | **0.228*** | **0.219*** |
| HARQF-RV-CJ | **0.228*** | **0.211*** |
| TVS-HAR | **0.711**** | **0.718**** |
| GARCH | 0.001 | 0.000 |
| TARCH | 0.002 | 0.000 |
| GJR | 0.000 | 0.000 |
| EGARCH | 0.000 | 0.000 |

注: 表中的数值表示 MCS 检验的 $p$ 值, $p$ 值越大, 表明该模型的预测精度越高; ** 表示模型属于置信水平为 25% 的模型置信集 $M^*_{0.75}$, * 表示模型属于置信水平为 10% 的模型置信集 $M^*_{0.90}$, 其中 $M^*_{0.75} \subset M^*_{0.90}$.

范围统计量下, 自助法模拟 1000 次的 MCS 检验结果, $p$ 值小于 0.1 (或 0.25) 的预测模型为样本外预测性能较差的模型, 将在 MCS 检验中被剔除, 而 $p$ 值大于 0.1 (或 0.25) 的预测模型则是样本外预测能力较好的模型, 将在 MCS 检验中保留下来, 即为显著水平为 10% 或 (25%) 的模型置信集 $M^*_{0.90}$ (或 $M^*_{0.75}$) 的元素. 若 $p$ 值等于 1, 则说明该预测模型是所有候选模型中最优的预测模型. 结合表 3 的样本外预测精度评价结果以及表 4 的 MCS 检验结果可知:

1) 基于低频数据的经典 GARCH 族模型: GARCH, TARCH, GJR 和 EGARCH 模型的损失函数值都明显高于基于高频数据的 HAR 模型及其扩展模型, 并且这几类经典 GARCH 模型的 MCS 检验 $p$ 值都小于 0.1, 均被排除在模型置信集之外, 说明对基于高频数据的已实现波动率进行预测建模能显著提高波动率的样本外预测精度, 这一结论与经典文献 Andersen, Torben and Bollerslev (2003), Corsi (2009) 的研究结论一致.

2) 基于特征选择的 Lasso-HAR、Lasso-HAR-J、RF-HAR 和 RF-HAR-J 模型在所有评价标准下对上证指数 RV 的样本外预测精度都明显优于经典的 HAR 模型和 HAR-J 模型, 并且在 MCS 检验中得以幸存的模型均是带有特征选择的 HAR 类模型, 而 HAR 模型和 HAR-J 模型的 MCS 检验 $p$ 值都小于 0.1, 说明特征选择对于已实现波动的样本外预测

极为重要. 其原因可能有两个方面: 一方面, 由于 HAR 模型和 HAR-J 模型中只包含了日线、周线和月线的波动率和跳跃, 这些因子可能并不是最好的预测因子; 另一方面, 近些年来国际金融市场的冲击、投资者情绪的变化以及一些政策性的影响, 上证指数的已实现波动率可能存在着结构性的突变, 导致不同时间段预测因子的预测性能可能不同, 经典 HAR 模型和 HAR-J 模型无法反映这些信息, 而通过特征选择可以根据不同的市场环境适时地选择出不同的最优预测因子, 类似于考虑了预测因子预测性能的时变性, 能较大程度地缓解结构突变对样本外波动率预测造成的影响. 因此, 投资者应结合市场行情和需要, 利用 Lasso 和 RF 等特征选择方法相机挑选出预测性能较好的预测因子, 进而构建融合特征选择方法的波动率模型来提高波动率的预测精度, 增加获利机会.

3) 相比于线性的基于特征选择的 HAR 类模型 (Lasso-HAR、Lasso-HAR-J、RF-HAR 和 RF-HAR-J 模型), 其对应形式的非线性 NN 模型 (Lasso-NN、Lasso-NN-J、RF-NN 和 RF-NN-J 模型) 对上证指数 RV 的样本外预测的评价指标值普遍更低, 并且在 MCS 检验中模型置信集所包含的基于特征选择的 HAR 类模型均是非线性的 NN 类模型 (Lasso-NN-J、RF-NN 和 RF-NN-J 模型的 MCS 检验 $p$ 值均大于 0.25, 都通过了 MCS 检验), 说明采用 NN 刻画波动率序列的非线性特征可以进一步提高预测模型的样本外预测精度, 这一结论与 Vortelinos (2017), Yang, Chen and Tian (2015) 的研究结论一致.

