基于 GARCH 方法的日内已实现波动率预测研究
——基于沪深 300 指数高频数据的实证分析

吴静雯，田博士

（湖南大学 工商管理学院，湖南 长沙 410082；湖南大学 工商管理学院，湖南 长沙 410082）

摘要：波动率在资产配置、期权定价和风险管理等多个领域起着关键作用。近年来，随着高频数据可获得性的提高，高频日内已实现波动率成为热点研究对象，其对于夏普比率和高频套利都具有重要意义。针对日内已实现波动率的非对称特性，本文提出一个非对称回归模型并引入非对称因子，进一步，运用沪深 300 指数高频数据对该模型进行实证分析，得到日内已实现核波动率能够显著证实非对称效应的存在，且滞后一天的收益率与波动率之间的负相关性最强的结论。在波动率预测方面，本文结合 GARCH 模型和已实现波动率模型，提出 GARCH-RK、ARMA-RK、GARCH-MRC 和 GARCH-RV 四种结合模型对高频金融数据日内波动率进行预测研究，运用沪深 300 指数高频日内已实现波动率序列对上述模型进行实证分析，样本内结果充分表明 GARCH-RK 的预测精度最高。

关键词：高频数据、高频日内已实现波动率、非对称效应、GARCH 模型、波动率预测

中图分类号： F830.91　　文献标识码：A

1.引言

金融资产收益率是金融资产定价研究的核心问题，其呈现以下四种特征：（1）尖峰厚尾特征；（2）异方差性；（3）波动聚集效应；（4）杠杆效应。资产收益率的波动性在金融衍生产品定价、风险管理和投资组合管理中具有重要作用。因此，对金融资产波动率的度量是金融计量领域研究的关键问题。Black 和 Scholes(1973)[1]提出可以通过一个价格公式，如 Black-Scholes 定价公式推出股票价格的波动率，该波动率称隐含波动率。进一步，Black(1976)[2]研究发现波动率对价格大幅上升和大幅下降的反应是不同的，后者对波动率的影响更大，并用非对称效应（又称杠杆效应）解释收益率和波动率的非对称关系。

为了反应收益率与波动率的非对称性，Nelson(1991)[3]基于 GARCH 模型提出了指数广义自回归条件异方差(EGARCH)模型，该模型条件方差为自然对数形式，因此系数可以为负，减弱了对系数的限制，并引入非对称参数，使得条件方差对正负冲击有不同程度的变化，从而可以描述冲击对价格波动的非对称影响。进一步，Glosten(1993)[4]发现在股票市场中，利好消息与利空消息对股票市场波动的影响程度不同，于是用 0 作为门限，大于门限值与小于门限值的冲击对价格扰动会产生不同的影响，从而提出门限广义自回归条件异方差(TGARCH)模型。

近期，朱婧等(2015)[5]将测量风险的 VaR 模型与非对称 GARCH 模型相结合对上证综合指数和深证成份指数进行研究，研究结果发现 EGARCH(1,1)模型能较好地拟合上证综合指数，而 TGARCH(1,1)模型能较好地拟合深证成份指数，从而表明以上两个数据样本均具有非对称性。Horpestad 等人(2019)[6]用 GARCH、Log-GARCH 和异质自回归已实现波动率模型(HAR-RV)三个对称波动率模型与 GJR-GARCH、TGARCH 和 LHAR-RV 三个非对称波动率模型对全

球 19 个股指的非对称波动性进行比较研究。结果表明样本内所有指数均表现出非对称性，且非对称波动率模型具有更好的拟合优度。

传统的 GARCH 模型对日、周、月、季度或者年度的低频数据进行分析，但低频数据由于采样频率低，易造成市场信息缺失，从而使对波动率的估计和预测出现较大偏差。由于信息技术的发展和交易电子化的趋势使高频数据的获取成为可能，高频数据不仅包含资产价格信息，还包括交易的时间间隔、交易量、买卖价差等，这些不同维度的信息对理解市场的价格形成、信息传递机制等市场微观结构方面的特征具有重大意义，国内外学者对于高频数据的研究主要集中在市场微观结构噪音、收益率波动率两个方面。为了对高频数据的日内波动率进行度量，Andersen、Bollerslev 等(1998)[7]将日内对数收益率的平方和定义为已实现波动率(Realized Volatility，RV)。进一步，Andersen、Bollerslev 等(2001)[8]研究指出在无微观结构噪音的假设下，已实现波动率依概率收敛于积分波动率。由于微观结构噪音的存在，使 RV 是积分波动的有偏估计量。在微观结构噪音存在的情况下，为了更好地对日内波动率进行有效估计，Barndorff-Nielsen 等(2008)[9]研究存在简单市场摩擦模型的情况下，如何利用已实现核(Realized Kernels)函数对波动率进行有效可行的推断，从而提出已实现核波动率估计。该方法充分利用高频数据中包含的信息，在最优时间窗口内构造已实现波动率的核函数，通过选择权值以获得最佳收敛速度。另一种方法是 Jacod 等(2009) [10]基于对日内高频数据局部移动提出的预平均已实现波动率，该方法通过移动平均消除噪音的影响，使高频序列更接近有效价格以得到更精确的波动率估计，结果表明在微观结构噪音存在的情况下，基于预平均方法估计的已实现波动率是积分波动率的稳健估计，其最优收敛速率为 $n^{-\frac{1}{4}}$。以上估计方法在金融资产波动率的估计中被广泛使用，如 Christensen 等（2019） [11]提出了金融资产对数收益率波动的非参数估计，对美国股票市场不同领域交易所交易基金的高频交易数据的波动率进行估计，并证明了在存在微观结构噪声的情况下，加入预平均方法后的波动率估计在渐近极限下与积分波动率是一致的。

已有很多学者使用混合频率模型对波动率进行预测，如 Santos 和 Ziegelmann (2014)[12]、Amendola 和 Storti (2015)[13]、Andreou(2016)[14]。为了充分发挥 GARCH 类模型能处理收益率序列异方差效应和已实现波动模型计算简便、无模型以及无偏性的优点，很多学者提出将 GARCH 类模型和已实现波动率进行结合提出混合频率模型对波动率进行研究。Ma 等人(2018)[15]将低频 GARCH 类模型和高频波动率模型如 HAR-RV(Corsi(2009)[16]) 和 ARFIMA-RV 型模型使用四种组合策略结合起来对上证综合指数以及标普 500 指数 5 年的五分钟高频数据进行波动率预测，结果表明在样本外采用组合模型预测精度较高。Zhang 等人(2019)[17]得出了类似的结论，他们基于可预测性动量的模型切换机制提出了一种新的混合频率方法来预测标准普尔 500 指数的波动率。模型置信集检验的结果表明，该混频模型在波动率预测方面明显优于单个模型。因此高频波动率模型与低频波动率模型的结合能够提高波动

