文章编号:1671-4229(2024)05-0054-15

基于高频数据的 GARCH 模型拟极大指数似然估计的一种 portmanteau Q 检验

陈燕珊[1], 张兴发[1*], 田 玥[2], 陈嘉卓[1]
(1. 广州大学 经济与统计学院, 广东 广州 510006; 2. 广东培正学院 经济学院, 广东 广州 510830)

摘 要: 已有研究表明,基于高频数据的 GARCH 模型的拟极大指数似然估计可以提升估计精度,但鲜有研究就该估计量的性质推导其对应的检验统计量。文章基于高频数据的 GARCH 拟极大指数似然估计性质,提出一种 portmanteau Q 检验统计量,通过模拟验证了该检验统计量的理论正确性,并选取沪深300、中证500和上证50等3个指数进行了具体应用。结果显示,在模型充分时,文章提出的检验统计量的分布更近似理论推导的分布,优于基于低频数据的检验统计量结果,且由于包含高频信息,该统计量能更好地捕捉高频残差自相关性;而当低频残差自相关性时,即使相关性较弱,该统计量也能识别模型是否充分,对 GARCH 模型的阶数识别有一定效果。实证研究也表明,该检验统计量能对高频信息有效利用,具有一定的实用性。

关键词: 高频数据; GARCH 模型; 拟极大指数似然估计; portmanteau Q 检验
中图分类号: O212.1 **文献标志码**: A

A portmanteau test for quasi maximum exponential likehood estimation of GARCH model based on high frequency data

CHEN Yan-shan[1], *ZHANG Xing-fa*[1*], *TIAN Yue*[2], *CHEN Jia-zhuo*[1]
(1. School of Economics and Statistics, Guanghou University, Guangzhou 510006, China;
2. School of Economics, Guang Dong Peizheng College, Guangzhou 510830, China)

Abstract: Previous studies have shown that the quasi maximum exponential likelihood estimation based on high frequency data can improve the estimation accuracy of GARCH model, but few studies have derived the corresponding test statistic for this estimator. In this paper, a portmanteau Q test statistic is proposed based on the asymptotic property of quasi maximum exponential likelihood estimation of GARCH model based on high-frequency data. The theoretical correctness of the test statistic is validated through simulation in this paper, and specific applications are provided by using the data of the CSI 300, CSI 500, and SSE 50 indices. The results show that when the model is adequate, the distribution of the test statistic proposed in this paper more closely follows the theoretically derived distribution, which is better than the results of the test statistic based on low-frequency data. Moreover, the statistic is able to capture high-frequency residual autocorrelation due to the inclusion of high-frequen-

收稿日期: 2024-01-03; **修回日期**: 2024-04-22
基金项目: 广东省自然科学基金面上资助项目(2022A1515010046);广州市校(院)联合资助项目(SL2022A03J00654)
作者简介: 陈燕珊(1997—),女,硕士研究生. E-mail:2112164127@e.gzhu.edu.cn
* 通信作者. xingfazhang@gzhu.edu.cn
引文格式: 陈燕珊, 张兴发, 田玥, 等. 基于高频数据的 GARCH 模型拟极大指数似然估计的一种 portmanteau Q 检验[J]. 广州大学学报(自然科学版),2024,23(5):54-68.

cy information. While for low-frequency residual autocorrelation, the statistic can also identify model non-sufficiency when the correlation is stronger, which is useful for order identification in GARCH model. Empirical research also indicates that the test statistic can identify the effective utilization of high-frequency information by the models based on high-frequency data, demonstrating a certain degree of practicality.

Key words: high frequency data; GARCH model; quasi-maximum likelihood estimation; portmanteau Q test

波动率是衡量股票市场风险的一个重要工具,波动率越大,代表市场风险越大[1]。针对波动率特征,研究者们已经提出了完善的波动率建模过程,其中主要包括建立模型、模型估计和模型检验 3 个部分。

针对金融市场波动率,最经典的时间序列模型有自回归条件异方差(ARCH)模型和广义自回归条件异方差(GARCH)模型。ARCH 是由 Engle[2]提出,它可以解释数据的异方差特征,但其只适用于异方差函数短期自相关过程。为了更能反映实际数据中的长期记忆性质,Bollerslev[3]提出了 GARCH 模型。GARCH 模型可以刻画波动率的聚集性和时变性,是一种常用的波动率模型。令 y_t 为某资产在第 t 天的对数收益率,则纯 GARCH(1,1) 模型可以表示如下:

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

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

其中,参数 $\omega > 0, \alpha \geqslant 0, \beta \geqslant 0$;$\varepsilon_t$ 是独立同分布,均值为 0 的随机变量,且对 $\forall s \leqslant t$, ε_t 与 y_s 独立;σ_t^2 表示资产在第 t 天的波动率,该波动率大于 0 且只依赖于 t 天之前的信息。GARCH 模型提出后,很多研究者在 GARCH 模型基础上进行改进,以此来提升模型对数据拟合的估计精度。例如 Nelson[4]提出指数 GARCH (EGARCH)模型,该模型中引入了加权扰动函数,可以对正、负扰动项进行非对称处理,从而解释数据的杠杆效应。更多 GARCH 类模型可见单整的 GARCH(IGARCH)模型[5]和门限 GARCH(TGARCH)模型[6]等。但这些模型都是基于低频数据(以天、周和月为时间间隔)的 GARCH 模型研究。

随着科学技术的发展,金融市场的日内交易数据变得可获得,这也意味着,GARCH 类模型的改进可以基于高频数据进行。基于此,Visser[7]提出基于高频数据的 GARCH 波动率代表模型,一方面提出了基于日内高频数据的尺度模型,另一方面,在尺度模型上使用波动率代表来解决模型的估计问题。结果显示,在高频数据的 GARCH 波动率代表模型下,使用极大似然估计(QMLE)可以极大地提升模型的估计精度。

在波动率模型估计方面,比较常见的估计方法有拟极大似然估计和拟极大指数似然估计,特别是对于 GARCH 类模型。传统极大似然估计要求序列服从正态分布,然而金融时间序列通常是厚尾分布。1992 年,Bollerslev 等[8]证明了尽管序列是厚尾分布,但高斯似然函数依旧适用。这种利用高斯似然函数的估计方法称为拟极大似然估计(QMLE)。随后,Lee 等[9]完善了 QMLE 的渐近性质。然而,当误差项的四阶矩不存在时, Hall 等[10]证明了拟极大似然估计可能不服从渐近正态分布,且收敛速度会低于$\sqrt{n}$。针对该情况,Peng 等[11]提出了 3 种最小绝对值偏差估计(LAD),并证明了不管四阶矩是否存在,这些估计的收敛速度都可以维持在$\sqrt{n}$。Peng 等还简单介绍了拟极大指数似然估计(QMELE),但未给出渐近性质推导。之后,Li 等[12]将 QMELE 应用于 ARFIMA-GARCH 模型,并给出该估计的渐近性质。借鉴 Visser[7]将 QMLE 应用于高频 GARCH 波动率模型的想法,黄金山等[13]使用 QMELE 求解基于高频数据的 GARCH 模型参数,削弱了四阶矩存在的条件。以上基于高频数据的 GARCH 模型的求解都需要假定公式(2)中的常数项 ω 已知。假设 ω 已知限制了模型的实际应用,因而李莉丽等[14]提出两步估计法,将黄金山等的结果推广到了一般的基于高频数据的 GARCH(1,1)波动率模型上。李

莉丽等通过结合日频 GARCH 模型的 QMELE 估计得到冗余参数的估计,从而得到所有参数的估计,并充分利用高频信息显著提高了 GARCH(1,1)模型估计精度。

portmanteau Q 检验是一种有效的检验工具,最早可以追溯到 Box 等[15]的研究,他们利用平方自相关系数构造检验统计量来检验序列是否白噪声,是否具有研究的意义。在此之后,Ljung 等[16]对 portmanteau Q 检验统计量进行修正,提出了 Ljung-Box (LB)检验统计量,并通过蒙特卡罗试验证明了修正后的统计量更接近卡方分布。拟合原序列数据后,会有一部分残差信息,如果残差为白噪声,则说明模型提取信息充分,因此,检验模型残差是否为白噪声,可以用来检验模型的充分性,如 McLeod 等[17]就将该检验统计量用于检验模型充分上。但对于 ARCH、GARCH 模型,Li 等[18]表明 Box 等提出的检验统计量并不适用,因此提出了带方差项的 portmanteau Q 检验统计量。这为后期 ARCH 及 GARCH 模型的充分性检验奠定了良好的基础。Ling 等[19]将带方差项的 portmanteau Q 检验统计量应用于多元 ARCH 模型,并运用了不带平方的残差自相关系数的检验统计量。后来,由于 LAD 和 QMELE 估计方法的出现,带绝对值的残差自相关系数构造的检验统计量也随之发展,两者都满足残差四阶矩不存在的更弱条件,实现了估计和检验的统一。Li 等[20]提出了绝对值相关系数构造的带方差的检验统计量,并证明了该统计量更稳健。至今仍有不少学者(Chen 等[21],Jiang 等[22]及 Li 等[23])将 Li 等[18]检验统计量的构造思想沿用到更多的模型检验上。

综上所述,带绝对值的 portmanteau Q 检验是检验 GARCH 模型的一种有效工具,而基于高频数据的 GARCH 模型的 QMELE 是提高估计精度的一种有效估计。但是将两者结合的检验是研究的空缺。因此,本文将基于高频数据的 QMELE 估计性质,提出一种带绝对值的 portmanteau Q 检验,从而实现基于高频数据的估计与检验的统一。

