文章编号:1671-4229(2023)03-0075-08

高频 GARCH 模型的最优抽样分析

程凌筠[a]，宋泽芳[a,b*]，张兴发[a,b]，李莉丽[a]
(广州大学　a. 经济与统计学院; b. 岭南统计科学研究院, 广东 广州　510006)

摘　要: 波动率代表是运用高频数据估计日频 GARCH 类模型时构造的一个重要统计量,文章针对运用高频数据估计日频 GARCH 模型的 3 种方法,基于估计量的渐近结果讨论了最优波动率代表的选择问题,并展开了在日内高频数据抽样中应用的讨论。通过对沪深 300 指数的高频数据实证分析发现:同一波动率代表在不同抽样频率下的表现有明显差异;在同一频率下,不同波动率代表有优劣之分;在不同估计方法下,每一个波动率代表的最优频率都不同。因此,日内高频数据的最优抽样频率应针对模型所用的不同估计方法加以区别对待。

关键词: 波动率代表; 抽样频率; GARCH 模型; 高频数据
中图分类号: O 212.1　　**文献标志码:** A

Optimal sampling analysis for GARCH model with high frequency data

CHENG Ling-jun[a], *SONG Ze-fang*[a,b*], *ZHANG Xing-fa*[a,b], *LI Li-li*[a]
(a. School of Economics and Statistics; b. Lingnan Research Institute of Statistical Science, Guangzhou University, Guangzhou 510006, China)

Abstract: Volatility proxies are an important statistic applied to estimate daily GARCH model by using high frequency data. This paper proposes the criterions for choosing an optimal volatility proxy by the asymptotic properties of three GARCH estimators based on the intraday high frequency data, and the applications of these criterions in high frequency data sampling are also discussed. The empirical study of the high-frequency data of CSI 300 index shows that: for the same volatility proxy, the performance of different frenquencies is obviously different; for the same frequency, different volatility proxies perform differently; each volatility proxy has a different optimal frequency under different estimation methods. Consequently, the optimal sampling frequency for intraday high frequency data should be treated differently among different estimation approaches.

Key words: volatility proxy; sampling frequency; GARCH model; high frequency data

作为资产收益变异程度的一种定量测度,波动率在金融时间序列的不同领域中都扮演着相当重要的角色,如衍生产品定价、对冲投资决策或风险价值(VaR)评估等,都与波动率密切相关。Engle[1]和 Bollerslev[2]提出的自回归条件异方差模型,即(G)ARCH 模型,是目前最成熟、最常用的波动率建模模型之一。它已被广泛用于刻画和预测股票价格、商品期货、通货膨胀率和外汇等金融产品的波动率。随着计算机技术的飞速发展,采集、存储数据的成本不断降低,日内高频金融数据的获取也越来越方便。如何使用这类数据推进金融市场波动率的研究成为备受关注的焦点。其中

收稿日期: 2021-12-31; **修回日期:** 2022-05-25
基金项目: 国家自然科学基金重点资助项目(11731015);国家自然科学基金青年资助项目(11701116);广州市基础与应用基金研究资助项目(202201010276)
作者简介: 程凌筠(1997—),女,硕士研究生. E-mail:www.xsn@qq.com
* 通信作者. E-mail:song_zefang@163.com
引文格式: 程凌筠, 宋泽芳, 张兴发, 等. 高频 GARCH 模型的最优抽样分析[J]. 广州大学学报(自然科学版),2023,22(3):75-82.

一个方向是对日内波动率的刻画,常见的是基于日内高频交易数据所估计的日内真实波动率,通常称为已实现波动率。学者们运用不同的非线性度量方法提出了许多已实现波动率指标,如已实现方差、已实现双幂次变差和已实现极差等[3-5]。另一个方向是利用日内高频数据来改进日频波动率模型的估计精度。Visser[6]将日内高频信息引入到GARCH模型中,提出尺度模型和波动率代表模型,改进了GARCH模型参数估计的渐近方差,提高了估计的准确性。越来越多的研究也表明,由于高频数据蕴含了更丰富的资产价格变动信息,将高频信息引入低频波动率模型可以有效地提高模型参数的估计精度[7-13]。

在进行高频数据分析处理时,首要面对的是数据抽样问题,不同的频率会对估计的准确性造成不同的影响,即抽样频率过高容易有太多噪音,抽样频率过低又没有充分利用信息,因而关于抽样频率的择优选取就显得尤为重要。徐正国等[14]定义了微观结构误差(MSE)作为最优抽样频率的选择准则,其实证研究表明已实现波动率估计在10 min间隔的抽样频率下MSE达到最小。郭名媛等[15]则考虑MSE和测量误差之和为择优标准,以总误差最小的60 min为最优抽样间隔来计算赋权已实现波动率。唐勇等[16]分别依据已实现波动和已实现极差波动与积分波动之间误差项的渐近分布,给出了最优抽样频率的选择方法。李胜歌等[17]基于已实现双幂次变差和赋权已实现波动,给出了最优抽样频率选择方法。闵素芹等[18]比较了3种已实现波动率的最优抽样频率选择方法。杨建辉等[19]研究了不同抽样间隔下创业板指数已实现波动的分布特征及其最优采样间隔。