率的预测准确性，已有研究使用的高频波动率模型均使为可以有效捕获金融市场中波动率的长记忆和多尺度行为的 HAR-RV 模型。

本文的研究基于高频数据日内波动率模型，利用 GARCH 模型有效检验波动率时间序列异方差的特征，提出 GARCH-日内波动率的模型对积分波动率进行预测。本文提出的方法与已有研究的不同主要体现为：（1）Ma 等(2018)[15]和 Zhang 等人(2019)[17]提出 HAR-RV 对波动率进行预测，但是 HAR-RV 模型中月度波动率没有考虑高频交易基本不隔夜持仓的特征，因此在高频交易中很难对套利策略有效评估。本文分别使用已实现波动率、预平均已实现波动率和已实现核估计波动率对日内积分波动率进行估计。（2）Ma 等人(2018) [18]的研究基于收益率的非对称性，通过指示函数将 RV 分解为正负已实现半方差(realized semi-variance)，研究结果表明加入正负半方差在内的波动率模型为中短期波动率预测提供了更高的准确性，但该研究仅考虑 ARFIMA 模型作为低频模型进行结合，并且 RV 在实际金融市场中存在微观结构噪音的情况下是积分波动的不一致估计。本文研究表明高频已实现日内波动率时间序列具有聚集效应，但是 ARFIMA 模型无法捕捉到该现象。由于高频已实现日内波动率时间序列具有条件方差和聚集效应，因此本文采用 GARCH 模型对波动率时间序列进行建模预测。

本文的主要贡献为：（1）本文使用预平均方法和已实现核方法对高频日内波动率进行估计，提出非对称因子对正负收益率对波动率的影响强度进行度量以研究高频金融时间序列的非对称性。进一步，本文以 2015 年股灾前后沪深 300 指数秒级高频交易数据为研究对象对上述非对称因子进行实证研究，研究发现正负收益率对已实现核估计波动率的非对称影响强度最大，从而说明日内收益率与高频日内波动率具有非对称性。（2）本文基于日内已实现波动率时间序列存在条件异方差性，使用 GARCH 模型对日内已实现波动率建模预测，进一步使用两种损失函数即均方根误差和平均绝对误差对模型的预测能力进行评估，得到非对称因子更大的日内已实现波动率模型能更好的对波动率进行短期预测，从而扩展高频与低频模型结合预测波动率的文献，为高频套利交易提供可靠指导。

本文其余部分的结构如下。第 2 节介绍了本文所使用的模型，包括日内已实现波动率估计模型如二次变差、去除微观结构噪音的已实现核估计法和预平均法以及 GARCH 类模型，第 3 节对本文研究的沪深 300 指数日收益率序列以及日内已实现波动率的统计性描述并引入非对称因子对非对称性强度进行度量，第 4 节展示了高频和低频结合的预测结果，第 5 节简要介绍了本文结论。

2. 模型与方法

2.1 微观结构噪音

在无套利的有效市场假设下，价格过程在概率空间$(\Omega,\mathcal{F},P)$下服从布朗半鞅过程。X_t表示资产在t时刻的对数价格，假设其服从如下连续半鞅模型：

$$X_t = X_0 + \int_0^t \mu_\tau d\tau + \int_0^t \sigma_\tau dW_\tau, t \geq 0$$

其中W是独立的一维布朗运动向量，μ_τ是一个有限方差的$c\grave{a}dl\grave{a}g$漂移过程，σ_τ是X_t的扩散项。二次变差就是已实现波动率，是高频领域中的一种全新的波动率度量方法，在没有微观结构噪音的情况下是积分波动率的无偏估计量，在高频领域中应用广泛。单资产对数价格向量X的二次变差定义如下：

$$RV = \sum_{t=1}^{N-1} (X_{t+1} - X_t)^2 = \sum_{t=1}^{N} \Delta X_t{}^2.$$

在实际中，微观结构噪音使观测到的资产价格偏离真实价格，观测到的价格不是X而是Y。噪音使真实价格失真，考虑过程Y的观测时间点$t = 1, \ldots, N$,

$$Y_t = X_t + \epsilon_t,$$

其中ϵ_t服从独立同分布高斯过程，且X_t与ϵ_t相互独立。由于微观结构噪音的存在，使得抽样频率越高已实现方差偏差越大。因此，采用高频数据估计波动率必须尽可能消除微观结构噪音的影响。

2.2 微观结构噪音处理方法

2.2.1 预平均估计方法

对于零均值独立同分布的微观结构噪音，通过光滑对数价格Y能够消除噪音的影响来近似X，因此 Jacod 等人(2009)[13]提出了预平均估计方法。其原理是通过局部加权平均高频收益来消除市场微观结构噪声，估计量是通过选择满足的整数序列k_n来构造的。对一列整数k_n，$\theta \in (0, \infty)$，$k_n = \theta\sqrt{n} + o(\sqrt[4]{n})$，其中，$n$为观测点数量。$k_n$就是进行预平均的带宽，即选取观测点进行预平均的数量，当k_n非整数时，向上取整。设权重函数$g(x)$为定义在(0，1)区间的分段连续且可微函数，$g(x) = min(x, 1-x)$并且$g(0) = g(1) = 0$，用$g(x)$对价格收益序列ΔY_j进行向前移动加权平均，产生新的收益序列$\overline{Y_i}$

$$\overline{Y_i} = \sum_{j=0}^{k_n-1} g(\frac{j}{k_n}) \Delta Y_{i+j},$$

用预平均后的收益率$\overline{Y_i}$估计积分波动，得到噪声修正的积分波动估计量C_t

$$C_t = \frac{\sqrt{\delta}}{\theta\varphi_2} \sum_{i=0}^{n-k_n+1} (\overline{Y_i})^2 - \frac{\varphi_1 \delta}{2\theta^2 \varphi_2} \sum_{i=1}^{n-1} (\Delta Y_i)^2,$$