本文总共包含 4 节,安排如下:第 1 节,介绍基于高频数据的 GARCH 模型和拟极大指数似然估计性质;第 2 节,推导基于高频信息的 portmanteau Q 检验统计量及其分布;第 3 节,展示 portmanteau Q 检验统计量的模拟结果;第 4 节,使用 3 个股票指数数据对提出的检验统计量进行实证分析,第 5 节,总结全文。

1 基于高频数据的拟极大指数似然估计

为引入高频数据,令第 t 天的日内对数收益率为 $Y_t(u)$,$0\leqslant u\leqslant 1$。将高频数据引入 GARCH(1,1)模型:

$$Y_t(u)=\sigma_t Z_t(u), \tag{3}$$

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

其中,σ_t 表示第 t 天的波动率,$Z_t(\cdot)$为独立同分布的随机过程,即对任意 $k\neq l$,$Z_k(\cdot)$与 $Z_l(\cdot)$相互独立,且具有相同的分布。记 $\boldsymbol{\theta}=(\omega,\alpha,\beta)$是模型(3)~(4)的待估参数。参数 u 表示将日内时间正则化到 0 到 1 的闭区间之间,$u=1$ 表示当天收盘时刻,则

$$Y_t(1)=y_t, Z_t(1)=\varepsilon_t, E|Z_t(1)|=1。 \tag{5}$$

等式(5)将日内高频 GARCH(1,1)模型与日间 GARCH(1,1)联系了起来。

为了估计参数,下面引入波动率代表函数 $H(\cdot)$。波动率代表是日内数据的一种统计量,对任意常数 $\rho(\rho>0)$和日内过程 $Y_t(u)$,其满足

$$H(\rho Y_t(u))=\rho H(Y_t(u))>0。$$

当时间 t 固定时,σ_t 可视为常数,则

$$H(Y_t(u))=H(\sigma_t Z_t(u))=\sigma_t H(Z_t(u))。$$

为得到更好的估计性质,在使用 QMELE 前,需要对模型进行一些变形。QMELE 的相合性要求模型误差项的绝对值期望为 1,当引入高频数据和波动率代表后,模型误差项 $H(Z_t(u))$不一定满足 $E|H(Z_t(u))|=1$ 的条件。因此,为了保证估计相合性,可以剔除误差项冗余部分,记为 μ_H,使得误差项剩余部分满足条件,余下部分记为 ε_t^*。为使得 $E|\varepsilon_t^*|=1$,求得冗余参数 $\mu_H=E(H(Z_t(u)))$,则 $\varepsilon_t^*=H(Z_t(u))/\mu_H$。

为了方便,令 $H_t\triangleq H(Y_t(u))$,则高频 GARCH(1,1)波动率模型可表示为

$$H_t=\sigma_t\mu_H\varepsilon_t^*, \tag{6}$$

$$\sigma_t^2=\omega+\alpha y_{t-1}^2+\beta\sigma_{t-1}^2。 \tag{7}$$

令 $\sigma_t^*=\sigma_t\mu_H$,则模型变为

$$H_t = \sigma_t^* \varepsilon_t^*, \tag{8}$$

$$\sigma_t^{*2} = \omega^* + \alpha^* y_{t-1}^2 + \beta^* \sigma_{t-1}^{*2}。 \tag{9}$$

为方便记录,记 $\boldsymbol{\theta} = (\omega, \alpha, \beta)$, $\boldsymbol{\theta}^* = (\omega^*, \alpha^*, \beta^*)$,其中,

$$\omega^* = \omega\mu_H^2, \alpha^* = \alpha\mu_H^2, \beta^* = \beta。 \tag{10}$$

在模型(8)~(9)下,QMELE 可被定义为

$$\hat{\boldsymbol{\theta}}^* = arg \max_{\boldsymbol{\theta}} - \left\{ \sum_{t=1}^{n} \log(\sigma_t \mu_H) + \frac{| H_t |}{\sigma_t \mu_H} \right\} =$$

$$arg \min_{\boldsymbol{\theta}^*} \sum_{t=1}^{n} \log(\sigma_t^*) + \frac{| H_t |}{\sigma_t^*} =$$

$$arg \min_{\boldsymbol{\theta}^*} \sum_{t=1}^{n} l_t^* = arg \min_{\boldsymbol{\theta}^*} L_n^*。 \tag{11}$$

由等式(8)可知,将 σ_t^* 看作一个整体,则高频波动率模型的似然函数结构与低频模型一致,因而可证在一定的正则条件下,$\boldsymbol{\theta}^*$ 存在渐近性质[14]如下:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0^*) \xrightarrow{d} N(0, \boldsymbol{\Sigma}^*), n \to \infty, \tag{12}$$

其中,

$$\boldsymbol{\Sigma}^* = 4(E\varepsilon_t^{*2} - 1)\boldsymbol{G}^{*-1},$$

$$\boldsymbol{G}^* = E\left(\frac{1}{\sigma_t^{*4}} \frac{\partial \sigma_t^{*2}}{\partial \boldsymbol{\theta}^*} \frac{\partial \sigma_t^{*2}}{\partial \boldsymbol{\theta}^{*\prime}}\right)。 \tag{13}$$

得到参数 $\boldsymbol{\theta}^*$ 的估计后,要进一步得到去“*”号的参数估计 $\boldsymbol{\theta}$,一种已有思路是通过对冗余参数 μ_H 估计后,再进一步得到 $\hat{\boldsymbol{\theta}}_0$。等式可以将高频波动率代表模型退化为低频 GARCH 模型,表明低频模型的参数估计与高频模型的参数估计存在联系。根据这一想法,李莉丽等[14]提出的关于 μ_H^2 的一种估计如下:

$$\widehat{\mu_H^2} = \frac{1}{n} \sum_{t=1}^{n} \frac{\widehat{\sigma_t^2}(\widehat{\boldsymbol{\theta}^*})}{\widetilde{\sigma_t^2}(\tilde{\boldsymbol{\theta}})}, \tag{14}$$

其中,$\widehat{\sigma_t^2}$是根据模型(8)~(9)和 QMELE 方法,即等式(11)得到的波动率估计,$\widetilde{\sigma_t^2}$是根据日间模型(1)~(2)使用 QMELE 方法得到的波动率估计。

根据等式(10)和(14)可以得到高频波动率代表模型的参数估计 $\hat{\boldsymbol{\theta}}$ 如下:

$$\hat{\omega} = \frac{\widehat{\omega^*}}{\widehat{\mu_H^2}}, \quad \hat{\alpha} = \frac{\widehat{\alpha^*}}{\widehat{\mu_H^2}}, \quad \hat{\beta} = \widehat{\beta^*}。 \tag{15}$$

记$\widehat{\omega^*}$、$\widehat{\alpha^*}$和$\widehat{\beta^*}$的渐近方差分别为 $\sigma_{\omega^*}^2$、$\sigma_{\alpha^*}^2$和 $\sigma_{\beta^*}^2$,则$\widehat{\omega^*}$、$\widehat{\alpha^*}$和$\widehat{\beta^*}$的渐近性质为

$$\sqrt{n}(\widehat{\omega^*} - \omega_0^*) \xrightarrow{d} N\left(0, \frac{\sigma_{\omega^*}^2}{\mu_H^4}\right), n \to \infty,$$

$$\sqrt{n}(\widehat{\alpha^*} - \alpha_0^*) \xrightarrow{d} N\left(0, \frac{\sigma_{\alpha^*}^2}{\mu_H^4}\right), n \to \infty,$$

$$\sqrt{n}(\widehat{\beta^*} - \beta_0^*) \xrightarrow{d} N\left(0, \frac{\sigma_{\beta^*}^2}{\mu_H^4}\right), n \to \infty。$$

2　portmanteau Q 检验

portmanteau Q 检验是一个检验 GARCH 类拟合模型是否充分的常用工具。该检验的统计量通常由残差平方自相关函数构成,但 QMELE 方法放宽了残差阶矩条件,因此,这里使用残差绝对值自相关函数。残差估计由波动率估计决定,而从前文可知,日间模型得到的波动率估计$\widetilde{\sigma_t}$,有别于日内模型即高频数据模型得到的波动率估计$\widehat{\sigma_t}$,而且参数的渐近性质也有差异。因此,在估计更准确的情况下,高频数据的引入也会使得 portmanteau Q 检验更精准。为了比较它们的差异,下面具体介绍这两种 portmanteau Q 检验。

2.1　传统的 portmanteau Q 检验

根据日频 GARCH 模型-可知,日频模型的残差估计为$\widetilde{\varepsilon_t} = \frac{y_t}{\widetilde{\sigma_t}}$,则样本残差绝对值自相关函数公式如下:

$$\widetilde{r_k} = \frac{\sum_{t=k+1}^{n} \left(\frac{| y_t |}{\widetilde{\sigma_t}} - 1\right)\left(\frac{| y_{t-k} |}{\widetilde{\sigma_{t-k}}} - 1\right)}{\sum_{t=1}^{n} \left(\frac{| y_t |}{\widetilde{\sigma_t}} - 1\right)^2},$$

$$k = 1, 2, 3, \cdots$$

当样本残差绝对值自相关函数的方差存在时,根据中心极限定理和 Wann-Wald 定理[19],可知残差绝对值自相关函数渐近正态。当 $n \to \infty$ 时,取$\widetilde{r_k}$的最大滞后阶数为 m,记$\widetilde{\boldsymbol{R}_M} = (\widetilde{r_1}, \widetilde{r_2}, \cdots, \widetilde{r_m})'$,

$$\sqrt{n}\widetilde{\boldsymbol{R}_M} \xrightarrow{d} N(0, \boldsymbol{V}),$$

其中,$\boldsymbol{V}$ 为$\widetilde{\boldsymbol{R}_M}$的渐近方差。由此可以构建卡方检

验统计量

$$\widetilde{Q} = n\,\widetilde{\boldsymbol{R}_M}\widetilde{\boldsymbol{V}}^{-1}\widetilde{\boldsymbol{R}_M}' \sim \chi^2(m)。$$