已有的这些研究中,学者们讨论高频数据的最优采样间隔大多是针对日内波动率刻画进行考虑,鲜有考虑日频波动率模型。本文进一步研究高频数据应用到日频波动率模型(GARCH)时的数据抽样问题。与传统的研究不同,本文的最优频率抽样问题可以比较方便地通过选择最优波动率代表来进行刻画。波动率代表是运用高频数据估计日频GARCH类模型时构造的一个重要统计量,它是由日内高频数据信息通过加工构造出的一个函数,不同的波动率代表对参数估计效果有直接影响。本文结合GARCH模型的3种估计方法,即基于对数正态分布的拟极大似然估计(log-Gaussian QMLE)、基于正态分布的拟极大似然估计(Gaussian QMLE)和基于拉普拉斯分布的拟极大指数似然估计(QMELE),讨论不同估计方法下最优波动率代表的选择问题及其在高频数据抽样频率的选择问题。

1 波动率代表模型

在日频数据下,使用GARCH(1,1)模型对波动率进行建模,其形式为

$$y_t = \sigma_t \varepsilon_t, \tag{1}$$

$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2, \tag{2}$$

其中,y_t为资产第t天的收益率;ε_t是服从均值为0,方差为1分布的一组独立同分布随机误差项,分布未知;对于$\forall t \geqslant s$,ε_t与y_s相互独立;参数$\omega > 0$,$\alpha \geqslant 0$,$\beta \geqslant 0$保证条件方差的非负性。

假设每天可观测到的金融资产价格过程为$P_t(u)$,$t = 1, \cdots, T$,将日内的交易时间设为$[0,1]$区间,$0 \leqslant u \leqslant 1$。当$u = 1$时,$P_t(1)$恰为第$t$天收盘价。定义第$t$天$u$时刻的高频对数收益率为

$$Y_t(u) = 100 \times [\log P_t(u) - \log P_{t-1}(1)],$$

即日内的收益过程。在模型(1)~模型(2)的基础上,Visser[6]考虑利用日内收益过程对日频GARCH模型进行扩展,得到如下尺度模型:

$$Y_t(u) = \sigma_t Z_t(u), u \in [0,1], \tag{3}$$

$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2, \tag{4}$$

其中,$Z_t(\cdot)$为标准过程,与σ_t相互独立;$\forall t \neq s$,$Z_t(\cdot)$与$Z_s(\cdot)$独立同分布;σ_t称为尺度参数。当$u = 1$时,$Y_t(1) = y_t$,$Z_t(1) = \varepsilon_t$,模型(3)~模型(4)即退化为日频GARCH模型。

在一般情形下,模型(3)~模型(4)含有高频数据$Y_t(u)$,$Z_t(u)$和低频数据σ_t^2,y_{t-1}^2,无法直接估计。因此,在尺度模型的基础上,Visser[6]进一步提出了波动率代表模型。一般地,波动率代表为正,且满足正齐次性。如果过程$Y_t(u)$乘以因子ρ,其中,$\rho > 0$,则H有

$$H(\rho Y_t(u)) = \rho H(Y_t(u)) > 0, \ \forall \rho > 0,$$

(第3.2节有举例详细介绍H函数)对每个交易日,根据正齐次性,由式(3)可得

$$H_t = H(Y_t(u)) = H(\sigma_t Z_t(u)) = \sigma_t H(Z_t(u)),$$

记$z_{H,t} = H(Z_t(u)) > 0$,由于$Z_t(u)$是独立同分布

的标准过程,因此,$z_{H,t}$是独立同分布的随机变量序列。于是,波动率代表模型可表达为

$$H_t = \sigma_t z_{H,t}, \tag{5}$$

$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2。 \tag{6}$$

上述模型中,所有变量都是同样的频率,日内高频数据信息体现在波动率代表量 H_t 上。当波动率代表量 $H_t = H(Y_t(u)) = H(Y_t(1)) = |Y_t(1)| = |y_t|$时,可以看出模型(5)~模型(6)和模型(1)~模型(2)是等价的。一般情形下,模型(5)~模型(6)引入了高频数据信息,同时又和模型(1)~模型(2)具有相同的模型参数。因此,基于波动率代表模型(5)~模型(6)估计的参数,用到了更多的信息,有望得到更为精确的估计。

2　估计方法

对于波动率代表模型,现有的研究主要采用 3 种估计方法对模型中的未知参数进行估计,分别为对数正态分布拟极大似然估计、正态分布拟极大似然估计和拟极大指数似然估计,下面依次介绍 3 种方法对模型估计的过程,及估计量的渐近结果。

2.1　对数正态分布拟极大似然估计

Visser[6] 给出了基于对数正态分布的拟极大似然估计(log-Gaussian QMLE)及其估计量的渐近正态结果,该方法是将拟极大似然估计(Gaussian QMLE)应用于对数波动率代表 $\log(H_t)$。

记模型(2)的待估参数向量为 $\theta = (\omega, \alpha, \beta)'$,假设 $Ez_{H,t}^2 = \mu_H$,并令 $\varepsilon_t^* = z_{H,t} / \sqrt{\mu_H}$,由模型(5)可得

$$H_t = \sigma_t z_{H,t} = \sigma_t \sqrt{\mu_H} \cdot \varepsilon_t^*。 \tag{7}$$