其中$\varphi_1 = \int_0^1 (g'(s))^2 ds = 1, \varphi_2 = \int_0^1 (g(s))^2 ds = 1/12$。

2.2.2 调整的预平均估计方法

Christensen 等根据 Jacod 等得到的预平均波动估计量，得到调整已实现方差:

$$MRC = \frac{n}{n-k_n+2}\frac{1}{\varphi_2 k_n}\sum_{i=0}^{n-k_n+1}(\overline{Y}_i)^2,$$

因子$\frac{n}{n-k_n+2}$是MRC中相对于样本大小 n 的真实求和数的有限样本校正。假设k_n是偶数，则

$$\hat{Y}_i = \frac{2}{k_n}\sum_{j=0}^{\frac{k_n}{2}-1} Y_{i+j},$$

这是对于$\frac{k_n}{2}$项Y的一个简单平均，由于这种预平均，$\hat{Y}_i$将更接近有效的价格X_i。 接下来，基于这些滤波增量来计算实现方差估计量，

$$\overline{Y}_i = \frac{1}{2}\left(\hat{Y}_{i+\frac{k_n}{2}} - \hat{Y}_i\right) = \frac{1}{k_n}\left(\sum_{j=\frac{k_n}{2}}^{k_n-1} Y_{i+j} - \sum_{j=0}^{\frac{k_n}{2}-1} Y_{i+j}\right).$$

2.2.3 已实现核估计方法

通过一个核函数捕捉潜在的微观结构噪声引起的收益与收益的串行相关性，Barndorff-Nielsen 等(2008)[12]提出已实现核估计，定义如下：

$$RK(H) = RV + \sum_{h=-H}^{H} K(\frac{h}{H+h})\gamma_h,$$

$$\gamma_h = \sum_{j=|h|+1}^{n} r_j r_{j-|h|},$$

$K(x)$为核函数，即权重函数，H为带宽，γ_h是收益率滞后h阶的已实现自协方差，$h = -H, \cdots, -1, 0, 1, \cdots H$。根据已实现波动率估计中不同降噪方法的比较分析及实证，本文选取有效性度量的数值最接近于 1，即相应估计最有效的 $\text{Tukey} - \text{Hanning}_2$ 核函数：

$$K(x) = sin^2\left\{\frac{\pi}{2}(1-x)^2\right\}.$$

在实证分析中，本文使用表达式$H = 5.76\zeta\sqrt{n}$来选择带宽参数。

2.3 GARCH 模型及其扩展模型

该部分简要介绍本文实证部分使用的经典 GARCH 模型和非对称 GARCH 模型。r_t表示某项资产在t时刻的时间序列数据，

$$r_t = \delta_0 + \sum_{i=1}^{m} \varphi_i r_{t-i} + a_t ,$$

a_t称为该资产在t时刻的扰动或新息，a_t是一个白噪声，a_t=$\sigma_t \varepsilon_t$且服从$N(0, {\sigma_t}^2)$。其中$\{\varepsilon_t\}$是均值为 0、方差为 1 的独立同分布(iid)随机变量序列。本文以以下方式对波动率进行建模：

（1） Bollerslev(1986)提出的广义 ARCH(GARCH)模型：

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

（2） Nelson(1991)提出的指数 GARCH(EGARCH)模型：

$$\ln(\sigma_t^2) = \alpha_0 + \alpha_1 \frac{|a_{t-1}|}{\sigma_{t-1}} + \gamma_1 \frac{a_{t-1}}{\sigma_{t-1}} + \beta_1 \ln(\sigma_{t-1}^2),$$

（3） Glosten(1993)等提出的门限 GARCH(TGARCH)模型：

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \gamma_1 N_{t-1} a_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

其中，N_{t-1}是关于α_{t-1}是否为负值的指示变量，即α_{t-1}＜0时$N_{t-1} = 1$。

3.高频波动率非对称性检验实证分析

3.1 数据选择与预处理

沪深 300 指数于 2005 年 4 月 8 日由上海证券交易所和深圳证券交易所联合发布，包含了上市交易 A 股中的 300 只样股，其样本覆盖了沪深市场约 60%的市值，具有良好的市场代表性。沪深 300 指数的推出，意味着中国证券市场有了反映整个 A 股市场全貌的指数。2015 年一场世界金融史上罕见的系统性崩溃骤然降临于中国大陆 A 股市场，这期间金融市场波动性较大，为此，本文以国泰安高频数据库沪深 300 指数 2014 年 1 月 2 日至 2016 年 12 月 30 日共 733 天的秒级高频数据为原始数据。受熔断机制影响，2016 年 1 月 7 日 A 股实际交易时间仅为 13 分钟，本文剔除该日的高频交易信息，进一步结合中国股市交易规则，参考 Barndorff-Nielsen 等使用的清洗程序，删除时间戳在交易所正常交易时间(9:30am-11:30am 和 13:00pm-15:00pm)以外的交易数据；为避免开盘效应，删除每个交易日开盘前 15 分钟的交易数据，最终共有 732 个交易日的高频数据参与本实证。

3.2 高频波动率特征分析

本节以 2015 年股灾前后沪深 300 指数秒级高频交易数据为研究对象，使用预平均方法、已实现核方法以及二次变差对高频日内波动率进行估计，提出非对称因子对正负收益率对波动率的影响强度进行度量以此研究高频金融时间序列的非对称性。

3.2.1 收益率的描述性统计

对收益率作时序图，如图 3.1 所示。由收益率时序图可知：波动在一些时间段内非常大（例如从 2015 年 4 月到八月），在有的时间段内（例如从 2014 年 4 月到 10 月和 2016 年 4

月到 7 月）相对较小，波动率聚集效应明显。并对该股指收益率进行统计描述性得到表 3.1，由表可知：收益率呈现出明显的“尖峰厚尾”形态，且 J-B 检验值拒绝该收益率序列服从正态分布的假设。一个序列是否平稳会极大地影响其行为和性质，运用非平稳数据会导致伪回归，而由表 3.1 的 ADF 检验值可知收益率序列平稳。