获得卡方检验统计量便可以对模型的充分性进行检验,而求解该检验统计量的重点在于求解$\widetilde{R_M}$的渐近方差。当$E\varepsilon_t^2 < \infty$且$E|\varepsilon_t| = 1$时,由大数定律可得

$$\frac{1}{n}\sum_{t=1}^{n}\left(\frac{|y_t|}{\widetilde{\sigma_t}} - 1\right)^2 \to E(|\varepsilon_t| - 1)^2 = var(|\varepsilon_t|)。$$

也就是说,$\widetilde{r_k}$的分母乘以$\frac{1}{n}$后收敛到常数,因此,若将$\widetilde{r_k}$分子分母同乘以$\frac{1}{n}$,即得到

$$\widetilde{r_k} = \frac{\frac{1}{n}\sum_{t=k+1}^{n}\left(\frac{|y_t|}{\widetilde{\sigma_t}} - 1\right)\left(\frac{|y_{t-k}|}{\widetilde{\sigma_{t-k}}} - 1\right)}{\frac{1}{n}\sum_{t=1}^{n}\left(\frac{|y_t|}{\widetilde{\sigma_t}} - 1\right)^2},$$
$$k = 1,2,\cdots,m。 \quad (16)$$

等式(16)分母收敛到常数,所以要考虑$\widetilde{R_M}$的渐近方差,考虑等式(16)分子的渐近方差即可。为方便计算,记

$$\widetilde{C_k} = \frac{1}{n}\sum_{t=k+1}^{n}\left(\frac{|y_t|}{\widetilde{\sigma_t}} - 1\right)\left(\frac{|y_{t-k}|}{\widetilde{\sigma_{t-k}}} - 1\right),$$
$$k = 1,2,\cdots,m。 \quad (17)$$

当等式(17)中$k=0$时,等式(17)退化为等式(16)的分母,记为$\widetilde{C_0}$。

当$\frac{|y_t|}{\sigma_t}$为白噪声过程时,$\sqrt{n}\boldsymbol{R}_M \xrightarrow{d} N(0,\boldsymbol{I}_m)$,$\boldsymbol{I}_m$表示$m\times m$的单位阵,则$\sqrt{n}\,\boldsymbol{C}_M \xrightarrow{d} N(0, var(|\varepsilon_t|)\boldsymbol{I}_m)$,其中,$\boldsymbol{C}_M = (C_1, C_2, \cdots, C_m)'$。因此,若要求$\widetilde{\boldsymbol{C}_M}$的渐近分布,可考虑$\widetilde{\boldsymbol{C}_M}$的泰勒一阶展开式

$$\widetilde{\boldsymbol{C}_M} = \boldsymbol{C}_M + \frac{\partial \boldsymbol{C}}{\partial \theta}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}),$$

其中,

$$\frac{\partial C_k}{\partial \boldsymbol{\theta}} = -\frac{1}{n}\sum_{t=k+1}^{n}\frac{1}{2}\frac{|y_t|}{\sigma_t}\frac{1}{\sigma_t^2}\frac{\partial \sigma_t^2}{\partial \boldsymbol{\theta}}\left(\frac{|y_{t-k}|}{\sigma_{t-k}} - 1\right) - \frac{1}{n}\sum_{t=k+1}^{n}\frac{1}{2}\left(\frac{|y_t|}{\sigma_t} - 1\right)\frac{|y_{t-k}|}{\sigma_{t-k}}\frac{1}{\sigma_{t-k}^2}\frac{\partial \sigma_{t-k}^2}{\partial \boldsymbol{\theta}}。 \quad (18)$$

由于$E(\varepsilon_t\varepsilon_{t-k}) = 0$且$E|\varepsilon_t| = 1$,上式(18)可简化为式(19),并将其记为$X_k, k = 1,2,\cdots,m$,则$\boldsymbol{X} = (X_1, X_2, \cdots, X_m)'$。

$$\frac{\partial C_k}{\partial \boldsymbol{\theta}} \approx -\frac{1}{2n_t}\sum_{t=k+1}^{n}\frac{1}{\sigma_t^2}\frac{\partial \sigma_t^2}{\partial \boldsymbol{\theta}}\left(\frac{|y_{t-k}|}{\sigma_{t-k}} - 1\right) \triangleq X_k。 \quad (19)$$

由于$\sqrt{n}\boldsymbol{C}_M$和$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$渐近分布已知,所以求解$\sqrt{n}\widetilde{\boldsymbol{C}_M}$的渐近方差的重点变为求解$\sqrt{n}\boldsymbol{C}_M$和$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$的协方差。要求解这两者的协方差,需对$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$进行等式变换。在所假设的正则条件下,

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -4\sqrt{n}\boldsymbol{G}^{-1}(1 + o_p(1))\frac{\partial L_n}{\partial \boldsymbol{\theta}}, \quad (20)$$

其中,似然函数为

$$L_n = \sum_{t=1}^{n}\log(\sigma_t) + \frac{|y_t|}{\sigma_t}。 \quad (21)$$

由于$\sqrt{n}\boldsymbol{C}_M$和$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$都是零均值分布,所以$cov(\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \sqrt{n}\boldsymbol{C}_M) = nE((\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\boldsymbol{C}_M')$,再根据等式(20),协方差可变换为$cov(\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \sqrt{n}\boldsymbol{C}_M) \approx -4n\boldsymbol{G}^{-1}E(\frac{\partial L_n}{\partial \boldsymbol{\theta}}\boldsymbol{C}_M')$。计算可得$E(\frac{\partial L_n}{\partial \boldsymbol{\theta}}C_k) = var(|\varepsilon_t|)X_k$。因此,

$$var(\sqrt{n}\widetilde{\boldsymbol{C}_M}) \approx \boldsymbol{X}var(\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0))\boldsymbol{X}' + var(\sqrt{n}\boldsymbol{C}_M) + 2\boldsymbol{X}cov(\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \sqrt{n}\boldsymbol{C}_M) = (var(|\varepsilon_t|))^2\boldsymbol{I}_m - 4var(|\varepsilon_t|)\boldsymbol{X}\boldsymbol{G}^{-1}\boldsymbol{X}'。$$

又由于$\sqrt{n}\widetilde{\boldsymbol{R}_M}$服从正态分布,所以$\sqrt{n}\widetilde{\boldsymbol{C}_M}$也服从正态分布。记$var(\sqrt{n}\widetilde{\boldsymbol{C}_M}) = \boldsymbol{D}$,$\widetilde{\boldsymbol{D}}$为$\boldsymbol{D}$在低频数据下的估计值,则下方统计量服从卡方分布:

$$\widetilde{\boldsymbol{Q}_c} = n\,\widetilde{\boldsymbol{C}_M}\widetilde{\boldsymbol{D}}^{-1}\widetilde{\boldsymbol{C}_M}' \sim \chi^2(m)。$$

2.2 基于高频信息的 portmanteau Q 检验

由于基于低频信息和高频信息得到的参数估计有差异,基于高频信息得到的波动率估计$\widehat{\sigma_t}$与基于低频信息得到的波动率估计$\widetilde{\sigma_t}$也有差异,因此,基于高频信息的样本残差绝对值自相关函数同样随之变化,具体公式如下:

$$\widehat{r_k} = \frac{\sum_{t=k+1}^{n}\left(\frac{|y_t|}{\widehat{\sigma_t}} - 1\right)\left(\frac{|y_{t-k}|}{\widehat{\sigma_{t-k}}} - 1\right)}{\sum_{t=1}^{n}\left(\frac{|y_t|}{\widehat{\sigma_t}} - 1\right)^2},$$
$$k = 1,2,3,\cdots \quad (22)$$

虽然基于高频信息的样本残差绝对值自相关函数与低频的有差异,但二者都为样本残差自相关函数,都代表模型提取充分信息后残差的自相关情况,所以当模型充分时,基于高频信息的样本残差绝对值自相关函数$\widehat{r_k}$满足渐近正态性。因此,取$\widehat{r_k}$的最大滞后阶数为 m,记$\widehat{\boldsymbol{R}_M}=(\widehat{r_1},\widehat{r_2},\cdots,\widehat{r_m})'$,$\boldsymbol{V}_h$ 为$\widehat{\boldsymbol{R}_M}$的渐近方差,

$$\sqrt{n}\widehat{\boldsymbol{R}_M}\xrightarrow{d}N(0,\boldsymbol{V}_h),n\to\infty。$$

与低频数据下的样本自相关函数处理同理,先对等式(22)分子分母同乘以$\frac{1}{n}$,得到变换后的等式如下:

$$\widehat{r_k}=\frac{\frac{1}{n}\sum_{t=k+1}^{n}(\frac{|y_t|}{\widehat{\sigma_t}}-1)(\frac{|y_{t-k}|}{\widehat{\sigma_{t-k}}}-1)}{\frac{1}{n}\sum_{t=1}^{n}(\frac{|y_t|}{\widehat{\sigma_t}}-1)^2},$$
$$k=1,2,\cdots,m。\qquad(23)$$

同理,等式(23)的分母收敛,因此,求解$\widehat{r_k}$的渐近分布转换成为求解$\widehat{C_k}$的渐近分布,其中,$\widehat{C_k}$具体公式如下:

$$\widehat{C_k}=\frac{1}{n}\sum_{t=k+1}^{n}(\frac{|y_t|}{\widehat{\sigma_t}}-1)(\frac{|y_{t-k}|}{\widehat{\sigma_{t-k}}}-1),$$
$$k=1,2,\cdots,m。$$