4) 通过比较基于特征选择的 HAR 类模型中不带跳跃成分模型和对应带跳跃成分模型, 我们发现: Lasso-HAR 模型与 Lasso-HAR-J 模型的所有损失函数值都相等, 在 MCS 检验中的 $p$ 值也几乎无差别, Lasso-NN-J 模型的所有损失值均小于 Lasso-NN 模型, 且 Lasso-NN-J 模型的 MCS 检验 $p$ 值等于 1, 而 Lasso-NN 模型的 MCS 检验 $p$ 值小于 0.1, 未能通过 MCS 检验; RF-HAR-J 模型的 $R^2$ 和 EV 值比 RF-HAR 模型小, 但两者的 MCS 检验 $p$ 值基本无差别, RF-NN-J 模型的 MAE、$R^2$ 和 EV 值均比 RF-NN 模型小, 且 RF-NN-J 模型的 MCS 检验 $p$ 值大于 0.1, 通过了 MCS 检验, 而 RF-NN 模型的 MCS 检验 $p$ 值小于 0.1, 未能通过 MCS 检验. 此外, MCS 检验的模型置信集中总共包含了 5 个模型, 其中 4 个为带跳跃成分的模型. 因此, 带跳跃成分的模型对上证指数 RV 的样本外预测精度在大多数情况下要高于不带跳跃成分的模型, 说明预测模型中引入跳跃成分可以进一步提高模型的样本外预测性能, 这一发现与 Andersen, Bollerslev and Diebold (2007), Patton and Sheppard (2015), 龚旭, 文凤华和黄创霞等 (2017) 以及马锋, 魏宇和黄登仕 (2017) 的研究结论一致. 因此, 投资者和相关的监管部门应密切关注宏观信息尤其是隔夜信息等引发的跳跃风险, 应将这些市场极端风险来源纳入金融风险预警机制, 并为系统性风险的监控和政策调控的针对性提供参考.

5) HARQ-RV-SJ 和 HARQF-RV-CJ 模型对上证指数 RV 的样本外预测精度明显高于 HAR 和 HAR-J 模型, 并且两者的 MCS 检验 $p$ 值均大于 0.1, 说明通过单级纠偏和多级纠偏可以进一步提高预测模型的样本外预测性能, 这一结论与陈声利, 李一军和关涛 (2018) 的研究结论一致; TVS-HAR 模型在所有情况下对上证指数 RV 的样本外预测表现很稳健, 其预测精度排名都在第 2 或者第 3 位, 并且在所有损失函数下, 该模型的 MCS 检验 $p$ 值均大于 0.25, 这与田凤平和杨科 (2016), 罗嘉雯和陈浪南 (2018) 的研究结论一致, 其原因可能是 TVS-HAR 模型采用时变稀疏度的方法同时考虑了 HAR 模型参数的时变性和预测因子的时变性, 一定程度上类似于特征选择; 在所有考察的预测模型中, Lasso-NN-J 模型的损失函数

值最低, 且 MCS 检验 $p$ 值等于 1, 说明该模型是对上证指数 RV 的样本外预测性能最优的模型. 因此, 投资者在基于高频数据的已实现波动率建模分析中, 应同时注重预测因子预测性能的时变特征以及波动率的非线性特征, 应结合市场行情和需要, 利用 Lasso 和 RF 特等特征选择方法在不同的市场环境下相机挑选出预测性能较好的预测因子, 并利用神经网络等方法捕获波动率序列的非线性特征, 进而构建融合特征选择方法和神经网络的波动率模型实现对股票市场波动率的准确预测, 增强经济决策的前瞻性和审慎性.

### 3.3 稳健性检验

为更充分地验证上述实证结果的准确性, 本文采用三种方式对模型预测性能进行稳健性检验: 1) 调整滚动窗口宽度重新进行预测, 并检验模型在不同滚动窗口下的预测性能; 2) 采用个股高频数据重新检验模型的预测性能; 3) 采用随机抽样模拟检验模型的预测性能.

#### 3.3.1 基于不同滚动窗的预测性能比较

本文进一步将固定滚动窗调整为 500 和 1500 重新进行样本外预测, 并继续采用基于 MSE 和 MAE 的 MCS 检验对预测结果进行稳健性检验, 表 5 报告了所有预测模型在滚动窗口分别为 500 和 1500 时 MCS 检验 $p$ 值.