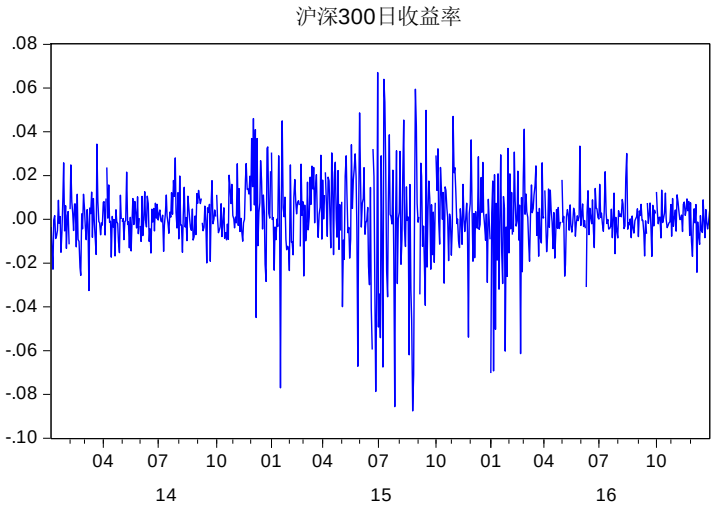

图 3.1　收益率时序图

表 3.1 收益率描述性统计表

	标准差	偏度	峰度	Jarque-Bera	ADF
沪深 300	0.179	-0.883	7.69	766.30***	-25.48***

（注：J-B 为 Jarque-Bera 的正态分布检验；ADF 检验采用无趋势项无常数项形式；***表示在 1%水平内显著拒绝原假设）

3.2.2 已实现波动率统计性描述

本小节运用 2.2 节提及的二次变差、调整预平均方法和已实现核方法三种日内波动率估计方法，得到沪深 300 指数的高频日内波动率，该过程结果均由 R 语言操作得出。由于估计方法不同，各波动率模型得到的结果具有不同的数量级，如果直接用原始值进行对比分析，可能会削弱数量级低的波动率在分析中的作用，因此，为保证对比结果的可靠性，需对原始波动率序列进行标准化处理。对原始数据进行 min-max 标准化(Min-max normalization)线性变换，使结果落到[0,1]区间，得到新的高频日内波动率序列，使之能在一个水平上进行比较分析。

图 3.2（a）、（b）、（c）、（d）为沪深 300 指数的二次变差、调整后预平均波动率以及已实现核估计波动率结果的直观对比图，由图可以看出三个序列均具有聚集效应，且已实现核估计和调整后预平均波动率的表现情况相似，但二次变差在收益率序列相对平稳时期（2014 年 1 月至 2017 年 7 月）的波动率较大，对波动率的描述不如已实现核估计和调整后预平均波动率。

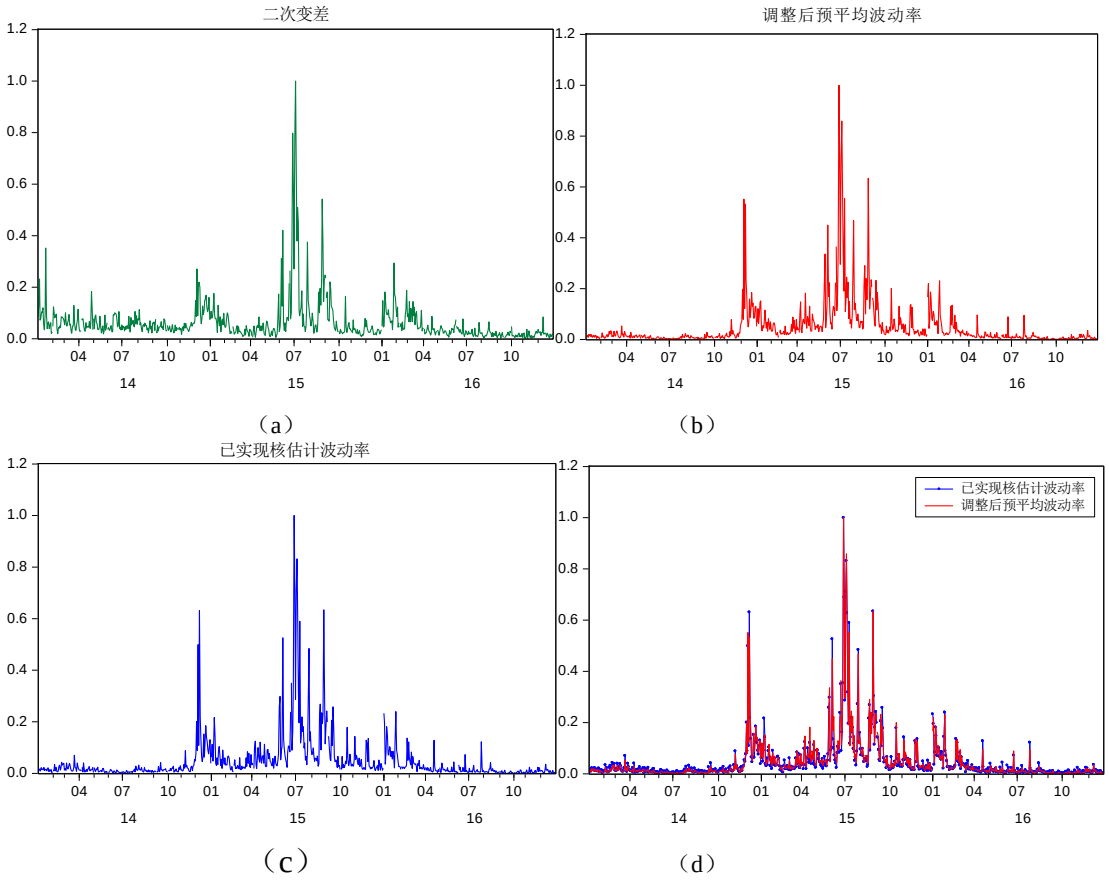

图 3.2 沪深 300 指数波动率对比图

表 3.2 为三种高频日内波动率的统计特征描述，可知：波动率均值均趋于 0，偏度都大于 0，说明序列分布有长的右拖尾；峰度高于正态分布的峰度值 3，说明收益率序列具有尖峰厚尾特征并且 J-B 统计量的值显著拒绝该序列服从正态分布的假设；单位根检验统计量的绝对值都显著大于临界值，所以波动率序列均为平稳时间序列，后续可直接进行建模，无需做进一步差分处理。

表 3.2 已实现波动率统计特征

	均值	偏度	峰度	Jarque-Bera	ADF
已实现核估计	0.063	5.564***	45.410***	58714.08***	-6.634***
调整后预平均	0.049	5.222***	37.985***	40711.35***	-5.870***
二次变差	0.049	5.117***	37.526***	39065.55***	-5.620***