对式(7)取对数,可得 $\log(H_t) = \log(\sigma_t \sqrt{\mu_H}) + \log(\varepsilon_t^*)$。

定义

$$\bar{\sigma}_t = \sigma_t \sqrt{\mu_H} \cdot \exp(E\log(\varepsilon_t^*)),$$

$$\mathcal{U}_{H,t} = \frac{\log(\varepsilon_t^*) - E\log(\varepsilon_t^*)}{\sqrt{Var(\log(\varepsilon_t^*))}},$$

$$\delta^2 = Var(\log(\varepsilon_t^*)),$$

于是有

$$\log(H_t) = \log(\bar{\sigma}_t(\theta)) + \delta \mathcal{U}_{H,t}, \tag{8}$$

其中,$\mathcal{U}_{H,t}$是均值为 0,方差为 1 的独立同分布随机变量。模型(8)的未知参数为 $\eta^* = (\theta', \delta)' = (\omega, \alpha, \beta, \delta)'$,参数真值记为 η_0^*。η^* 的 log-Gaussian QMLE 定义为

$$\eta^* = \underset{\eta^*}{\operatorname{argmin}} \sum_{t=1}^{T} \left\{ \log(h_t(\eta)) + \frac{[\log(H_t) - \mu_t(\eta)]^2}{h_t(\eta)} \right\},$$

其中,$\mu_t(\eta) = \log(\bar{\sigma}_t(\theta))$,$h_t(\eta) = \delta \mathcal{U}_{H,t}$。

根据 Visser[6] 的证明,可知估计量 $\hat{\eta}^*$ 具有渐近正态性,即

$$\sqrt{T}(\hat{\eta}^* - \eta_0^*) \xrightarrow{L} N(0, V^*),\ T \to \infty, \tag{9}$$

其中,

$$V^* = 4Var(\log(\varepsilon_t^*)) \boldsymbol{G}(\boldsymbol{\theta}),$$

这里,$\boldsymbol{G}(\boldsymbol{\theta})$是关于 θ 和$\mathcal{U}_{H,t}$的矩阵,与 $Var(\log(\varepsilon_t^*))$ 无关,具体证明见文献[6]。

2.2　正态分布拟极大似然估计

GARCH 模型的常用估计方法是基于正态分布的拟极大似然估计(Gaussian QMLE)。为使用 QMLE 来估计 $\theta = (\omega, \alpha, \beta)'$,需要对残差项 $z_{H,t}$进行标准化,并对模型(5)~模型(6)稍作调整。

令 $Ez_{H,t}^2 = \mu_H$, $\varepsilon_t^* = z_{H,t} \sqrt{\mu_H}$, $\sigma_t^* = \sigma_t \sqrt{\mu_H}$,根据式(5)可得 $H_t = \sigma_t z_{H,t} = \sigma_t^* \varepsilon_t^*$。于是,模型(5)~模型(6)可以分别表示为

$$H_t = \sigma_t^* \varepsilon_t^*, \tag{10}$$

$$\sigma_t^{*2} = \omega^* + \alpha^* y_{t-1}^2 + \beta^* \sigma_{t-1}^{*2}, \tag{11}$$

其中,ε_t^* 为独立同分布随机变量序列,且 $E\varepsilon_t^{*2} = 1$;参数满足:$\omega^* = \omega \mu_H$, $\alpha^* = \alpha \mu_H$, $\beta^* = \beta$。

依据模型(10)~模型(11),θ^* 的 QMLE 定义为

$$\hat{\theta}^* = \underset{\theta^*}{\operatorname{argmin}} \sum_{t=1}^{T} \left\{ \log(\sigma_t^{*2}(\theta^*)) + \frac{H_t^2}{\sigma_t^{*2}(\theta^*)} \right\}。 \tag{12}$$

假设 θ^* 的真值为 θ_0^*,根据 QMLE 的渐近理论(Straumann 等[20]),$\hat{\theta}^*$ 的渐近分布:

$$\sqrt{T}(\hat{\theta}^* - \theta_0^*) \xrightarrow{L} N(0, \Sigma^*),\ T \to \infty,$$

其中,

$$\Sigma^* = Var(\varepsilon_t^{*2}) \left(E \left(\frac{1}{\sigma_t^{*4}(\theta_0^*)} \frac{\partial \sigma_t^{*2}(\theta_0^*)}{\partial \theta_i^*} \frac{\partial \sigma_t^{*2}(\theta_0^*)}{\partial \theta_j^*} \right) \right)^{-1}。 \tag{13}$$

利用 QMLE 方法估计得到 $\hat{\theta}^*$,只需估计出

μ_H，即可得到基于高频数据 θ 的估计如式(14)：

$$\hat{\theta}=(\hat{\omega},\hat{\alpha},\hat{\beta})'=\left(\frac{\hat{\omega}^*}{\hat{\mu}_H},\frac{\hat{\alpha}^*}{\hat{\mu}_H},\hat{\beta}^*\right)', \quad (14)$$

其中，估计 μ_H 的方法即依据 $\sigma_t^*=\sigma_t\sqrt{\mu_H}$，可知 $\mu_H=\sigma_t^{*2}/\sigma_t^2$。利用高频数据构造的 $\{H_t\}$ 序列值估计出 $\hat{\theta}^*$ 后，由 $\hat{\theta}^*$ 和模型(11)可拟合得到 $\{\hat{\sigma}_t^{*2}\}$ 序列。然后，令 $H_t=|y_t|$，此时模型(10)～模型(11)即为模型(1)～模型(2)，那么由式(12)得到的 $\hat{\theta}^*$ 即为模型(1)～模型(2)的 $\theta=(\omega,\alpha,\beta)'$ 估计，记为 $\hat{\theta}=(\tilde{\omega},\tilde{\alpha},\tilde{\beta})'$，则

$$\hat{\mu}_H=\frac{1}{T}\sum_{t=1}^{T}\frac{\hat{\sigma}_t^{*2}}{\tilde{\sigma}_t^2}。$$