记$\widehat{\boldsymbol{C}_M}=(\widehat{C_1},\widehat{C_2},\cdots,\widehat{C_m})'$,则同理,若要求$\widehat{\boldsymbol{C}_M}$的渐近方差,可以考虑对其进行一阶泰勒展开:

$$\widehat{\boldsymbol{C}_M}=\boldsymbol{C}_M+\frac{\partial\boldsymbol{C}}{\partial\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})。$$

由高频数据下使用的似然函数(11)可知,其是先通过求解 $\boldsymbol{\theta}^*$ 的估计值进而去求解原参数 $\boldsymbol{\theta}$。而根据等式(15)可知,$\boldsymbol{\theta}^*$ 与 $\boldsymbol{\theta}$ 存在以下转换关系式:

$$\hat{\boldsymbol{\theta}}=\boldsymbol{T}\widehat{\boldsymbol{\theta}^*},\boldsymbol{T}=diag(\frac{1}{\mu_H^2},\frac{1}{\mu_H^2},1)。$$

由此,上方的泰勒展开可以转换为

$$\widehat{\boldsymbol{C}_M}=\boldsymbol{C}_M+\frac{\partial\boldsymbol{C}}{\partial\boldsymbol{\theta}}\boldsymbol{T}(\widehat{\boldsymbol{\theta}^*}-\boldsymbol{\theta}^*)。$$

已知$\sqrt{n}\boldsymbol{C}$ 和$\sqrt{n}(\widehat{\boldsymbol{\theta}^*}-\boldsymbol{\theta}^*)$的渐近分布,要求$\sqrt{n}\widehat{\boldsymbol{C}}$的方差,则需求协方差 $cov(\sqrt{n}\boldsymbol{C},\sqrt{n}(\widehat{\boldsymbol{\theta}^*}-\boldsymbol{\theta}^*))$。求计算协方差前,要对$\sqrt{n}(\widehat{\boldsymbol{\theta}^*}-\boldsymbol{\theta}^*)$进行等式变换。

考虑到极大似然估计值在似然函数的一阶偏导数为0,所以可以尝试对一阶偏导数进行泰勒展开,得到

$$0=\frac{\partial L_n^*(\widehat{\boldsymbol{\theta}_0^*})}{\partial\boldsymbol{\theta}^*}=\frac{\partial L_n^*(\boldsymbol{\theta}_0^*)}{\partial\boldsymbol{\theta}^*}+\frac{\partial^2 L_n^*(\boldsymbol{\theta}_0^*)}{\partial\boldsymbol{\theta}^*\partial\boldsymbol{\theta}^{*'}}(\widehat{\boldsymbol{\theta}^*}-\boldsymbol{\theta}_0^*)+\boldsymbol{o}_p(1)。$$

整理,可得

$$\sqrt{n}(\widehat{\boldsymbol{\theta}^*}-\boldsymbol{\theta}_0^*)=-4\sqrt{n}\boldsymbol{G}^{*-1}(1+\boldsymbol{o}_p(1))\frac{\partial L_n^*}{\partial\boldsymbol{\theta}^*}。$$

进而可求协方差

$$E(\frac{\partial L_n^*}{\partial\boldsymbol{\theta}^*}C_k)=E[\sum_{t=1}^{n}\frac{1}{2}\frac{1}{\sigma_t^{*2}}\frac{\partial\sigma_t^{*2}}{\partial\boldsymbol{\theta}^*}(1-\frac{|H_t|}{\sigma_t^*})\times\frac{1}{n}\sum_{t=k+1}^{n}(\frac{|y_t|}{\sigma_t}-1)(\frac{|y_{t-k}|}{\sigma_{t-k}}-1)]=$$
$$-E[\frac{1}{2n}\sum_{t=k+1}^{n}\frac{1}{\sigma_t^{*2}}\frac{\partial\sigma_t^{*2}}{\partial\boldsymbol{\theta}^*}(\frac{|H_t|}{\sigma_t^*}-1)(\frac{|y_t|}{\sigma_t}-1)(\frac{|y_{t-k}|}{\sigma_{t-k}}-1)]=$$
$$-\frac{1}{2n}\sum_{t=k+1}^{n}\frac{1}{\sigma_t^{*2}}\frac{\partial\sigma_t^{*2}}{\partial\boldsymbol{\theta}^*}(\frac{|y_{t-k}|}{\sigma_{t-k}}-1)E[(\frac{|H_t|}{\sigma_t^*}-1)(\frac{|y_t|}{\sigma_t}-1)]。$$

所以,

$$var(\sqrt{n}\widehat{\boldsymbol{C}})\approx\boldsymbol{X}var(\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0))\boldsymbol{X}'+var(\sqrt{n}\boldsymbol{C}_M)+2\boldsymbol{X}cov(\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0),\sqrt{n}\boldsymbol{C}_M)=$$
$$\boldsymbol{XT}var(\sqrt{n}(\widehat{\boldsymbol{\theta}^*}-\boldsymbol{\theta}_0^*))\boldsymbol{T}'\boldsymbol{X}'+var(\sqrt{n}\boldsymbol{C})+2\boldsymbol{XT}cov(\sqrt{n}(\widehat{\boldsymbol{\theta}^*}-\boldsymbol{\theta}_0^*),\sqrt{n}\boldsymbol{C}_M)=$$
$$4var(|\varepsilon_t^*|)\boldsymbol{XTG}^{*-1}\boldsymbol{T}'\boldsymbol{X}'+var(|\varepsilon_t|)\boldsymbol{I}_m-8\boldsymbol{XTG}^{*-1}E(\frac{\partial L_n^*}{\partial\boldsymbol{\theta}^*}\boldsymbol{C})。$$

$\sqrt{n}\widehat{\boldsymbol{R}_M}$服从正态分布,则$\sqrt{n}\widehat{\boldsymbol{C}_M}$也服从正态分布。记方差 $var(\sqrt{n}\widehat{\boldsymbol{C}_M})=\boldsymbol{D}$,$\widehat{\boldsymbol{D}}$为 $\boldsymbol{D}$ 在低频数据下的估计值,那么下方统计量服从卡方分布:

$$\widehat{Q_c}=n\widehat{\boldsymbol{C}_M}\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{C}_M}'\sim\chi^2(m)。$$

3　模拟分析

本小节通过模拟符合波动率模型的证券高频

数据来验证理论的正确性。已知要给模型参数 $\boldsymbol{\theta}$ 赋值以生成日内对数收益率 $Y_t(u)$,根据模型(3)~(4)可知,生成 $Y_t(u)$ 前,需先生成日内随机过程 $Z_t(u)$。借鉴 Visser[7] 可根据以下公式生成符合假设条件的随机过程 $Z_t(u)$:

$$\mathrm{d}\Gamma_t(u) = -\delta(\Gamma_t(u)-\mu_\Gamma)\mathrm{d}u+\sigma_\Gamma \mathrm{d}B_t^{(2)}(u),$$
$$\mathrm{d}\Psi_t(u) = \exp(\Gamma_t(u))\mathrm{d}B_t^{(1)}(u),\quad u\in[0,1],$$
$$m = E|\Psi_t(1)|,\quad Z_t(u)=\frac{\Psi_t(u)}{m},\qquad (24)$$

其中,$B_t^{(1)}(u)$ 和 $B_t^{(2)}(u)$ 数据由两个不相关的布朗运动生成。做初始化,令 $Z_t(0)=0$,$\Gamma_t(0)$ 从正态分布 $N(\mu_\Gamma,\sigma_\Gamma^2)$ 中随机产生,设

$$\delta=\frac{1}{2},\ \mu_\Gamma=\frac{1}{4},\ \sigma_\Gamma=-\frac{1}{16}。$$

对于 u,在中国证券市场每天开放 240 min 的背景下,令 $\mathrm{d}u=1/240$ 可以实现模拟市场每分钟的交易情况。最后为了满足 $E|Z_t(1)|=1$ 的矩条件,设定等式。在以上设定下,可顺利生成序列 $Z_t(u)$。接下来,分别设定参数 $\boldsymbol{\theta}_0=(0.01,0.08,0.8)'$ 和 $\boldsymbol{\theta}_0=(0.03,0.06,0.9)'$ 两种情况来生成 $Y_t(u)$。

进一步地,要使用高频模型-,需先选择波动率代表对高频序列 $Y_t(u)$ 进行处理。这里选取已实现波动率(RV)作为波动率代表,并分别选取 5 min、15 min 和 30 min 作为采样频率,即分别记 $H_t=\mathrm{RV5}$、$H_t=\mathrm{RV15}$ 和 $H_t=\mathrm{RV30}$。记 $u_1,u_2,\cdots,u_{240}$ 为不同时间间隔点,表示开盘每分钟的时间点,同时初始化 $Y_t(u_0)=Y_t(0)=0$。以 RV5 为例,波动率代表的计算公式具体为

$$H_t=\mathrm{RV5}=\sqrt{\sum_{i=1}^{48}\left[Y_t(u_{5i})-Y_t(u_{5(i-1)})\right]^2}。$$

为了以低频数据下的模型效果作为参照,也取 $H_t=|Y_t(1)|=|y_t|$ 作为对比。设原假设分布的显著性水平为 0.05,取滞后自相关函数的阶数 m 为 6,则卡方检验的自由度为 6。根据第 2 节的估计方法可得参数的估计结果,进而可计算检验统计量的数值。重复 1 000 次试验,记录检验统计量超过原假设卡方检验统计量 0.95 分位数的比例,即可得到检验统计量的显著性水平大小。最终,各波动率代表模型使用 QMELE 估计后的检验统计量的显著性水平结果如表 1 所示。