3.2.3 波动率的非对称性检验

因我国股市存在非对称效应，使用 EGARCH 模型和 TGARCH 模型对收益率序列估计。由于高维 GARCH 模型其参数呈几何形式上涨，而且一般实证检验中也只用到 GARCH(1,1)模型，因此，这里主要考虑 EGARCH(1,1)模型和 TGARCH(1,1)模型。

表 3.3 非对称模型系数

	α_0	α_1	γ_1	β_1

沪深 300	TGARCH	0.001***	0.073***	0.024***	0.937***
		(0.000)	(0.011)	(0.013)	(0.007)
	EGARCH	-0.113***	0.122***	-0.019***	0.997***
		(0.023)	(0.014)	(0.010)	(0.002)

（注：***表示在 1%显着性水平上拒绝零假设。）

表 3.3 为沪深 300 指数日收益率 EGARCH(1,1)模型和 TGARCH(1,1)模型估计结果，由表可知： TGARCH(1,1)模型中的非对称性参数γ_1大于 0 且高度显著，说明与相同程度的正向冲击相比，负向冲击会导致下一期出现更大的条件方差; EGARCH(1,1)模型中非对称性参数γ_1也高度显著，其符号为负，但可以得出相同结论。以上实证研究表明，沪深 300 指数收益率存在非对称效应，投资者对收益率下跌的反应往往高于相同程度收益率上涨的反应，即收益率的下跌对市场的影响更大。为了直观描述用日数据得到的非对称 GARCH 模型波动率和用日内数据得到的日内高频已实现波动率的优劣，对非对称 GARCH 模型波动率进行归一化处理，使之能与日内已实现波动率在同一水平进行对比，并将其与收益率进行做图比较。

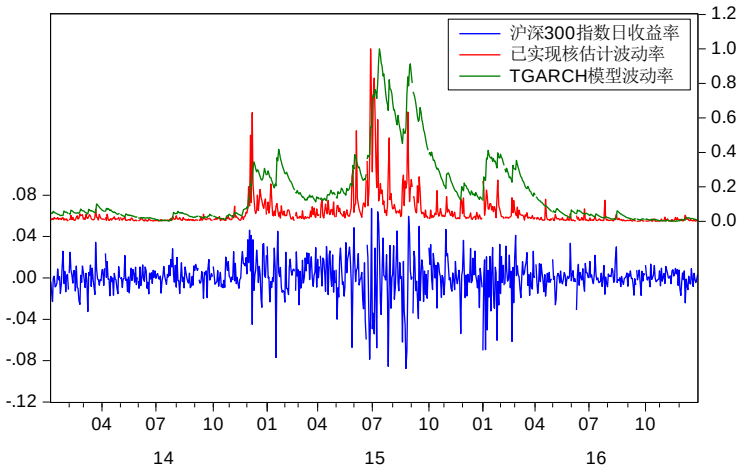

图 3.4 沪深 300 指数波动率收益率对比图

由图 3.4 的对比图可看出，在非对称 GARCH 模型中，方差预测结果为线性，波动率对于收益率正负变化的反应并不及时，当短期内收益率变化较大时，由于波动率估计结果是线性变化并不能很快地对此做出相应改变。并且在图 3.4 中，当收益率变化相对平稳时期，如2014年一月至三月，非对称GARCH模型高估了波动率的值，已实现波动率比非对称GARCH模型得出的波动率表现更好。

3.3.4 非对称性检验

非对称 GARCH 模型虽能体现样本股的非对称性，但已实现波动率比非对称 GARCH 模型得出的波动率在对收益率变化的反应情况上表现更好。本小节考虑日内波动率是否也能体现收益率与波动率之间的非对称关系。选取国泰安数据库中沪深 300 指数的日个股回报率，将其按天分为正收益率(r^+)和负收益率（r^-）（如若当天收益为-0.0727，则处理后负收益率序列的当天值为-0.0727，正收益率序列当天值以 0 填充）。

表 3.4 正负收益率的描述性统计

	均值	标准差	偏度	峰度	Jarque-Bera	ADF
r^+	0.006	0.010	2.362***	9.572***	2000.805***	-24.061***
r^-	-0.006	0.012	-3.693***	19.006***	9490.602***	-9.427***

（注：J-B 统计量(Jarque&Bera，1987)检验了序列分布中正态性的零假设。 ADF 是基于最小 AIC 的增强 Dickey-Fuller 统计量(Cheung&Lai，1995)。***表示在 1%显着性水平上拒绝零假设。)

为描述不同滞后天数正负收益率对于当期日内高频已实现波动率的影响程度，本文在回归中同时使用当天、滞后一天至滞后四天的正负收益率数据分别对三种日内已实现波动率进行回归估计，回归模型如下：

$$\sigma_t^2 = C + C_1 * r_t^+ + C_2 * r_t^- + C_3 * r_{t-1}^+ + C_4 * r_{t-1}^- + C_5 * r_{t-2}^+ + C_6 * r_{t-2}^- + C_7 * r_{t-3}^+ + C_8 * r_{t-3}^- + C_9 * r_{t-4}^+ + C_{10} * r_{t-4}^- + \varepsilon_t,$$

其中 C 是常数，ε_t是一个白噪声，且$\varepsilon_t \sim N(0, \sigma_t^2)$。进一步对模型变量使用逐步回归模型进行确定，经过逐步回归可使最后保留在模型中的解释变量既是重要的，又没有严重多重共线性，最终回归结果保留的解释变量有r_t^+、r_t^-、r_{t-1}^+、r_{t-1}^-、r_{t-2}^+、r_{t-2}^-，即最终模型为：

$$\sigma_t^2 = C + C_1 * r_t^+ + C_2 * r_t^- + C_3 * r_{t-1}^+ + C_4 * r_{t-1}^- + C_5 * r_{t-2}^+ + C_6 * r_{t-2}^- + \varepsilon_t.$$

为直观度量不同解释变量对波动率的具体影响，引入非对称因子ρ_1、ρ_2和ρ_3来分别度量当天、滞后一天和滞后两天的正负收益率对于波动率的不同影响程度。$\rho_1 = C_2/C_1$表示当天的负收益率对波动率影响C_2与正收益率对波动率影响C_1的程度之比，因非对称效应的存在，ρ_1应小于-1。同理，$\rho_2 = C_4/C_3$和$\rho_3 = C_6/C_5$分别表示滞后一天和滞后两天的负收益率与正收益率对波动率影响程度之比。