2.3 拟极大指数似然估计

在实际数据分析中，模型残差的分布是未知的，为了弱化矩条件，残差项常被假定为服从标准双指数分布(Laplace 分布)，因而基于该分布的拟极大指数似然估计(QMELE)也是常用的估计方法。

若采用 QMELE 来估计 θ，则只需要残差项的一阶绝对矩存在。假设 $Ez_{H,t}=\nu_H$，并令 $e_t^*=z_{H,t}/\nu_H$，则 $Ee_t^*=1$。相应地，令 $\sigma_t^*=\sigma_t\nu_H$，则模型(5)～模型(6)可改写为

$$H_t=\sigma_t^* e_t^*, \quad (15)$$

$$\sigma_t^{*2}=\omega^*+\alpha^* y_{t-1}^2+\beta^*\sigma_{t-1}^{*2}, \quad (16)$$

其中，参数 $\omega^*=\omega\nu_H^2,\alpha^*=\alpha\nu_H^2,\beta^*=\beta$。

依据模型(15)～模型(16)，$\theta^*=(\omega^*,\alpha^*,\beta^*)'$的 QMELE 定义为

$$\hat{\theta}^*=\underset{\theta^*}{\operatorname{argmin}}\sum_{t=1}^{T}\left\{\log(\sigma_t^*(\theta^*))+\frac{|H_t|}{\sigma_t^*(\theta^*)}\right\},$$

根据 QMELE 的渐近理论(Andersen 等[21])，容易得到 $\hat{\theta}^*$ 的渐近分布为

$$\sqrt{T}(\hat{\theta}^*-\theta_0^*)\xrightarrow{L}N(0,\Omega^*),T\to\infty,$$

其中，

$$\Omega^*=4(Ee_t^{*2}-1)\left(E\left(\frac{1}{\sigma_t^{*4}(\theta_0^*)}\frac{\partial\sigma_t^{*2}(\theta_0^*)}{\partial\theta^*}\frac{\partial\sigma_t^{*2}(\theta_0^*)}{\partial\theta^{*\prime}}\right)\right)^{-1}。$$

类似于 QMLE 方法，利用估计量 $\hat{\theta}^*$ 和 $\hat{v}_H^2$，便可以得到 QMELE 方法下基于高频数据的 θ 估计为

$$\hat{\theta}=(\hat{\omega},\hat{\alpha},\hat{\beta})'=\left(\frac{\hat{\omega}^*}{\hat{v}_H^2},\frac{\hat{\alpha}^*}{\hat{v}_H^2},\hat{\beta}^*\right)',$$

其中，$\hat{v}_H^2=T^{-1}\sum_{t=1}^{T}\hat{\sigma}_t^{*2}/\tilde{\sigma}_t^2$，具体估计过程类似于第2.2节中的 μ_H 估计，故不再赘述。

3 波动率代表的选择方法

3.1 波动率代表选择

从第2节可知，参数估计是通过极小化关于波动率代表的似然函数得到的，所以使用不同的波动率代表得到估计量的有效性是不同的，同时基于不同的估计方法，又会存在差异。因而选择合适的波动率代表是获取准确估计的重要前提。

(1)对于对数正态拟极大似然估计，其估计量的渐近分布由式(9)给出，从中可看出，$Var(\log(\varepsilon_t^*))$越小，$\hat{\theta}$ 的渐近方差也越小。Visser[6]在文献中已给出了对应此估计方法的一个波动率代表选择标准：

$$\lambda=Var(\log(H_t)|F_{n-1}), \quad (17)$$

如果 λ 越小，参数估计的渐近方差就越小，即对应的波动率代表越好。因而，对于 log-Gaussian QMLE，寻找最优的波动率代表即为寻找最小的 λ 值。

(2)对于正态分布拟极大似然估计，Visser[6]并未给出适用于其估计量的波动率代表选择方法。为此，在估计量渐近分布的基础上进行分析，以给出针对 QMLE 的波动率代表选择标准。

由式(13)易知，$Var(\varepsilon_t^{*2})$ 越小，意味着参数估计的渐近方差越小，估计就越有效。已知 $\varepsilon_t^*=z_{H,t}/\sqrt{\mu_H}=z_{H,t}/\sqrt{Ez_{H,t}^2}$，代入 $Var(\varepsilon_t^{*2})$ 中可得

$$Var(\varepsilon_t^{*2})=Var\left(\frac{z_{H,t}^2}{Ez_{H,t}^2}\right)=\frac{Ez_{H,t}^4-(Ez_{H,t}^2)^2}{(Ez_{H,t}^2)^2}=\frac{Ez_{H,t}^4}{(Ez_{H,t}^2)^2}-1\geqslant 0,$$