从表 5 可以看出, 预测检验结果与前文基本一致, 在不同的滚动窗口下, 从 MCS 检验幸

**表 5 MCS 检验结果 (预测滚动窗口为 500 和 1500)**

| | 预测滚动窗口为 500 | | 预测滚动窗口为 1500 | |
|---|---|---|---|---|
| 模型 | MSE | MAE | MSE | MAE |
| HAR | 0.000 | 0.001 | 0.000 | 0.006 |
| HAR-J | 0.001 | 0.000 | 0.004 | 0.000 |
| Lasso-HAR | 0.005 | 0.001 | 0.000 | 0.000 |
| Lasso-HAR-J | 0.000 | 0.000 | 0.000 | 0.000 |
| Lasso-NN | 0.000 | 0.002 | 0.002 | 0.002 |
| Lasso-NN-J | **0.897**** | **0.901**** | **1.000**** | **1.000**** |
| RF-HAR | 0.000 | 0.000 | 0.004 | 0.000 |
| RF-HAR-J | 0.000 | 0.000 | 0.003 | 0.001 |
| RF-NN | **0.314**** | **0.318**** | **0.289**** | **0.280**** |
| RF-NN-J | **0.749**** | **0.718**** | **0.904**** | **0.896**** |
| HARQ-RV-SJ | **0.305**** | **0.302**** | **0.253**** | **0.220*** |
| HARQF-RV-CJ | **0.305**** | **0.301**** | **0.252**** | **0.221*** |
| TVS-HAR | **0.647**** | **0.652**** | **0.711**** | **0.718**** |
| GARCH | 0.000 | 0.000 | 0.003 | 0.000 |
| TARCH | 0.001 | 0.000 | 0.000 | 0.001 |
| GJR | 0.000 | 0.000 | 0.000 | 0.000 |
| EGARCH | 0.001 | 0.000 | 0.000 | 0.001 |

注: 表中的数值表示 MCS 检验的 $p$ 值, $p$ 值越大, 表明该模型的预测精度越高; ** 表示模型属于置信水平为 25% 的模型置信集 $M^*_{0.75}$, * 表示模型属于置信水平为 10% 的模型置信集 $M^*_{0.90}$, 其中 $M^*_{0.75} \subset M^*_{0.90}$.

存下来的模型置信集中包含了 Lasso-NN-J 模型、RF-NN 模型和 RF-NN-J 模型, 其 MCS 检验 $p$ 值分别为 0.897、0.314 和 0.749, 而经典的 HAR 模型和 HAR-J 模型均未通过 MCS 检验, 说明在 HAR 模型基础上同时融入特征选择方法、波动率的非线性特征以及跳跃成分可以提高已实现波动率的样本外预测精度; 在所有考察模型中, 新提出的 Lasso-NN-J 模型在所有损失函数下的 MCS 检验 $p$ 值均为最大值 (其 $p$ 值等于 1 或者接近 1), 说明该模型是所有考察模型中预测性能表现最好的模型.

### 3.3.2 基于个股高频数据的预测性能比较

本文选择了我国股市最具代表性的 10 只规模最大、流动性较好的个股, 包括浦发银行、上海机场、民生银行等, 具体的个股名称缩写参见表 6 第 1 行, 时间跨度为 2007 年 12 月 12 日至 2019 年 12 月 31 日, 共 2934 个交易日, 仍采用固定滚动窗为 990 对各个预测模型进行滚动样本外预测, 表 6 报告了各个预测模型滚动样本外预测的基于 MSE 的 MCS 检验 $p$ 值及其预测精度排名均值. 从表 6 可以看出, 对于所有个股而言, 从 MCS 检验幸存下来的模型置信集中均包含了 Lasso-NN-J 模型、RF-NN 模型和 RF-NN-J 模型, 且这些模型的排名均值都优于经典的 HAR 模型和 HAR-J 模型, 而经典的 HAR 模型和 HAR-J 模型均未通过 MCS 检验, 再次验证了在 HAR 模型基础上同时融入特征选择方法、波动率的非线性特征以及跳跃成分可以提高已实现波动率的样本外预测精度; Lasso-NN-J 模型的 MCS 检验 $p$ 值在所有考察模型中再次显示为最大值 (等于 1 或者接近于 1), 进一步说明 Lasso-NN-J 模型是所有考察模型中预测性能最优的模型.