表 3.5 回归估计结果

	沪深 300 指数		
	RK^*	MRC	RV
C	-0.024***	0.007*	-0.024***
	(0.004)	(0.004)	(0.004)
C_1	1.234***	0.539***	0.968***
	(0.261)	(0.291)	(0.260)
C_2	-2.412***	-1.588***	-2.178***
	(0.214)	(0.213)	(0.213)
C_3	1.807***	0.933***	1.971***
	(0.251)	(0.250)	(0.250)
C_4	-3.814***	-3.219***	-3.890***
	(0.214)	(0.213)	(0.213)
C_5	1.595***	1.391***	1.786***
	(0.248)	(0.247)	(0.247)

C_6	-1.775***	-2.007***	-1.900***
	（0.224）	（0.222）	（0.222）
R^2	0.539	0.442	0.471
ρ_1	-1.955	-2.946	-2.25
ρ_2	-2.119	-3.450	-1.974
ρ_3	-1.113	-1.443	-1.064

（注：***表示在 1%显着性水平上拒绝零假设）

表 3.5 为回归估计结果，由表可知已实现核估计波动率的回归拟合优度最大，为 0.539。进一步，在已实现核估计波动率拟合最优的情况下，对比当前、滞后一天、滞后两天的非对称因子，发现滞后一天的非对称因子即ρ_2显著小于-1 且小于ρ_1和ρ_3，前一天的负收益率比正收益率对波动率的冲击更大，说明高频股票数据的去噪波动率估计方法中，已实现核估计能够更好的拟合沪深 300 指数的积分波动，且由于滞后一天非对称因子值更小，更能体现收益率与波动率的非对称效应，所以在后续预测部分使用超前一步预测。

4.高频波动率预测分析

4.1 数据与描述性统计

由本文第三节实证结果得出已实现核估计波动率对于非对称效应体现最优，因此本节以沪深 300 指数自 2014 年 1 月 2 日到 2016 年 12 月 30 日的归一化后日内已实现核估计波动率共 732 个数据为样本进行预测。

表 4.1 波动率描述性统计

	均值	标准差	偏度	峰度	Jarque-Bera	ADF
已实现核估计	0.049	0.094	5.222***	37.985***	40711.35***	-5.870***

（注：J-B 统计量(Jarque&Bera，1987)检验了序列分布中正态性的零假设。 ADF 是基于最小 AIC 的增强 Dickey-Fuller 统计量(Cheung&Lai，1995)。***表示在 1%显着性水平上拒绝零假设。）

表 4.1 为沪深 300 已实现核估计波动率序列的描述性统计，由表可知：样本均值较小，偏度显著大于 0，说明序列为右偏；且峰度显著大于 3，意味着序列具有厚尾；J-B 值表明收益率序列为显著非正态分布；ADF 检验结果表明在 1%显著性水平下，拒绝原假设，认为序列平稳，可以直接建模而无须进一步变换。

4.2 建模与预测

为检验高频日内波动率作为时间序列是否具有自相关性以及为后续波动率预测的均值方程建模，本节对其进行自相关与偏自相关函数检验。具体检验结果如下图 4.1 和 4.2 所示。从 ACF 图可以看出序列出现拖尾现象；而偏自相关系数在三阶或者五阶存在较强的相关性，因此进行 AR(1)至 AR(5)建模。

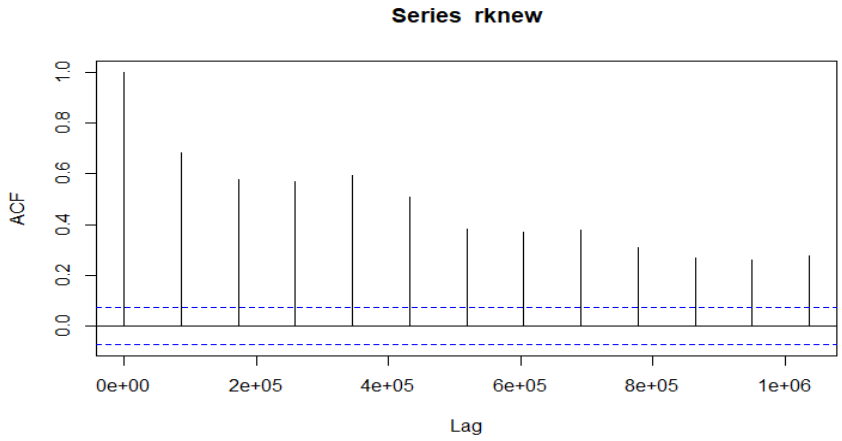

图 4.1 已实现核估计自相关系数

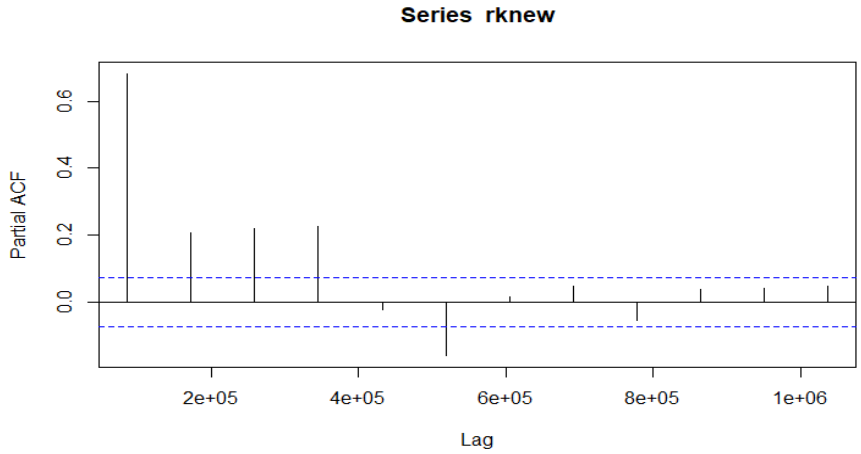

图 4.2 已实现核估计偏自相关系数

表 4.2 不同模型系数估计结果

	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
φ_0	0.048***	0.049***	0.049***	0.049***	0.049***
	（0.008）	（0.010）	（0.012）	（0.016）	（0.015）
φ_1	0.683***	0.540***	0.494***	0.444***	0.449***
	（0.027）	（0.036）	（0.036）	（0.036）	（0.037）
φ_2		0.208***	0.088**	0.068*	0.071*
		（0.036）	（0.040）	（0.039）	（0.040）
φ_3			0.222***	0.110***	0.112***
			（0.036）	（0.040）	（0.040）
φ_4				0.226***	0.235***
				（0.036）	(0.040)
φ_5					-0.021
					(0.037)
ARCHtest	37.032***	28.001***	17.342***	16.128***	16.258***
R^2	0.466	0.489	0.514	0.539	0.539
AIC	-2.510	-2.549	-2.596	-2.645	-2.641