表 1 基于 QMELE 的各波动率代表模型的检验显著性水平

Table 1 Empirical size of each volatility proxy model based on QMELE

参数真值	波动率代表	$n=1\,000$	$n=1\,500$	$n=2\,000$
$\boldsymbol{\theta}_0=(0.01,0.08,$	$\lvert y_t\rvert$	0.155 0	0.163 0	0.142 0
$0.8)'$	RV30	0.048 0	0.058 0	0.057 0
	RV15	0.043 0	0.050 0	0.051 0
	RV5	0.048 0	0.048 0	0.047 0
$\boldsymbol{\theta}_0=(0.03,0.06,$	$\lvert y_t\rvert$	0.080 0	0.118 0	0.104 0
$0.9)'$	RV30	0.046 0	0.048 0	0.061 0
	RV15	0.044 0	0.052 0	0.054 0
	RV5	0.043 0	0.051 0	0.050 0

从表 1 可以看出,基于高频数据模型的检验显著性水平明显更接近 0.05。检验的显著性水平越接近 0.05,说明该检验统计量 0.95 的分位数越接近卡方分布 0.95 的分位数。检验统计量显著性水平的大小是一个判断统计量的分布是否符合本文所假设的卡方分布的标准。基于高频信息的检验统计量的显著性水平越接近 0.05,说明在一定程度上,基于高频数据模型的检验统计量更近似理论推导的卡方分布。

选取的卡方检验统计量由残差自相关函数构成,当模型充分时,残差为白噪声过程,则统计量服从推导的卡方分布。那么相对地,当残差自相关时,统计量不会服从推导的卡方分布。因此,在模拟时,可以设定生成的高频随机误差项具有自相关关系,从而确定所提出的统计量是否能识别不充分的模型。一般可以设定高频误差项存在自相关关系如下:

$$Z_t(u)=0.1Z_{t-1}(u)+a_t,\qquad (25)$$

其中,a_t 由服从标准正态分布的随机数产生。特别地,当 $u=1$ 时,误差项也存在如下自相关关系:

$$\varepsilon_t=Z_t(1)=0.1Z_{t-1}(1)+a_t=0.1\varepsilon_{t-1}++a_t。$$

依据上面公式(25)生成 $Z_t(u)$ 数据,即生成高频误差项线性自相关的模拟数据。重复 1 000 次试验,计算高频残差自相关情况下检验统计量大于卡方分布 0.95 的分位数的比例,便可得到高频误差项自相关情况下检验统计量的功效,具体得到

检验功效结果如表 2 所示。

表 2　基于高频残差线性自相关的各波动率代表模型的检验功效

Table 2　Power of each volatility proxy model based on linear autocorrelation of high-frequency residuals

参数真值	波动率代表	$n=1\ 000$	$n=1\ 500$	$n=2\ 000$
$\boldsymbol{\theta}_0=(0.01,0.08,0.8)'$	$\lvert y_t\rvert$	0.155 0	0.157 0	0.148 0
	RV30	1.000 0	1.000 0	1.000 0
	RV15	1.000 0	1.000 0	1.000 0
	RV5	1.000 0	1.000 0	1.000 0
$\boldsymbol{\theta}_0=(0.03,0.06,0.9)'$	$\lvert y_t\rvert$	0.098 0	0.084 0	0.108 0
	RV30	1.000 0	1.000 0	1.000 0
	RV15	1.000 0	1.000 0	1.000 0
	RV5	1.000 0	1.000 0	1.000 0

从表 2 可以明显看出,基于高频数据检验统计量的功效明显优于基于低频数据的结果,也就是说,当高频误差项存在线性自相关时,基于高频数据的 portmanteau Q 检验能更好地识别其相关性。这是符合预想的,因为基于高频数据的检验统计量包含了高频残差信息,当高频时刻的残差之间都存在明显的相关性时,基于高频信息的检验统计量能够更好地捕捉其相关性。

值得说明的是,在这里添加的白噪声选择使用低频噪声 a_t,而不是高频,一方面是为了对比,另一方面由于引入高频数据的模型是使用波动率代表降到一维进行求解的,即模型(8)~(9),其中高频残差也是通过波动率代表处理的。也就是说,最终的计算是一维层面的,而实际模型使用的是噪声,是波动率代表降维后的误差 $H(Z_t(u))$。根据波动率代表的性质,以等式(25)为例,当使用已实现波动率作为波动率代表时,降维后的高频残差相关性如下:

$$H(Z_t(u))=H(0.1Z_{t-1}(u)+a_t)=0.1H(Z_{t-1}(u))。\tag{26}$$

从式(26)可以看出,这里添加的白噪声 a_t 在使用波动率代表时,噪声会被消除,因此,在这种情况下,使用高频信息检验统计量的残差自相关会增强,这也是基于高频数据检验的优势。不过这是比较强的假设,若添加的是乘性噪声,则添加的白噪声不会被消除。添加乘性噪声可使高频残差满足如下关系:

$$Z_t(u)=0.1Z_{t-1}(u)a_t。$$

当使用波动率代表时,

$$H(Z_t(u))=0.1H(Z_{t-1}(u))a_t。$$

对于乘性白噪声也计算了检验统计量的功效结果。但在这个假设下,计算统计量的过程中,需要求逆的矩阵接近奇异值,因此,检验统计量的结果都远大于 0.95 分位数,从而导致所有检验功效结果都为 1。这使得基于高频数据的结果和基于低频数据的没有差异。但求逆矩阵接近奇异值也是一种自相关性表现的情况,所以,这也验证了本文所提出的检验统计量是具有识别高频残差相关能力的。

不管是基于加性白噪声还是基于乘性白噪声的高频误差项自相关假设,都是比较强的假设,而实际过程可能自相关性是低频的,这可能不利于基于高频数据的模型,但这是可能存在的情况,所以,也应当观察这种情形下的检验功效结果。可以假设低频随机误差项之间存在的线性关系如下:

$$Z_t(1)=\varepsilon_t=b\varepsilon_{t-1}+a_t=bZ_{t-1}(1)+a_t,\tag{27}$$

其中,b 为自相关系数。在这里,初步设置低频噪声 t 时刻与 $t-1$ 时刻的自相关系数 b 为 0.1,而除了 $u=1$ 时刻,其他 u 高频时刻还由原来的过程产生。最终得到基于高频信息 RV5 波动率代表模型的检验统计功效与低频模型的结果对比如表 3 所示。因为各高频波动率代表模型的差异不大(图 1),所以表 3 只列举了 RV5 的检验功效结果。

表 3　基于低频残差线性自相关的波动率代表模型的检验功效

Table 3　Power of each volatility proxy model based on linear autocorrelation of low-frequency residuals

参数真值	波动率代表	$n=1\ 000$	$n=1\ 500$	$n=2\ 000$
$\boldsymbol{\theta}_0=(0.01,0.08,0.8)'$	$\lvert y_t\rvert$	0.159 0	0.129 0	0.131 0
	RV5	0.060 0	0.069 0	0.065 0
$\boldsymbol{\theta}_0=(0.03,0.06,0.9)'$	$\lvert y_t\rvert$	0.100 0	0.106 0	0.101 0
	RV5	0.069 0	0.064 0	0.076 0

从表 3 可以看到,基于高频信息检验功效结果并不优于基于低频信息的,而且基于高频信息的检验不能很好地识别低频残差自相关性,明显有一部分原因是高频信息的检验统计量包含更多噪声信息,但这里的自相关性是低频的。对比检验显著性水平可以看出,基于高频信息的检验功

效结果略大于检验显著性水平。另外,基于高频的检验功效结果不佳可能还与相关性较小有关,设想如果相关性够强,若随着相关性的增大,统计量的分布与原假设的分布应当会更有差异,而检验功效的结果也会逐渐趋向于1。为了验证这一点,将低频噪声相关系数 b 分别增大为0.2、0.3 和0.4,重复1 000次试验,观察基于高频信息检验统计量的功效变化,得到具体结果如图1所示。

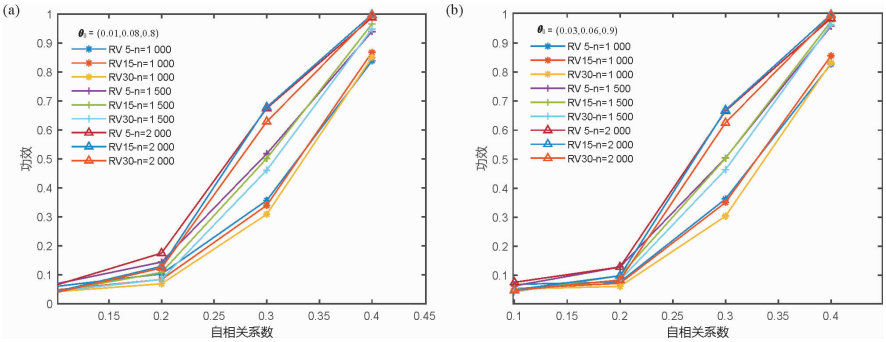


图1 基于低频残差线性自相关的高频波动率代表模型的检验功效变化图

Fig. 1 Power of the high-frequency volatility proxy models based on linear autocorrelation of low-frequency residuals

图1(a)为设定参数 $\boldsymbol{\theta}_0=(0.01,0.08,0.8)'$ 时,图1(b)为设定参数 $\boldsymbol{\theta}_0=(0.03,0.06,0.9)'$ 时,当式(27)中 b 分别为0.1、0.2、0.3 和0.4 时,基于各高频波动率代表 RV5、RV15 和 RV30 的检验功效变化图。从图1中可以看出,随着自相关系数增大,各高频波动率代表模型的检验功效变化明显,且逐渐向1靠拢,符合预期。此外,还可以看出,随着样本量的增大,检验功效的结果也逐渐增大,并且各波动率代表特征一致。