### 3.3.3 随机抽样模拟检验

第三种稳健性检验方式是根据 Beyaztas, Firuzan and Beyaztas (2017) 的研究, 采用区组长度为 480 的 SONBB 随机抽样方法 (具体抽样步骤参见 Beyaztas, Firuzan and Beyaztas (2017), 于孝建和王秀花 (2018)) 对实证研究中采用的上证指数 5 分钟高频数据进行重复抽样 5000 次, 利用抽样得到的高频数据计算已实现波动率、跳跃及其所需的滞后值, 然后对各预测模型的样本外波动率预测损失函数进行 MCS 检验, 表 7 报告了在损失函数 MSE、MAE 和 EV 下的 MCS 检验 $p$ 值大于 0.9 的次数 (表中未带括号的数值) 及其中位数 (表中带括号的数值). 因为在每个损失函数评价标准下得到的实证结果基本一致, 本文以 MSE 为例进行具体分析. 在损失函数 MSE 评价标准下, Lasso-HAR、Lasso-HAR-J、Lasso-NN、Lasso-NN-J、RF-HAR、RF-HAR-J、RF-NN 和 RF-NN-J 模型 MCS 检验 $p$ 值大于 0.9 的次数分别为 796、1359、1024、4592、788、957、2411 和 3281, 其中 Lasso-NN-J、RF-NN 和 RF-NN-J 模型 MCS 检验 $p$ 值的中位数大于 0.25, 分别为 0.941、0.485 和 0.794, 而经典的 HAR 模型和 HAR-J 模型 MCS 检验 $p$ 值大于 0.9 的次数仅为 103 和 335, 对应的中位数都接近 0, 说明在 HAR 模型基础上融入特征选择方法、波动率的非线性特征以及跳跃成分确实可以提高已实现波动率的样本外预测精度. 此外, 无论在哪一个损失函数下, 新构建的 Lasso-NN-J 模型的 MCS 检验 $p$ 值大于 0.9 的次数及其中位数 (分别为 4592 和 0.941) 在所有考察模型中均为最大值, 表明 Lasso-NN-J 模型在所有考察模型中是预测性能最佳的模型.

综合表 5、表 6 和表 7 给出的预测评价和比较结果可知, 在不同的预测滚动窗口宽度、不同的个股高频数据以及随机抽样模拟检验下, 检验结果与前文得到的实证结果一致: 采用

**表 6　MCS 检验结果 (个股样本)**

| | PFYH | SHJC | MSYH | CPCC | ZXZQ | ZSYH | BLDC | ZGLT | SQJT | FXYY | 排名均值 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HAR | 0.000 | 0.000 | 0.000 | 0.006 | 0.000 | 0.001 | 0.001 | 0.001 | 0.000 | 0.006 | 12.6 |
| HAR-J | 0.002 | 0.000 | 0.002 | 0.000 | 0.001 | 0.000 | 0.007 | 0.000 | 0.002 | 0.000 | 12.3 |
| Lasso-HAR | 0.008 | 0.003 | 0.000 | 0.000 | 0.005 | 0.001 | 0.003 | 0.001 | 0.000 | 0.000 | 10.4 |
| Lasso-HAR-J | 0.001 | 0.004 | 0.002 | 0.000 | 0.002 | 0.010 | 0.008 | 0.000 | 0.004 | 0.000 | 7.8 |
| Lasso-NN | 0.000 | 0.002 | 0.002 | 0.002 | 0.000 | 0.004 | 0.002 | 0.001 | 0.002 | 0.001 | 8.5 |
| Lasso-NN-J | **0.882**** | **0.912**** | **1.000**** | **0.940**** | **0.992**** | **0.835**** | **1.000**** | **0.794**** | **1.000**** | **0.922**** | 1.1 |
| RF-HAR | 0.001 | 0.002 | 0.004 | 0.002 | 0.000 | 0.000 | 0.002 | 0.000 | 0.009 | 0.004 | 10.6 |
| RF-HAR-J | 0.002 | 0.001 | 0.003 | 0.008 | 0.003 | 0.002 | 0.008 | 0.002 | 0.001 | 0.000 | 9.4 |
| RF-NN | **0.414**** | **0.354**** | **0.217*** | **0.272**** | **0.300**** | **0.353**** | **0.286**** | **0.251**** | **0.301**** | **0.296**** | 4.6 |
| RF-NN-J | **0.701**** | **0.676**** | **0.822**** | **0.830**** | **0.765**** | **0.704**** | **0.765**** | **0.702**** | **0.883**** | **0.853**** | 2.6 |
| HARQ-RV-SJ | **0.312**** | **0.384**** | **0.238*** | **0.231*** | **0.274**** | **0.314**** | **0.324**** | **0.316**** | **0.230**** | **0.249*** | 5.2 |
| HARQF-RV-CJ | **0.321**** | **0.275**** | **0.260**** | **0.272**** | **0.354**** | **0.311**** | **0.276**** | **0.240*** | **0.241**** | **0.206*** | 5.2 |
| TVS-HAR | **0.710**** | **0.677**** | **0.711**** | **0.725**** | **0.848**** | **0.791**** | **0.812**** | **0.824**** | **0.760**** | **0.698**** | 2.4 |
| GARCH | 0.000 | 0.001 | 0.003 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.001 | 0.000 | 16.8 |
| TARCH | 0.001 | 0.002 | 0.000 | 0.001 | 0.002 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 16.2 |
| GJR | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 14.7 |
| EGARCH | 0.002 | 0.001 | 0.001 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.003 | 14.3 |