(注：本表显示波动率的均值方程建模估计结果。 ***、**和*分别表示在 1%、5%和 10%显著性水平上拒绝零假设。)

表 4.2 为不同模型系数估计结果，可知 AR(5)与 AR(4)系数的显著性相似，但滞后五阶的自相关系数不显著，因此对于波动率建模的均值方程选取 AR(4)。又因 ARCH 检验结果均显著拒绝原假设，说明残差存在自回归条件异方差结构，因此本文首次使用 GARCH 模型对高频日内已实现核估计模型进行建模，并提出 GARCH-RK 模型，其表述如下：

$$RK_t = \varphi_0 + \varphi_1 RK_{t-1} + \varphi_2 RK_{t-2} + \varphi_3 RK_{t-3} + +\varphi_4 RK_{t-4} + a_t ,$$

$$a_t = \sigma_t \varepsilon_t ,$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

a_t为已实现核估计波动率在t时刻的扰动或新息。{ε_t }是均值为 0、方差为 1 的独立同分布(iid)随机变量序列，且假定ε_t服从标准正态分布或标准学生t分布。

本文同时考虑前文得到的另外两种日内高频波动率估计，并进一步构建 GARCH-MRC、GARCH-RV 以及 ARMA-RK。对以上四种模型通过添加新的一天和删除估计窗口的第一天来将估计周期向前滚动对波动率进行滚动预测，可以保证用于估计模型的样本量大小保持不变，且预测不重叠。滚动窗口为 250 个样本点，即 1-250 个样本点预测第 251 个数据点，2-251 个样本点预测第 252 个数据点，以此类推，最终得到预测结果。将预测结果与波动率序列进行对比得到图 4.3 和图 4.4，图 4.3 为 GARCH-RK 模型预测结果、图 4.4 为 GARCH-RV 模型预测结果，可直观比较出在波动率较平稳时期 GARCH-RV 模型波动率预测的结果拟合度并不好，为了定量比较各模型预测能力的显著差异，采用以下两个损失函数：

$$RMSE = \sqrt{\sum_{i=251}^{n} (\sigma_i^2 - \hat{\sigma}_i^2)^2} ,$$

$$MAE = \sum_{i=251}^{n} |\sigma_i^2 - \hat{\sigma}_i^2|,$$

其中：σ_i^2为原时间序列差分的平方，作为波动率的代理变量；$\hat{\sigma}_i^2$为模型估计出的条件方差。均方根误差（$RMSE$）和平均绝对误差（MAE）的值越小，说明预测值和实际值的差异越小，模型的预测效果也越好。

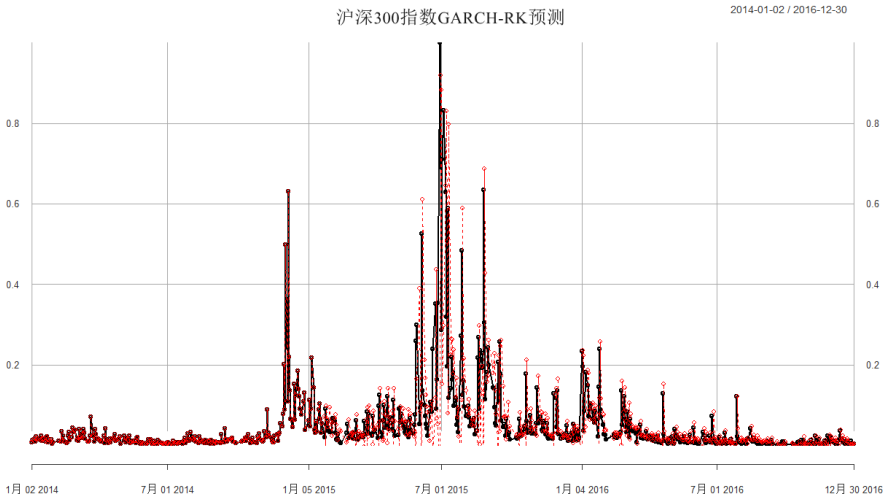

图 4.3 已实现核估计波动率预测结果

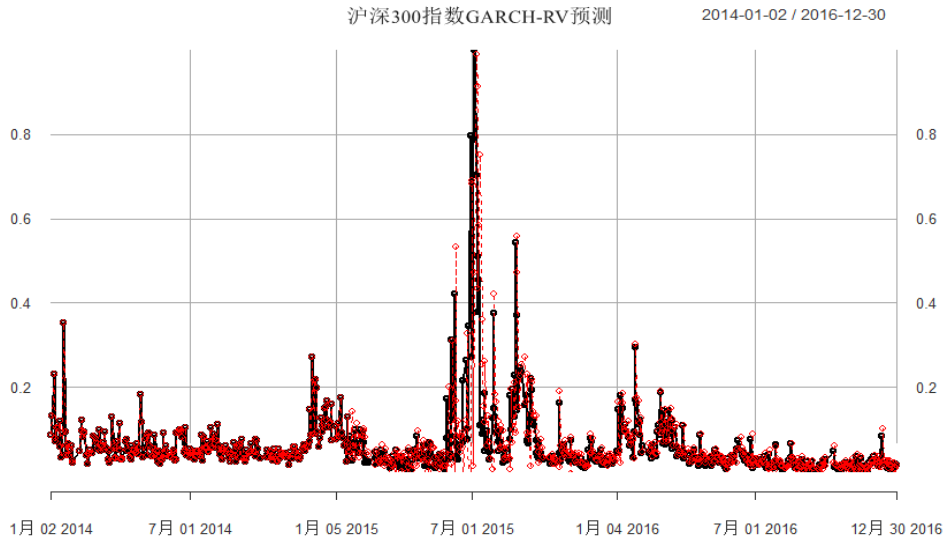

图 4.4 二次变差预测结果

表 4.4 沪深 300 指数预测结果

	GARCH-RK	ARMA-RK	GARCH-MRC	GARCH-RV
$RMSE$	0.065	0.078	0.083	0.089
MAE	0.027	0.032	0.037	0.040

表 4.4 为四种模型预测结果，由表可知：对于沪深 300 指数，使用 GARCH-RK 模型进行预测时，$RMSE$和MAE两个损失函数值最小，该模型对波动率的预测效果最优。