设定自相关等式(27)是因为该等式直观显示了低频误差项的线性自相关性,但实际过程中,残差自相关可能更复杂,比如由于 GARCH 模型的阶数选择错误导致的误差自相关。GARCH 模型的阶数由公式确定,由于这是一个低频过程,所以阶数选择错误导致的自相关性也应当是低频的,这也是笔者引入低频残差自相关研究的一个原因。下面将以 GARCH(1,2) 为例,即以等式(28)代替等式(4)生成模拟数据:

$$\sigma_t^2=\omega+\alpha y_{t-1}^2+\beta\sigma_{t-1}^2+\beta_2\sigma_{t-2}^2, \qquad (28)$$

其中,为了对比,$(\omega,\alpha,\beta)'$ 依旧分别设为$(0.01,0.08,0.8)'$和$(0.03,0.06,0.9)'$,同时,为了比较检验功效在残差自相关性强弱下的变化,β_2 分别设置为0.1、0.2、0.3、0.4 和0.5。生成模拟数据后,再用 GARCH(1,1) 模型拟合数据,重复1 000次试验,最终得到检验统计量的功效结果,具体如表4所示。

表4 基于 GARCH(2,1)下的波动率代表模型的检验功效

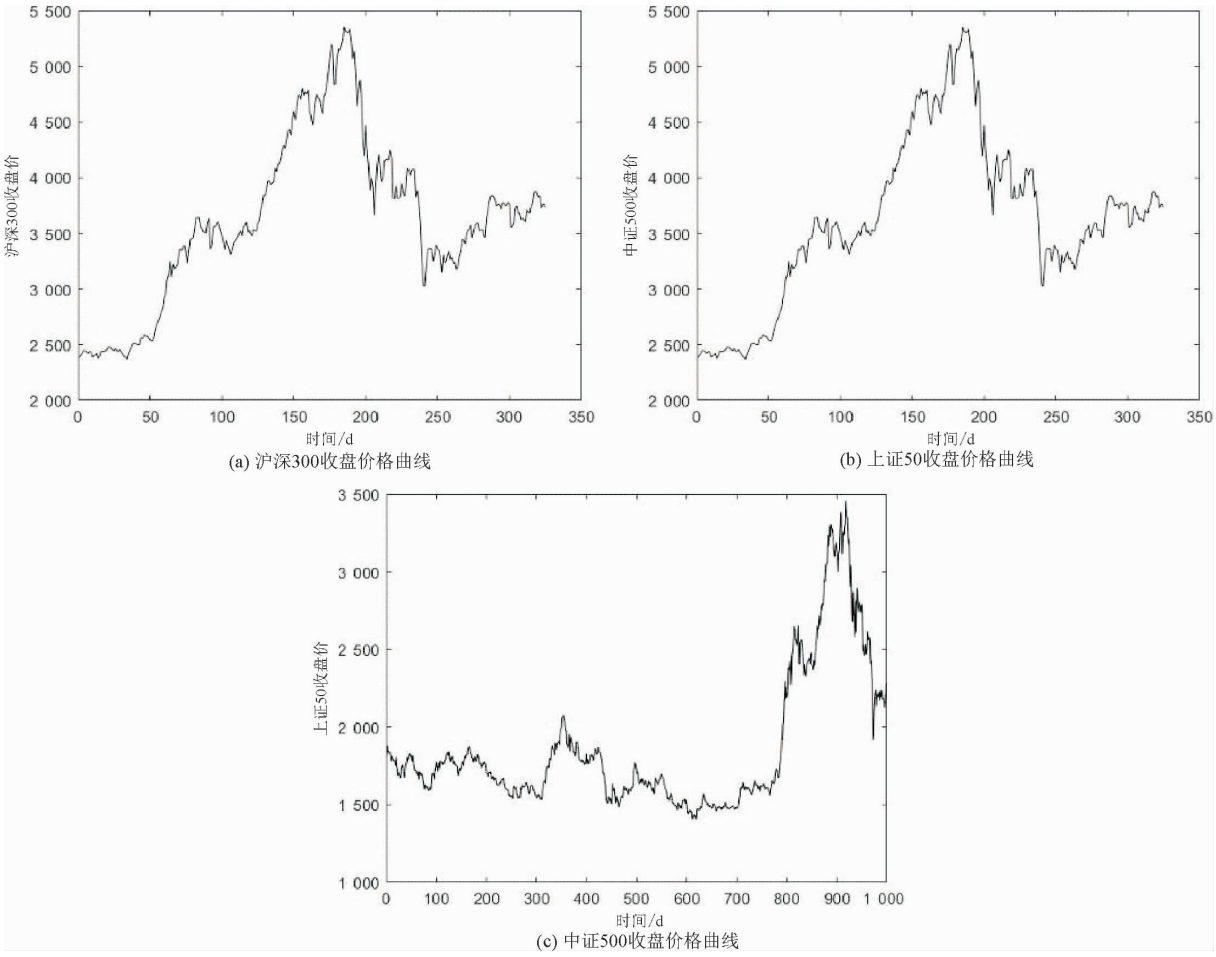
Table 4 Power of the volatility proxy models based on GARCH(2,1)

参数	$\boldsymbol{\theta}_0=(0.01,0.08,0.8,\beta_2)'$					$\boldsymbol{\theta}_0=(0.03,0.06,0.9,\beta_2)'$				
	β_2	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
$n=1\,000$	$\lvert y_t\rvert$	0.105	0.052	0.514	0.982	1.000	0.052	0.583	0.989	1.000	1.000
	RV30	0.055	0.047	0.248	0.789	0.985	0.047	0.294	0.824	0.990	0.998
	RV15	0.057	0.050	0.258	0.799	0.984	0.050	0.299	0.827	0.992	0.998
	RV5	0.059	0.049	0.259	0.801	0.985	0.051	0.301	0.826	0.993	0.998
$n=1\,500$	$\lvert y_t\rvert$	0.125	0.053	0.723	1.000	1.000	0.043	0.801	1.000	1.000	1.000
	RV30	0.045	0.043	0.427	0.950	1.000	0.042	0.501	0.980	1.000	1.000
	RV15	0.043	0.047	0.433	0.950	1.000	0.041	0.511	0.981	1.000	1.000
	RV5	0.041	0.048	0.431	0.952	1.000	0.041	0.509	0.978	1.000	1.000
$n=2\,000$	$\lvert y_t\rvert$	0.108	0.070	0.879	1.000	1.000	0.054	0.912	1.000	1.000	1.000
	RV30	0.053	0.065	0.611	0.997	1.000	0.050	0.645	0.997	1.000	1.000
	RV15	0.050	0.068	0.625	0.997	1.000	0.051	0.654	0.997	1.000	1.000
	RV5	0.052	0.065	0.626	0.997	1.000	0.052	0.657	0.997	1.000	1.000

从表 4 可以看出,当低频残差相关性较小时,各波动率代表模型都不能很好地辨别 GARCH(1,1)模型和 GARCH(1,2)模型,但当 β_2 比较显著时,特别是样本量较大的时候,可以很好地识别。此外,虽然基于低频数据检验统计量的功效结果较大,但当 $\beta_2 \leqslant 0.1$ 时,它也不能很好地识别模型阶数。因此,总体来说,受噪声影响,当低频残差自相关性较小时,检验统计量都有一定的误判率。此外,尽管在低频残差自相关情形下,基于高频信息的检验统计量并不占优势,但当低频残差相关性较大时,在一定的样本量规模下,本文所提出的检验统计量是可以判别 GARCH 模型阶数选择错误问题的,具有实用性。

4　实证分析

本节研究所提出的检验在实际数据的实践情况。研究数据分别为 2014 年 9 月 2 日至 2015 年 12 月 31 日沪深 300 指数(CSI 300)、2014 年 1 月 7 日至 2015 年 10 月 13 日中证 500 指数(CSI 500)以及 2011 年 8 月 24 日至 2015 年 10 月 13 日上证 50 指数(SSE 50)的每日 1 min 间隔价格数据,记第 t 天的 1 min 间隔价格数据为 $P_t(u)$,每天 240 个观测值,分别包含 325、429、1 000 个交易日。而 $P_t(1)$ 表示第 t 天的收盘价格,3 个指数每天收盘价格见图 2。