显然，$Var(\varepsilon_t^{*2})$ 越小，对应 $Ez_{H,t}^4/(Ez_{H,t}^2)^2$ 也越小。但 $z_{H,t}$ 是不可观测的，不能直接计算得到 $Var(\varepsilon_t^{*2})$，需要作进一步推导。考虑到 $H_t=\sigma_t H(Z_t(u))=\sigma_t z_{H,t}$，且 σ_t 与 $z_{H,t}$ 相互独立，可以推得

$$EH_t^4=E(\sigma_t^4 z_{H,t}^4)=E(\sigma_t^4)E(z_{H,t}^4),$$

$$(EH_t^2)^2=[E(\sigma_t^2 z_{H,t}^2)]^2=[E(\sigma_t^2)]^2[E(z_{H,t}^2)]^2。$$

相应地，将上述两式进行相除，即得

$$\frac{EH_t^4}{(EH_t^2)^2}=\frac{E(\sigma_t^4)E(z_{H,t}^4)}{[E(\sigma_t^2)]^2[E(z_{H,t}^2)]^2}=\frac{E(\sigma_t^4)}{[E(\sigma_t^2)]^2}\cdot\frac{E(z_{H,t}^4)}{[E(z_{H,t}^2)]^2}=\mathrm{c}\cdot\frac{E(z_{H,t}^4)}{[E(z_{H,t}^2)]^2},$$

其中,$c = E(\sigma_t^4)/[E(\sigma_t^2)]^2$ 为一常数。

定义

$$MH_{qmle} = \frac{EH_t^4}{(EH_t^2)^2}, \tag{18}$$

从而有 MH_{qmle} 越小 $\leftrightarrow Ez_{H,t}^4/(Ez_{H,t}^2)^2$ 越小 $\leftrightarrow Var(\varepsilon_t^{*2})$越小,这表明,$MH_{qmle}$越小,QMLE 的渐近方差就越小,即估计就越准确。因此,QMLE 下寻找最优的波动率代表即转化为寻找最小的 MH_{qmle}。

(3)对于拟极大指数似然估计,从渐近分布的角度出发同样可以得到选择最优波动率代表的标准。

于是,依据同样的推导思想,可以推得 QMELE 下波动率代表的选择标准为

$$MH_{qmele} = \frac{EH_t^2}{(EH_t)^2}, \tag{19}$$

以及有 MH_{qmele}越小 $\leftrightarrow Ez_{H,t}^2/(Ez_{H,t})^2$越小 $\leftrightarrow Ee_t^{*2}$越小,从而,$MH_{qmele}$越小,QMELE 的渐近方差就越小。因此,QMELE 下寻找最优的波动率代表即转化为寻找最小的 MH_{qmele}。

3.2　常见波动率代表

从高频对数收益率的表达式可以看出,计算波动率代表需要通过离散性抽样数据来得到,也就是需要固定一个时间间隔来采集日内的高频数据。令 k 表示日内时间间隔(单位:min),m 表示在抽样频率 k 下一天内总的收益个数,$Y_t(u_{ik})$ 表示第 i 次抽样的收益率,$i = 1,\cdots,m$。本文考虑以下 4 种波动率代表,作为模型中主要的选择比较:

①已实现波动率(RV):

$$RV_t(k) = \sqrt{\sum_{i=1}^{m}[Y_t(u_{ik}) - Y_t(u_{(i-1)k})]^2},$$

②日内收益绝对值之和(RAV):

$$RAV_t(k) = \sum_{i=1}^{m}|Y_t(u_{ik}) - Y_t(u_{(i-1)k})|,$$

③已实现极差波动率($RVHL$):

$$RVHL_t(k) = \sqrt{\sum_{i=1}^{m}[\max_{\Delta_i}(Y_t(u)) - \min_{\Delta_i}(Y_t(u))]^2},$$

$$k(i-1) < \Delta_i < ki,$$

④日内极差之和($RAVHL$):

$$RAVHL_t(k) = \sum_{i=1}^{m}\{\max_{\Delta_i}(Y_t(u)) - \min_{\Delta_i}(Y_t(u))\}, \quad k(i-1) < \Delta_i < ki,$$

其中,$Y_t(u_0)$的值用 $Y_t(0) = 0$ 代替;$\max_{\Delta_i}(Y_t(u))$和 $\min_{\Delta_i}(Y_t(u))$分别为第 i 个时间段 Δ_i 中的收益率最大值和最小值。

以上 RV、RAV、$RVHL$ 和 $RAVHL$ 都是波动率代表 H_t 的具体例子,容易看出 H_t 的具体值会依赖于离散化的数据量个数,而将一天内的交易时间等分为多少段最合适即为最优抽样频率的问题。对于同一波动率代表,最优频率对应的 H_t 才是最优的。因而第 3.1 节给出的波动率代表选择标准可以提供一个选取最优抽样频率的方法,即波动率代表选择标准达到最小的时间频率为该波动率代表的最佳频率。

4　实　证

本文以沪深 300 指数为研究对象,基于不同估计方法对波动率代表及其频率的最优选择问题进行了实证研究。选取 2017 年 9 月 1 日 - 2019 年 7 月 12 日共 466 个交易日的 1 min 收盘价数据,每天有 241 个价格观测值。令沪深 300 指数的价格序列为$\{P_t(u), t \in [0,466], u \in [0,1]\}$,根据高频价格数据计算出 1 min 高频收益率$\{Y_t(u_i)\}_{i=1}^{240}$。