注: 表中的数值表示 MCS 检验的 $p$ 值, $p$ 值越大, 表明该模型的预测精度越高; ** 表示模型属于置信水平为 25% 的模型置信集 $M^*_{0.75}$, * 表示模型属于置信水平为 10% 的模型置信集 $M^*_{0.90}$, 其中 $M^*_{0.75} \subset M^*_{0.90}$.

**表 7　随机抽样模拟检验结果**

| | MSE | MAE | $R^2$ | EV |
|---|---|---|---|---|
| HAR | 103 | 154 | 208 | 246 |
| | (0.000) | (0.000) | (0.001) | (0.000) |
| HAR-J | 335 | 392 | 371 | 287 |
| | (0.001) | (0.001) | (0.001) | (0.001) |
| Lasso-HAR | 796 | 874 | 857 | 904 |
| | (0.001) | (0.001) | (0.001) | (0.002) |
| Lasso-HAR-J | 1359 | 1281 | 1302 | 1115 |
| | (0.003) | (0.008) | (0.005) | (0.005) |
| Lasso-NN | 1024 | 1105 | 986 | 929 |
| | (0.001) | (0.002) | (0.002) | (0.002) |
| Lasso-NN-J | 4592 | 4124 | 4035 | 4033 |
| | (0.941) | (0.934) | (0.901) | (0.900) |
| RF-HAR | 788 | 892 | 905 | 893 |
| | (0.001) | (0.001) | (0.001) | (0.002) |
| RF-HAR-J | 957 | 1032 | 1237 | 1189 |
| | (0.003) | (0.002) | (0.003) | (0.001) |
| RF-NN | 2411 | 2856 | 2903 | 2844 |
| | (0.485) | (0.405) | (0.358) | (0.372) |
| RF-NN-J | 3281 | 3397 | 2679 | 3048 |
| | (0.794) | (0.816) | (0.727) | (0.715) |
| HARQ-RV-SJ | 2015 | 2189 | 2614 | 2149 |
| | (0.385) | (0.326) | (0.318) | (0.283) |
| HARQF-RV-CJ | 2024 | 2207 | 2960 | 2367 |
| | (0.396) | (0.310) | (0.289) | (0.307) |
| TVS-HAR | 3495 | 3102 | 3369 | 3206 |
| | (0.873) | (0.815) | (0.796) | (0.810) |
| GARCH | 0 | 0 | 0 | 0 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| TARCH | 0 | 0 | 0 | 0 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| GJR | 10 | 0 | 0 | 21 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| EGARCH | 0 | 0 | 0 | 0 |
| | (0.000) | (0.000) | (0.000) | (0.000) |

基于高频数据的已实现波动率预测建模能显著提高金融资产未来波动率的样本外预测精度, 基于特征选择的 Lasso-HAR、Lasso-HAR-J、RF-HAR 和 RF-HAR-J 模型的样本外预测精度明显优于经典的 HAR 模型和 HAR-J 模型, 具有非线性特征的 Lasso-NN、Lasso-NN-J、RF-NN 和 RF-NN-J 模型的样本外预测性能在大多数情况下都优于线性的 HAR 类模型, 带跳跃成分的模型的样本外预测精度在大多数情况下要高于不带跳跃成分的模型, 并且在所有考察的预测模型中, Lasso-NN-J 模型的样本外预测性能最强.