(a) 沪深300收盘价格曲线　(b) 上证50收盘价格曲线

(c) 中证500收盘价格曲线

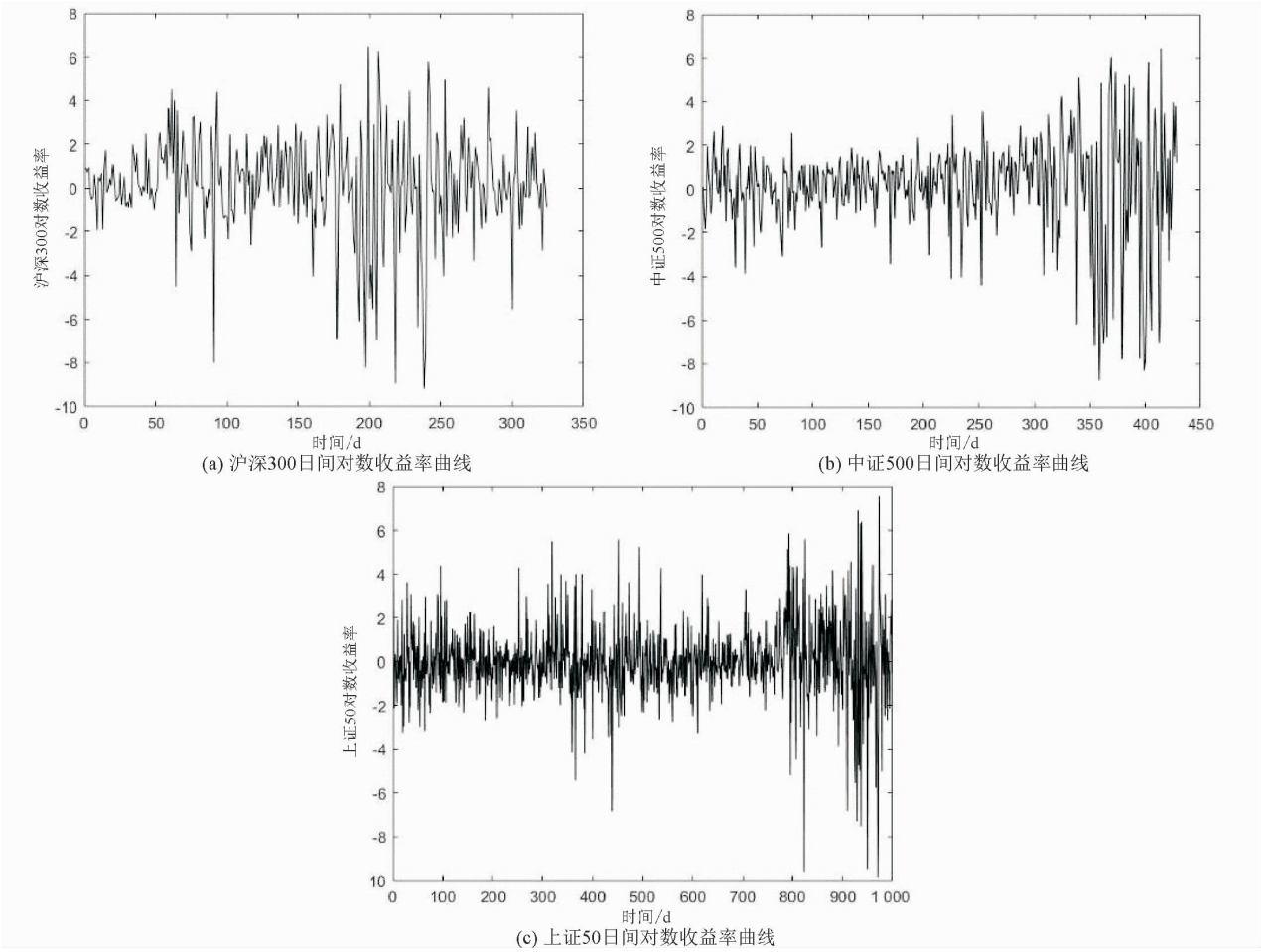
图 2　各指数收盘价格曲线图

Fig. 2　Closing price curves of various indices

从图 2 可以看出,3 个指数的收盘价格序列都不平稳,但可通过取对数收益率对其进行平稳化处理。同样,对高频数据也采用对数收益率公式[7],具体对数收益率计算公式如下:

$$Y_t(u) = [\log P_t(u) - \log P_{t-1}(u)] \times 100。 \quad (29)$$

以收盘价格为例,取 $u=1$,使用公式(29)后,可以得到3个指数的日间收盘价对数收益率,结果如图3所示。



(a) 沪深300日间对数收益率曲线 (b) 中证500日间对数收益率曲线 (c) 上证50日间对数收益率曲线

图3 各指数日间对数收益率曲线图

Fig. 3 Intraday log-return curves of various indices

从图3可以看出,使用对数收益率公式后,平稳化效果明显。3个指数的日间对数收益率都在0附近震荡,符合平稳序列特征,但从中也可以看出,3个指数的对数收益率在不同时间段波动存在明显差异,有异方差特征,可以使用GARCH(1,1)模型尝试对数据进行拟合。

在进行模型估计前,需采用波动率代表对高频数据进行降维,与模拟一致,选择已实现波动率,并分别选取5 min、15 min和30 min的抽样频率,得到波动率代表 H_t,分别记为RV5、RV15和RV30。同时,选取波动率代表 $H_t = |y_t|$ 作为日间模型,与之进行对比。经过处理后,便可采用文中的QMELE估计方法得到估计参数结果,见表5。

表5 各指数各波动率代表模型估计结果表

Table 5 Estimations of each volatility proxy model of various indices

指数	波动率代表	$\hat{\omega}$	$\hat{\alpha}$	$\hat{\beta}$
CSI 300	$\|y_t\|$	0.067 1	0.054 3	0.882 1
	RV30	0.091 4	0.124 9	0.765 0
	RV15	0.104 3	0.124 6	0.758 6
	RV5	0.069 9	0.126 5	0.774 2
CSI 500	$\|y_t\|$	0.081 9	0.096 3	0.818 8
	RV30	0.121 3	0.189 6	0.675 7
	RV15	0.111 1	0.175 3	0.700 2
	RV5	0.058 9	0.171 5	0.738 9
SSE 50	$\|y_t\|$	0.021 3	0.033 5	0.923 5
	RV30	0.014 6	0.080 7	0.852 9
	RV15	0.018 7	0.079 6	0.851 0
	RV5	0.013 5	0.079 2	0.856 4

以沪深 300 指数为例,选取日间波动率代表 $|y_t|$ 进行建模估计,得到的模型结果为 $y_t=\sigma_t\varepsilon_t,\sigma_t^2=0.067\ 1+0.054\ 3y_{t-1}^2+0.882\ 1\sigma_{t-1}^2$。而选取日内波动率代表,以 RV5 为例,进行建模估计,得到的模型结果为 $y_t=\sigma_t\varepsilon_t,\sigma_t^2=0.069\ 9+0.126\ 5y_{t-1}^2+0.774\ 2\sigma_{t-1}^2$。

对得到的波动率估计值进行可视化,可得到具体的波动率估计曲线,见图 4。

由于基于高频数据的各日间波动率代表之间的结果相近,波动率估计曲线会互相重合,所以图 4 选取基于日间波动率代表 RV5 的估计结果与基于日内波动率代表的结果进行对比。从图 4 可以看出,基于高频数据的波动率代表模型可以捕捉更多波动信息,估计结果更为准确。

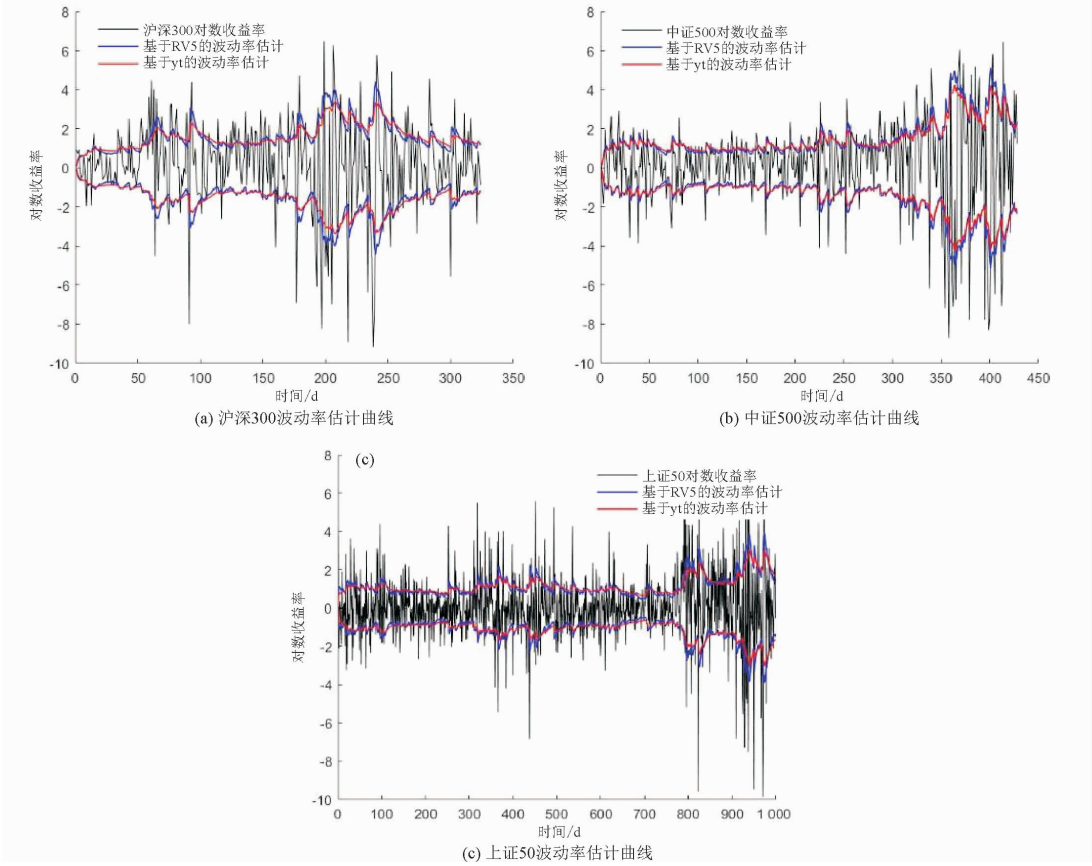


图 4　各指数选取日内波动率和日间波动率代表下波动率估计效果图

Fig. 4　Estimation curves of intraday and daily volatility proxy models for various indices

得到估计结果后,便可计算检验统计量的值,同样取残差自相关函数最大滞后阶数 $m=6$,具体结果见表 6。

表 6　各指数各波动率代表

Table 6　Test statistics results for each index and volatility proxy

指数	$\widehat{Q}_c(\lvert y_t\rvert)$	$\widehat{Q}_c$(RV30)	$\widehat{Q}_c$(RV15)	$\widehat{Q}_c$(RV5)
CSI 300	3.318 9	2.899 9	2.703 4	2.573 1
CSI 500	10.712 8	5.700 6	4.855 6	6.052 6
SSE 50	1.086 0	1.043 5	0.877 1	0.766 1