5.结论

对波动率准确地预测是风险管理、投资组合管理和资产定价等领域研究的重要问题，并且与日常和低频数据相比高频数据包含更丰富的市场信息，因此，基于高频数据的研究是高频金融计量研究的热点。本文基于高频数据日内波动率模型，引入非对称因子对正负收益率

对不同日内波动率模型的影响强度进行度量。进一步利用 GARCH 模型有效检验波动率时间序列异方差的特征，提出 GARCH-日内波动率的模型对积分波动率进行预测，得出以下结论：首先，基于沪深 300 指数高频金融数据的实证结果表明，使用核估计方法对微观结构噪音进行处理后得到的日内已实现核波动率能够显著地证实非对称效应的存在，且滞后一天的收益率与波动率之间的负相关性最强。其次，高频日内已实现波动序列表现出显著异方差性，支持对已实现波动率进行 GARCH 模型建模。基于沪深 300 指数高频已实现波动率的样本内实证结果表明，使用 GARCH-RK 模型对波动率进行预测具有较高的预测精度，有助于对沪深 300 指数的日内波动进行预测。最后，本文的研究为使用高频波动率模型与低频波动率模型结合进行股票收益率的波动性预测提供新的方法，对基于高频数据进行投资组合、风险管理和高频套利等决策的相关人员提供有利的方法支撑。

参考文献

[1] Black F,Scholes M S . The Pricing of Options and Corporate Liabilities[J]. Journal of Political Economy, 1973, 81(3):637-654.

[2] Black F."Studies of Stock Price Volatility Changes."[J]. proceedings of the 1976 meetings of the business& economics statistics section, american statistical association, 1976, 81:177-181.

[3] Nelson, Daniel. Conditional Heteroskedasticity in Asset Returns: A New Approach[J]. Modelling Stock Market Volatility, 1991, 59(2):347-0.

[4] Glosten R, Jagannathan R, Runkle E. On the relation between the expected value and the volatility of the nominal excess return on stocks[J]. Journal of Econometrics, 1993, 48(5):1779-1801.

[5] 朱婧, 张静, 付云鹏. 基于非对称 GARCH 模型的股票指数 VaR 的度量与实证分析[J]. 市场周刊, 2015.

[6] Horpestad J, Lyocsa S, P Molnár, et al. Asymmetric Volatility in Equity Markets Around the World[J]. Social Science Electronic Publishing.

[7] Andersen T G, Bollerslev T. Answering the Critics: Yes, ARCH Models Do Provide Good Volatility Forecasts[J]. NBER Working Papers, 1997.

[8] Andersen T G, Bollerslev T, Diebold F X, et al. The distribution of realized stock return volatility[J]. Journal of Financial Economics, 2001, 61(1):43-76.

[9] Barndorff-Nielsen O E, Hansen P, Lunde A , et al. Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise[J]. Econometrica, 2008, 76(6):1481-1536.

[10] Jacod J, Li Y, Mykland P A, et al. Microstructure noise in the continuous case[J]. Technical Reports, 2009, 119(7):2249-2276.

[11] Christensen K, yrsgaard M, Eliyev B. The realized empirical distribution function of stochastic

variance with application to goodness-of-fit testing[J]. Journal of Econometrics, 2019, 212.

[12] Santos D G, Ziegelmann F A. Volatility Forecasting via MIDAS, HAR and their Combination: An Empirical Comparative Study for IBOVESPA[J]. Journal of Forecasting, 2014, 33(4):284–299.

[13] Amendola A, Storti G. Model Uncertainty and Forecast Combination in High-Dimensional Multivariate Volatility Prediction[J]. Journal of Forecasting, 2015.

[14] Andreou, Elena. On the use of high frequency measures of volatility in MIDAS regressions[J]. Journal of Econometrics, 2016, 193(2):367-389.

[15] Ma F, Li Y, Liu L, et al. Are low-frequency data really uninformative? A forecasting combination perspective[J]. North American Journal of Economics & Finance, 2017, 44(APR.):92-108.

[16] Corsi F. A Simple Approximate Long-Memory Model of Realized Volatility[J]. Journal of Financial Econometrics, 2004(2):174-196.

[17] Zhang Y, Ma F, Wang T, et al. Out‐of‐sample volatility prediction: A new mixed‐frequency approach[J]. Journal of Forecasting, 2019.

[18] Ma F, Wei Y, Chen W, et al. Forecasting the volatility of crude oil futures using high-frequency data: further evidence[J]. Empirical Economics, 2017.

Research On Intra-day Realized Volatility Prediction Based on GARCH Method
——Empirical analysis based on high frequency data of CSI 300 index

WU Jing-wen, TIAN Bo-shi

(Business School, HUNAN University, Changsha 410082, China; Business School, HUNAN University, Changsha 410082, China)

Abstract: Volatility plays an important role in asset allocation, option pricing, and risk management. In recent years, with the improvement of the availability of high-frequency data, high-frequency intraday volatility has become a hot research object, which is of great significance for sharp ratio and high-frequency arbitrage. Aiming at the asymmetric characteristics of intraday realized volatility, we propose an asymmetric regression model and introduce an asymmetric factor. Further, we use the high-frequency data of CSI 300 index to make an empirical analysis of the model. It is concluded that intraday realized nuclear volatility can significantly confirm the existence of asymmetric effect, and the negative correlation between one-day lagged return and volatility is the strongest. In the aspect of volatility prediction, this paper combines GARCH model and realized volatility model, and proposes four combined models to predict intraday volatility of high-frequency financial data. The empirical analysis of the above models is carried out by using the high-frequency intraday realized volatility series of CSI 300 index. The results in the sample fully show that GARCH-RK has the highest prediction accuracy.

Keywords: High frequency data; High frequency intraday realized volatility; Asymmetric effect; GARCH model; Volatility prediction