## 4 主要结论及启示

本文将机器学习中的 Lasso 和 RF 融入 HAR 模型中进行模型特征选择, 采用 NN 刻画波动率序列的非线性特征, 构建了几类崭新的已实现波动率模型, 并采用 MCS 检验实证评价和比较了各类已实现波动率预测模型的预测性能. 此外, 本文还通过调整滚动窗口宽度、采用不同的研究样本 (个股高频数据) 和随机抽样模拟检验方法对模型样本外预测性能进行了稳健性检验. 研究结论表明, 采用基于高频数据的已实现波动率预测建模能显著提高波动率的样本外预测精度, 模型中引入跳跃成分可以进一步提高模型的样本外预测性能, 分别融入特征选择方法和考虑波动率的非线性特征的 HAR 类模型对已实现波动率的样本外预测性能明显优于经典的 HAR 模型和 HAR-J 模型, 同时融入 Lasso 特征选择、神经网络和跳跃成分的 Lasso-NN-J 模型是所有考察的预测模型中样本内和样本外预测性能最佳的模型. 这些研究结论在不同的预测滚动窗口宽度、不同的个股高频数据以及随机抽样模拟检验下都是稳健的. 基于以上研究结论, 本文得到如下启示: 首先, 相关监管部门在建立金融风险预警机制时应充分结合隔夜信息等引发的跳跃风险, 及时预判未来市场风险的走势, 提高政策调控的针对性和时效性, 避免市场的过度波动, 实现国内金融市场的稳定发展; 其次, 投资者在优化资产配置和规避市场风险时, 应结合市场行情和需要, 可利用 Lasso 和 RF 等特征选择方法在不同市场环境下相机挑选出预测性能较好的预测因子, 并利用 NN 等方法捕获波动率序列的非线性特征, 进而实现对股票市场波动率的准确预测, 增强经济决策前瞻性和审慎性. 此外, 还需综合运用不同预测性能检验方法, 避免方法上的误差, 提高风险预判和风险管理效 果.

本文在高频数据下初步探讨了通过融合机器学习和 HAR 模型来提高金融资产波动率的预测精度, 以此为基础可以进一步拓展的研究方向包括: 1) 使用更加全面的机器学习模型与 HAR 模型融合, 以进一步提高金融波动率的预测精度, 尤其是最近金融机器学习中的有监督数据降维方法, 如 Huang, Jiang and Tu et al. (2015) 的 PLS 模型, Huang, Jiang and Li et al. (2022) 的 Scaled PCA 模型等; 2) 参照 Huang, Jiang and Tu et al. (2015), Jiang, Lee and Martin et al. (2019), 姜富伟, 孟令超和唐国豪 (2021), 姜富伟, 胡逸驰和黄楠 (2021) 以及闵峰, 文凤华和吴楠 (2021) 的研究, 从流动性冲击、信息冲击和情绪冲击等角度进一步解释金融资产价格波动的经济学来源.

## 参 考 文 献

陈声利, 李一军, 关涛, (2018). 基于四次幂差修正 HAR 模型的股指期货波动率预测 [J]. 中国管理科学, 26(1): 57–71.

Chen S L, Li Y J, Guan T, (2018). Forecasting Realized Volatility of Chinese Stock Index Futures Based on Approved HAR Models with Median Realized Quarticity[J]. Chinese Journal of Management Science, 26(1): 57–71.

龚旭, 曹杰, 文凤华, 杨晓光, (2020). 基于杠杆效应和结构突变的 HAR 族模型及其对股市波动率的预测研究 [J]. 系统工程理论与实践, 40(5): 1113–1133.

Gong X, Cao J, Wen F H, Yang X G, (2020). The HAR-TYPE Models with Leverage and Structural Breaks and Their Applications to the Volatility Forecasting of Stock Market[J]. Systems Engineering — Theory & Practice, 40(5): 1113–1133.