考虑使用高频收益数据$\{Y_t(u_i)\}_{i=1}^{240}$来构造波动率代表。采用 4 种类型的波动率代表,分别为 $H_1 = RV_t(k)$,$H_2 = RAV_t(k)$,$H_3 = RVHL_t(k)$,$H_4 = RAVHL_t(k)$。利用高频收益数据,根据不同的时间间隔 k 取样,可以构造不同频率的 H_1、H_2、H_3 和 H_4。由于所得样本的最高频率为 1 min,无法得到 1 min 内的极值或极差,因此,$RVHL$ 和 $RAVHL$ 的最高频率为 2 min。考虑尽可能多的抽样频率,即在 2 ~ 60 min 中选取所有能满足等间隔抽样的整数 k 值。利用第 3.1 节中所提出的波动率代表选择标准,根据式(17) ~ 式(19)分别计算不同波动率代表在不同频率下的 λ、MH_{qmle}和 MH_{qmele}估计值,λ、MH_{qmle}和 MH_{qmele}估计值最小所对应的抽样频率代表了使用该波动率代表和对应估计方法的最优频率,表 1 汇总了不同波动率代表在不同估计方法下选择得到的最优抽样频率结果。同时,为了更直观地展示波动率代表与抽样频率的关系,基于不同估计准则下、不同频率上的曲线趋势图,如图 1 所示。图 1 中的左边 3 幅子图代表 RV 和 RAV 在不同估计方法下的频率关系图,右边 3 幅子图代表 $RVHL$ 和 $RAVHL$ 下不同频率的关系图。

表 1 不同波动率代表在不同估计方法下的最优抽样频率

Table 1 The optimal sampling frequencies of different volatility represents under different estimation methods

估计方法	最优抽样频率/min			
	RV	*RAV*	*RVHL*	*RAVHL*
log-Gaussian QMLE	8(0.181 04)	8(0.138 67)	10(0.123 90)	10(0.124 17)
Gaussian QMLE	15(2.982 62)	2(1.751 89)	2(1.533 28)	2(1.593 36)
QMELE	8(1.260 59)	2(1.150 64)	8(1.127 27)	8(1.129 64)

注:括号中的值是最优抽样频率在对应估计方法下的准则值。

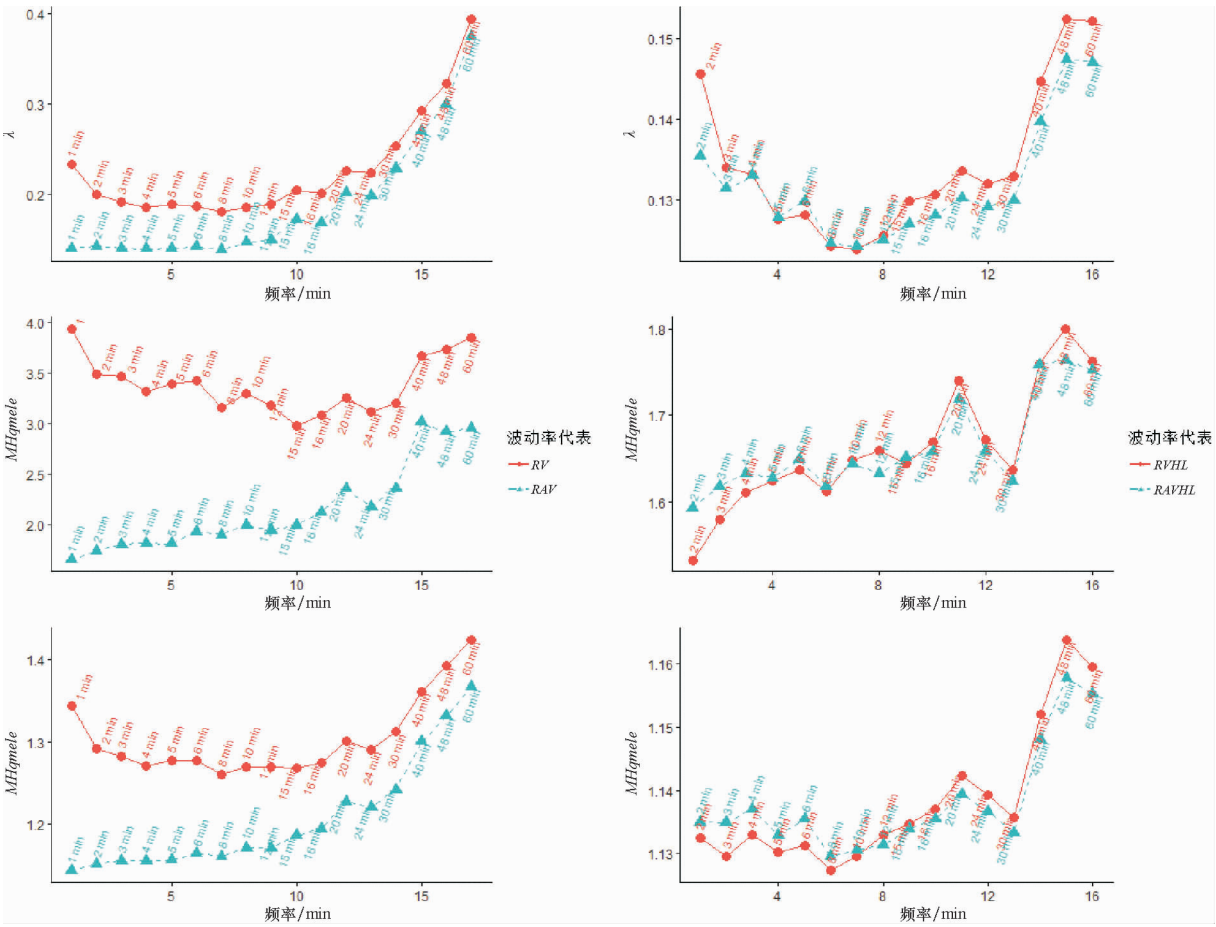

图 1 波动率代表与抽样频率的关系图

Fig. 1 Relationship between volatility representation and sampling frequency

由表 1 和图 1 可看出,在不同波动率代表、不同频率、不同估计方法下,得到的估计量有效性均明显不同。通过比较分析发现:

(1)在同一频率 3 种估计方法准则下,比较 *RV*、*RAV*、*RVHL* 和 *RAVHL* 后发现,*RV* 的有效性最差,*RAVHL* 和 *RVHL* 是较优的波动率代表函数。可以看到,在频率较高时(2 ~ 8 min)*RAV* 和 *RVHL* 是最优的波动率代表函数,在频率较低时(10 min 以上)*RVHL* 的有效性表现更优。从指标的总体性看,两者相差不大,因此,*RVHL* 和 *RAV* 都可作为最优的波动率函数。

(2)对同一波动率代表分别比较不同频率的表现可以发现,*RV* 在不同频率下的有效性波动(数值大小变动)最大,随着抽样频率的增加,每种方法对应的数值大多是先递减后递增;相比之下,*RVHL* 和 *RAVHL* 在不同频率下的有效性表现较为稳定,数值变动较小。

(3)比较最优抽样频率下的不同波动率代表发现,采用 log-Gaussian QMLE 时的最优波动率代表为 10 min 的 *RVHL*(λ 值最小),采用 Gaussian

QMLE 时,则是 2 min 的 *RVHL* 最优(MH_{qmle} 值最小),而采用 QMELE 时是 8 min 的 *RVHL* 最优(MH_{qmele} 值最小),意味着在最优抽样频率下,无论使用什么估计方法,最优的波动率代表为 *RAV*。

(4)每个波动率代表在相应频率上达到了最低点。不同波动率代表的曲线趋势存在显著差异。其中,*RV* 的 3 条曲线均呈"低谷"状态;而 *RAV* 的 3 条曲线整体上均呈现递增趋势;*RVHL* 和 *RAVHL* 随时间频率的变化趋势相同,或呈曲折递增,或大致地递减后再递增。

基于表 1 的结果,选择以 2 min 为间隔的 *RVHL* 进一步使用高斯 QMLE 估计出 GARCH(1,1)模型的参数。该波动率代表具有最小的 MH_{qmle} 值,那么根据式(12)和式(14)得到的估计量是最有效的 QMLE,拟合沪深 300 指数收益率的 GARCH(1,1)模型为

$$y_t = \sigma_t \varepsilon_t, \tag{20}$$

$$\sigma_t^2 = 0.072\,9 + 0.108\,9 y_{t-1}^2 + 0.859\,4 \sigma_{t-1}^2, \tag{21}$$

于是,基于模型(20)~模型(21)可以获得更为准确的波动率估计。另外,若选用 QMELE 方法或对数正态分布 QMLE 方法,则分别需要以 8 min 和 10 min 的频率来构造 *RVHL*,这样得到的 QMELE 和 log-Gaussian QMLE 的估计有效性是其中最好的,便于更准确地建立模型和估计波动率。

为了检验结果的稳健性,本文再将全样本分成了两个子样本(2017 年 9 月 1 日 - 2018 年 6 月 30 日和 2018 年 7 月 1 日 - 2019 年 7 月 12 日),进行同样的波动率代表和频率选择分析。汇总结果见表 2。

表 2　子样本在最优波动率代表下的最优抽样频率

Table 2　Optimal sampling frequency of subsamples represented by optimal volatility

估计方法	2017 年 9 月 1 日至 2018 年 6 月 30 日				2018 年 7 月 1 日至 2019 年 7 月 12 日			
	最优抽样频率/min				最优抽样频率/min			
	RV	*RAV*	*RVHL*	*RAVHL*	*RV*	*RAV*	*RVHL*	*RAVHL*
log-Gaussian QMLE	10(0.171 91)	1(0.131 42)	10(0.127 83)	8(0.124 84)	4(0.132 11)	1(0.076 32)	3(0.069 11)	2(0.075 42)
Gaussian QMLE	24(3.122 72)	1(2.179 72)	2(1.891 71)	30(1.869 90)	15(2.671 61)	1(1.386 13)	2(1.293 62)	2(1.349 52)
QMELE	12(1.260 31)	1(1.172 71)	10(1.156 42)	10(1.150 11)	8(1.214 10)	1(1.084 41)	2(1.069 83)	2(1.080 11)

由表 2 可以发现:

(1)同频率两个子样本下,*RV* 波动率代表函数的有效性是最差的,*RVHL* 和 *RAVHL* 仍然是两个差别不大的最优波动率代表函数。

(2)针对两个不同的样本,基于最优准则下选择的最优波动率代表会有所不同,同时在不同估计方法下选择的最优频率也会有所差异,表 2 中列出了两个子样本选择的最优波动率函数,以及不同估计方法下,基于最优波动率函数选择的最优频率。在 2017 年 9 月至 2018 年 6 月的第一个子样本中,选择的是 *RAVHL* 为最优波动率函数,3 个估计方法下所对应的最优频率分别为 8 min、30 min 和 10 min;在 2018 年 7 月至 2019 年 7 月的第二个子样本中,选择的是 *RVHL* 为最优波动率函数,3 个估计方法下所对应的最优频率分别为 3 min、2 min 和 2 min。