从表 6 可以看出,若取显著性水平为 0.05,则基于高频数据构建的模型和基于低频数据构建的模型的检验统计量结果都小于临界值 $\chi^2_{0.95}(6)=12.591\ 6$,都接受了原假设,即认为模型选择正确。此外,基于高频数据的检验统计量结果都比基于低频数据的小,也就是说,在估计效果更好的情况下,检验统计量更倾向于接受原假设,这是符合预期的。假设原假设为真,那么估计结果越接近真值,其检验统计量也应当越远离临界值。因为原假设中检验统计量为卡方分布,所以越远离临界值对应越小的检验统计量结果。下面通过对比不同滞后阶数的残差自相关图进行进一步验证,具体见图 5。

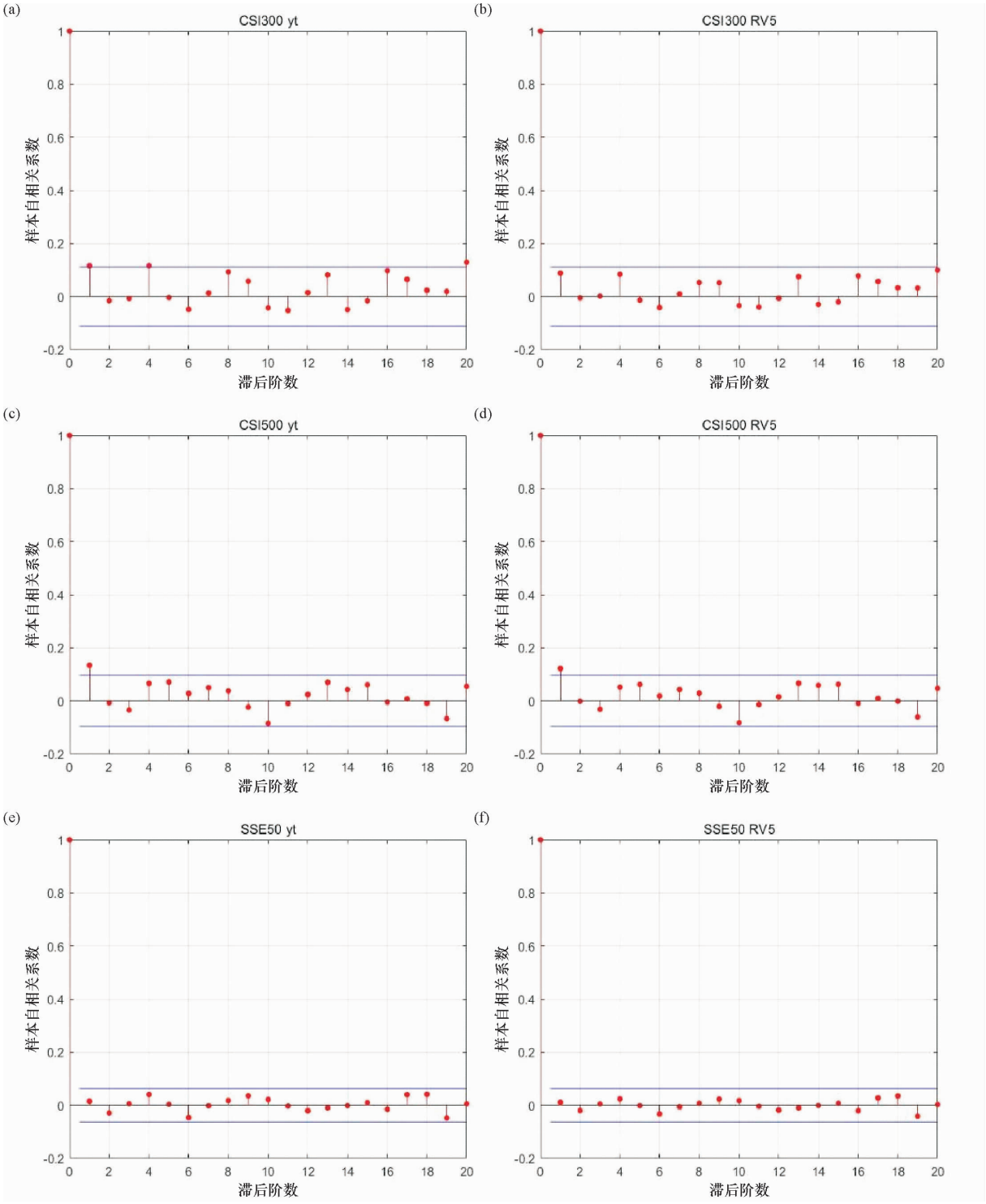


图 5 各指数选取日间波动率和日内波动率代表下的残差自相关图

Fig. 5 Residual autocorrelation plots for each index based on intraday and daily volatility

同样,此处选择基于 RV5 为日间波动率代表的残差自相关函数与日内波动率代表进行对比,从图 5 可以看出,3 个指数使用高频数据构建模型的残差自相关性对比基于低频数据都有所减小,也就是说,基于高频数据的模型确实更好地提取了有用信息,这也预示了基于高频数据的检验统计量结果应当会更小,验证了前面的检验统计量结果。特别地,结合模拟中基于高频数据检验

统计量的检验显著性水平更接近 0.05 这一结论可知,基于高频信息的检验犯第一类错误的概率更小,所以当模型选择正确时,基于高频的检验统计量更小,即检验对接受原假设的倾向性更明显是非常合理的。这也说明所推导的检验统计量是一个有效的模型诊断工具,能反映有用信息的提取效果,具有实际意义。

5　结　语

本文根据基于高频信息的波动率代表 GARCH 模型的 QMELE 估计性质,推导出对应的 portmanteau Q 检验统计量,并通过模拟得到了该检验统计量的显著性水平和功效。从模拟结果可以看出,基于高频信息检验统计量的显著性水平结果优于基于低频信息的检验统计量结果,也就是说,所推导的检验统计量的 0.95 分位数更接近理论的卡方分布 0.95 分位数。另外,从检验功效结果可以看出,当高频残差存在相关性时,对比基于低频信息的检验统计量,基于高频信息检验统计量在识别上更具有优势。而当低频残差存在相关性时,当相关性较大时,基于高频信息的检验统计量也可以识别其相关性,这说明所提出的检验统计量对 GARCH 模型的阶数选择有一定的识别作用。最后,以沪深 300、中证 500 和上证 50 等 3 个指数为例进行实证分析发现,当日内波动率代表和日间波动率代表模型的检验结果都为接受原假设时,基于日间波动率代表模型的检验统计量结果更小。对应基于日间数据的估计效果更好这一结果,表明该检验统计量确实能显示信息的有效提取效果,具有实用性。

参考文献:

[1] Odean T. Volume, volatility, price, and profit when all traders are above average[J]. The Journal of Finance, 1998, 53(6): 1887-1934.

[2] Engle R F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation[J]. Econometrica, 1982, 50(4): 987-1007.

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

[4] Nelson D B. Conditional heteroscedasticity in asset returns: a new approach[J]. Econometrica, 1991, 59(2): 347-370.

[5] Engle R F, Bollerslev T. Modelling the persistence of conditional variances[J]. Econometric Reviews, 1986, 5(1): 1-50.

[6] Zakoïan J M. Threshold heteroskedastic models[J]. Journal of Economic Dynamics and Control, 1994, 18(5): 931-955.

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

[8] Bollerslev T, Wooldridge J M. Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances[J]. Econometric Reviews, 1992, 11(2): 143-172.

[9] Lee S W, Hansen B E. Asymptotic theory for the GARCH (1,1) quasi-maximum likelihood estimator[J]. Econometric Theory, 1994, 10(1): 29-52.

[10] Hall P, Yao Q W. Inference in ARCH and GARCH models with heavy-tailed errors[J]. Econometrica, 2003, 71(1): 285-317.

[11] Peng L, Yao Q W. Least absolute deviations estimation for ARCH and GARCH models[J]. Biometrika, 2003, 90(4): 967-975.

[12] Li G D, Li W K. Least absolute deviation estimation for fractionally integrated autoregressive moving average time series models with conditional heteroscedasticity[J]. Biometrika, 2008, 95(2): 399-414.

[13] 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.

[14] 李莉丽, 张兴发, 邓春亮, 等. 基于高频数据的 GARCH 模型拟极大指数似然估计[J]. 应用数学学报, 2022, 45

(5):652-664.

[15] Box G E P, Pierce D A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models[J]. Journal of the American Statistical Association, 1970, 65(332): 1509-1526.

[16] Ljung G M, Box G E P. On a measure of lack of fit in time series models[J]. Biometrika, 1978, 65(2): 297-303.

[17] McLeod A I, Li W K. Diagnostic checking ARMA time series models using squared-residual autocorrelations[J]. Journal of Time Series Analysis, 1983, 4(4): 269-273.

[18] Li W K, Mak T K. On the squared residual autocorrelations in non-linear time series with conditional heteroskedasticity[J]. Journal of Time Series Analysis, 1994, 15(6): 627-636.

[19] Ling S Q, Li W K. Diagnostic checking of nonlinear multivariate time series with multivariatearch errors[J]. Journal of Time Series Analysis, 1997, 18(5): 447-464.

[20] Li G D, Li W K. Diagnostic checking for time series models with conditional heteroscedasticity estimated by the least absolute deviation approach[J]. Biometrika, 2005, 92(3): 691-701.

[21] Chen M, Zhu K. Sign-based portmanteau test for ARCH-type models with heavy-tailed innovations[J]. Journal of Econometrics, 2015, 189(2): 313-320.

[22] Jiang F Y, Li D, Zhu K. Non-standard inference for augmented double autoregressive models with null volatility coefficients [J]. Journal of Econometrics, 2020, 215(1): 165-183.

[23] Li M Y, Zhang Y F. Bootstrapping multivariate portmanteau tests for vector autoregressive models with weak assumptions on errors[J/OL]. Computational Statistics & Data Analysis, 2022,165:107321.

【责任编辑:卓祯雨】