龚旭, 文凤华, 黄创霞, 杨晓光, (2017). HAR-RV-EMD-J 模型及其对金融资产波动率的预测研究 [J]. 管理

评论, 29(1): 19–32.

Gong X, Wen F H, Huang C X, Yang X G, (2017). The HAR-RV-EMD-J Model and Its Application to Forecasting the Volatility of Financial Assets[J]. Management Review, 29(1): 19–32.

罗嘉雯, 陈浪南, (2018). 基于 TVS-MHAR 模型金融市场高频多元波动率的预测 [J]. 系统工程理论与实践, 38(7): 1677–1689.

Luo J W, Chen L N, (2018). Multivariate Realized Volatility Forecasts of Financial Markets Based on TVS-MHAR Model[J]. Systems Engineering — Theory & Practice, 38(7): 1677–1689.

姜富伟, 胡逸驰, 黄楠, (2021). 央行货币政策报告文本信息、宏观经济与股票市场 [J]. 金融研究, (6): 95–115.

Jiang F W, Hu Y C, Huang N, (2021). Textual Information of Central Bank Monetary Policy Report, Macroeconomy and Stock Market Performance[J]. Journal of Financial Research, (6): 95–115.

姜富伟, 孟令超, 唐国豪, (2021). 媒体文本情绪与股票回报预测 [J]. 经济学季刊, 21(4): 1323–1344.

Jiang F W, Meng L C, Tang G H, (2021). Media Textual Sentiment and Chinese Stock Return Predictability[J]. China Economic Quarterly, 21(4): 1323–1344.

马锋, 魏宇, 黄登仕, (2017). 基于符号收益和跳跃变差的高频波动率模型 [J]. 管理科学学报, 20(10): 31–43.

Ma F, Wei Y, Huang D S, (2017). Forecasting the Realized Volatility Based on the Signed Return and Signed Jump Variation[J]. Journal of Management Sciences in China, 20(10): 31–43.

闵峰, 文凤华, 吴楠, (2021). 货币政策和财政政策对中国消费和投资的有效性评估 [J]. 计量经济学报, 1(1): 94–113.

Min F, Wen F H, Wu N, (2021). Assessing the Effectiveness of Monetary and Fiscal Policies on Chinese Investment and Consumption[J]. China Journal of Econometrics, 1(1): 94–113.

瞿慧, 沈微, (2020). 基于 LSTHAR 模型的投资者关注对股市波动影响研究 [J]. 中国管理科学, 28(7): 23–34.

Qu H, Shen W, (2020). The Impact of Investor Attention on Market Volatility Based on the LSTHAR Model[J]. Chinese Journal of Management Science, 28(7): 23–34.

田凤平, 杨科, (2016). 基于 TVS-HAR 模型的农产品期货市场已实现波动率的预测研究 [J]. 系统工程理论与实践, 36(12): 3003–3016.

Tian F P, Yang K, (2016). Forecasting Realized Volatility of Agricultural Commodity Futures Using TVS-HAR Model[J]. Systems Engineering — Theory & Practice, 36(12): 3003–3016.

魏宇, (2010). 沪深 300 股指期货的波动率预测模型研究 [J]. 管理科学学报, 13(2): 66–76.

Wei Y, (2010). Volatility Forecasting Models for CS1300 Index Futures[J]. Journal of Management Sciences in China, 13(2): 66–76.

于孝建, 王秀花, (2018). 基于混频已实现 GARCH 模型的波动预测与 VaR 度量 [J]. 统计研究, 35(1): 104–116.

Yu X J, Wang X H, (2018). Mix Frequency Realized GARCH Models: The Forecast of Volatility and Measure of VaR[J]. Statistical Research, 35(1): 104–116.

Andersen T G, Bollerslev T, (1998). Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts[J]. International Economic Review, 39(4): 885–905.

Andersen T G, Bollerslev T, Diebold F X, (2007). Roughing It Up: Including Jum Components in the Measurement, Modeling and Forecasting of Return Volatility[J]. The Review of Economics and Statistics, 89(4): 701–720.

Andersen T G, Bollerslev T, Huang X, (2011). A Reduced form Framework for Modeling Volatility of Speculative Prices Based Realized Variation Measures[J]. Journal of Econometrics, 160(1): 176–189.