总的来说,本文可以得到一个稳健的结论是在全样本和子样本下,*RVHL* 和 *RAVHL* 都可考虑选择为最优波动率函数,且不同频率下它们的表现模式也很相似。在选择最优频率时,除了考虑选择最优波动率函数,还要考虑不同的估计方法,同时也会依赖不同样本下的情况,但就整体而言,最优频率在 2 min、8 min 和 10 min 中选择,也启示了研究者和实践应用学者们在实际中对抽样频率进行谨慎选择。

5　结　论

通过波动率代表,可以将日内高频数据应用于改进 GARCH 模型的参数估计,不同的波动率代表提高估计精度的效果不同。本文介绍了波动率代表模型的 3 种估计方法,针对不同的估计方法给出了波动率代表的选择标准,并将这些准则进一步应用于解决高频数据抽样频率的选择问题。最后,采用沪深 300 指数的高频数据做实证研究,通过比较研究发现,不同波动率代表、不同抽样频率都会对 GARCH 参数估计量的有效性造成明显差异,而且在不同估计方法下的表现也不同。

已实现极差波动率(RVHL)和日内极差波动率(RAVHL)是有效性表现较优的波动率代表,但它们都会依赖于抽样频率,其最优频率也会依赖具体的估计方法,但主要是在2~8 min中选择,具体的选择也需考虑更多的因素,未来可以考虑一个自适应样本的方法对最优频率进行选择。

参考文献:

[1] Engle R F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation[J]. Econometrica, 1982, 50(4):987-1008.

[2] Bollerslev T. Generalized autoregressive conditional heteroskedasticity[J]. Journal of Econometrics, 1986, 31(3):307-327.

[3] McAleer M, Medeiros M C. Realized volatility: A review[J]. Econometric Reviews, 2008,27(1/2/3):10-45.

[4] Aït-Sahalia Y, Fan J Q, Xiu D C. High-frequency covariance estimates with noisy and asynchronous financial data[J]. Journal of the American Statistical Association, 2010,105(492):1504-1517.

[5] Jing B Y, Liu Z, Kong X B. Estimating volatility functionals with multiple transactions[J]. Econometric Theory, 2017,33(2):331-365.

[6] Visser M P. GARCH parameter estimation using high-frequency data[J]. Journal of Financial Econometrics, 2011, 9(1): 162-197.

[7] 黄金山, 陈敏. 基于高频数据的GARCH模型的伪极大指数似然估计[J]. 应用数学学报, 2014, 37(6):1005-1017.

[8] Huang J S, Wu W Q, Chen Z, et al. Robust M-estimate of GJR model with high frequency data[J]. Acta Mathematicae Applicatae Sinica, English Series, 2015, 31(3): 591-606.

[9] 樊鹏英, 兰勇, 陈敏. 高频数据下基于PGARCH模型的VaR估计方法及应用[J]. 系统工程理论与实践, 2017, 37(8):2052-2059.

[10] 吴思鑫, 冯牧, 张虎, 等. 基于高频数据的非平稳GARCH(1,1)模型的拟极大指数似然估计[J]. 中国科学(数学), 2018, 48(3): 443-456.

[11] Wang M, Chen Z, Wang C D, et al. Composite quantile regression for GARCH models using high-frequency data[J]. Econometrics and Statistics, 2018, 7: 115-133.

[12] Deng C L, Zhang X F, Li Y A, et al. On the test of the volatility proxy model[J]. Communications in Statistics-Simulation and Computation, 2022,51(12):7390-7403.

[13] Liang X, Zhang X F, Li Y A, et al. Daily nonparametric ARCH (1) model estimation using intraday high frequency data[J]. AIMS Mathematics, 2021, 6(4): 3455-3464.

[14] 徐正国,张世英. 高频时间序列的改进"已实现"波动特性与建模[J]. 系统工程学报, 2005,20(4): 344-350.

[15] 郭名媛,张世英. 赋权已实现波动及其长记忆性,最优频率选择[J]. 系统工程学报, 2006,21(6): 568-573.

[16] 唐勇, 张世英. 已实现波动和已实现极差波动的比较研究[J]. 系统工程学报, 2007,22(4): 437-442.

[17] 李胜歌, 张世英. 金融高频数据的最优抽样频率研究[J]. 管理学报, 2008,5(6): 801-806,840.

[18] 闵素芹, 柳会珍. "已实现"波动率中最优抽样频率的选择[J]. 统计与决策, 2009(13): 13-15.

[19] 杨建辉, 鲁旭芬. 不同抽样频率下创业板指数波动的测度[J]. 统计与决策, 2012(23): 13-16.

[20] Straumann D, Mikosch T. Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach[J]. The Annals of Statistics, 2006, 34(5): 2449-2495.

[21] Andersen T G, Bollerslev T, Diebold F X, et al. Modeling and forecasting realized volatility[J]. Econometrica, 2003,71(2): 579-625.

【责任编辑: 卓祯雨】