Andersen T G, Torben T, Bollerslev T, (2003). Modeling and Forecasting Realized Volatility[J]. Econometrica, 71(2): 579–625.

Bandi F M, Russell J R, (2006). Separating Microstructure Noise from Volatility[J]. Journal of Financial Economics, 79(3): 655–692.

Barndorff-Nielsen O E, (2004). Econometrics of Testing for Jumps in Financial Economics Using Bipower Variation[J]. Journal of Financial Econometrics, 4(1): 1–30.

Beyaztas B H, Firuzan E, Beyaztas U, (2017). New Block Bootstrap Methods: Sufficient and/or Ordered[J]. Communications in Statistics: Simulation and Computation, 46(5): 3942–3951.

Blair B J, Poon S H, Taylor S J, (2001). Forecasting S&P100 Volatility: Incremental Information Content of Implied Volatility and High Frequency Index Returns[J]. Journal of Econometrics, 45(2): 195–213.

Bollerslev T, (1986). Generalized Autoregressive Conditional Heteroskedasticity[J]. Journal of Econometrics, 31(3): 307–327.

Breiman L, (2001). Random Forests[J]. Machine Learning, 45: 5–32.

Chen X, Diebold F X, (2011). News-good or Bad and Its Impact on Volatility Predictions Over Multiple Horizons[J]. Review of Financial Studies, 24(1): 1–37.

Corsi F, (2009). A Simple Approximate Long-Memory Model of Realized Volatility[J]. Journal of Financial Econometrics, 7(2): 174–196.

Hansen P R, Lunde A, (2006). Realized Variance and Market Microstructure Noise[J]. Journal of Business & Economic Statistics, 24(2): 127–161.

Hansen P R, Lunde A, Nason J M, (2011). The Model Confidence Set[J]. Econometrica, 79(2): 453–497.

Harvey A C, Shephard N, (1996). Estimation of an Asymmetric Stochastic Volatility Model for Asset Returns[J]. Journal of Business & Economic Statistics, 14(4): 429–434.

Huang D S, Jiang F W, Li K P, Tong G S, Zhou G F, (2022). Scaled PCA: A New Approach to Dimension Reduction[J]. Management Science, 68(3): 1678–1695.

Huang D S, Jiang F W, Tu J, Zhou G F, (2015). Investor Sentiment Aligned: A Powerful Predictor of Stock Returns[J]. The Review of Financial Studies, 28(3): 791–837.

Jiang F W, Lee J S, Martin X, Zhou G F, (2019). Manager Sentiment and Stock Returns[J]. Journal of Financial Economics, 132(1): 126–149.

Kalli M, Griffin J E, (2014). Time-varying Sparsity in Dynamic Regression Models[J]. Journal of Econometrics, 178(2): 779–793.

Kock A B, (2012). On the Oracle Property of the Adaptive Lasso in Stationary and Nonstationary Autoregressions[R]. CREATES Research Papers 5, Aarhus University.

Koopman S J, Jungbacker B, Hol E, (2005). Forecasting Daily Variability of the S&P100 Stock Index Using Historical, Realized and Implied Volatility Measurement[J]. Journal of Empirical finance, 12(3): 445–475.

Mandelbrot B B, (1999). A Multifractal Walk Down Wall Street[J]. Scientific American, 280(2): 70–73.

Martens M, Zein J, (2004). Predicting Financial Volatility: High-Frequency Time-series Forecasts Vis Implied Volatility[J]. Journal of Future markets, 24(11): 1005–1028.

Patton A J, Sheppard K, (2015). Good Volatility, Bad Volatility: Signed Jumps and the Persistence of Volatility[J]. The Review of Economics and Statistic, 97(3): 683–697.

Tibshirani R, (1996). Regression Shrinkage and Selection Via the Lasso[J]. Journal of the Royal Statistical Society, 58(1): 267–288.

Vortelinos D I, (2017). Forecasting Realized Volatility: Har Against Principal Components Combining, Neural Networks and GARCH[J]. Research in International Business and Finance, 39: 824–839.

Yang K, Chen L N, Tian F P, (2015). Realized Volatility Forecast of Stock Index Under Structural Breaks[J]. Journal of Forecasting, 34(1): 57–82.

Yu J, (2005). On Leverage in a Stochastic Volatility Model[J]. Journal of Econometrics, 127(2): 165–178.