

高频数据中内生测量时间的存在性

肖鸿民, 叶立

(西北师范大学数学与统计学院, 甘肃 兰州 730070)

[摘要] 基于高频数据估计积分波动率时, 一般假设测量时间和价格过程无关, 但这个假设并不符合实际情况, 本文在考虑噪音污染的影响下, 为一般的受内生测量时间影响的实波动率建立了中心极限定理, 同时用真实的股票数据呈现了这种内生性.

[关键词] 内生性, 高频数据, 实波动率

[中图分类号] O212.2 **[文献标志码]** A **[文章编号]** 1001-4616(2015)04-0036-06

Existence of Endogenous Sampling High Frequency Data

Xiao Hongmin, Ye Li

(College of Mathematics and Statistics, Northwest Normal University, Lanzhou 730070, China)

Abstract: In view of the high frequency data, especially the ultra-high frequency data, when handling with integral volatility, simplifying assumptions are usually imposed on the relationship between the observation times and the price process. They are generally considered to be independent, but this assumption does not conform to the actual situation. With market microstructure noise, this paper establishes a central limit theorem for the realized volatility in a general endogenous time setting, and documents that this endogeneity can be presented in real stock data.

Key words: endogenous, high frequency data, realized volatility

20世纪90年代以前, 人们对金融时间序列的研究都是针对日、周、月、季度或者年度数据进行的, 这种金融数据在金融计量学研究领域通常称为低频数据. 近年来, 随着计算工具和计算方法的发展, 极大地降低了数据记录和存储的成本, 使得对更高频率的金融数据进行研究成为可能. 在金融市场中, 高频率采集的数据可以分为两类: 高频数据(high frequency data)和超高频数据(ultra high frequency data). 高频数据即日内数据, 是指在开盘时间和收盘时间之间进行抽样的交易数据, 主要是以小时、分钟、甚至秒为抽样频率的、按时间顺序排列的时间序列. 超高频数据则是指交易过程中实时采集的数据. 高频数据和超高频数据两者之间的最大区别是: 前者是等时间间隔的, 后者的时间间隔是时变的. 对这些数据进行各种分析、建立模型和相关的研究, 都极大地推动了市场微观结构理论和金融计量学的发展, 从而大大地丰富和推广了金融工程学和金融计量学的研究领域和视角.

一般而言, 金融市场中的信息是连续的影响股票价格的运动过程的, 采用离散模型考察资产的价格行为必然会造成信息的丢失, 数据的采集频率越低, 信息丢失越多; 反之, 数据的采集频率越高, 获取的市场信息也就越多. 因此, 在股票的交易过程中, 记录出来的高频数据和超高频数据包含了更多的实时信息, 因而能更加准确地捕捉到市场发生的微小的变化过程, 所以利用高频数据和超高频数据的特性研究资产价格的相关特征与潜在过程比采用低频数据具有更多的优势.

20世纪90年代以来, 高频数据和超高频数据成为金融市场研究的全新手段, 它们从根本上改变了以往对市场波动性的测量和应用. Bollerslev and Zhao^[1]基于高频数据的特点, 提出了不需要模型的“已实现”波动率作为积分波动率的非参估计量, 该估计方法计算简便且精度较高, 由此也引发了关于高频数据相关特征研究的一个热潮. 但其实质就是充实渐进理论使高频数据的可用性增加. 具体来说, 相关渐进理论

收稿日期: 2014-09-15.

基金项目: 国家自然科学基金(71261023).

通讯联系人: 肖鸿民, 教授, 研究方向: 金融统计与保险数学. E-mail: xiaohm9@126.com

是建立在两个收敛的结果之上,分别为伊藤公式 $dX_t = \Delta_t d_t + \sigma_t dW_t$ [2] 和观测时间 $t_{n,i}, i=0,1,\dots$

首先,如果观测时间 $t_{n,i}$ 是停时,即网络分割 $\max |t_{n,i} - t_{n,i-1}|$ 依概率趋近于 0,已实现波动率 $[X, X]_T = \sum_{t_{n,i} \leq T} (X_{t_{n,i}} - X_{t_{n,i-1}})^2$ 是二次变分 $\langle X, X \rangle_T = \int_0^T \sigma^2 ds$ 的一致点估计. 其次,在某些关于随机次数 $t_{n,i}$ 的假设之下,也就是所谓“二次变分时间”的过程收敛 $\lim_{n \rightarrow \infty} n \sum (t_{n,i} - t_{n,i-1})^2 = H_t$. H_t 是一个适应过程,次数 $t_{n,i}$ 关于 X 过程是独立的.

距离相等的情况可以推广到“次数变化”的情形,这时测量次数导致了某种程度的内生性. 本文关注的焦点是内生的观测时间很重要,这意味着 $n^{\frac{1}{2}}([X, X, X]_T)$ 有非零的极限,主要的理论成果是,给出了内生采样次数存在时的中心极限定理,同时通过对真实股票数据的 3 个实验,用假设检验的方法呈现了高频数据中的内生性.

1 预备知识

在此简略地介绍一下估计. 首先,我们考虑对数价格过程 $X_t, 0 \leq t \leq 1$, 并且此过程满足伊藤过程: $X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s$.

事实上,我们所观察到的数据是受到微观结构噪音影响的数据,即我们得到的是 Y_t 而非 X_t . $Y_t = X_t + \varepsilon_t$, 其中, ε_t 是在 t 时刻的微观结构噪音. 现在就是要利用 Y_t 来估计预测 ISV [3]. 我们人为地划分区间 $[0, 1]$ 为 m 等子区间,令 $K = \left\lfloor \frac{n}{m} \right\rfloor$, 并记: $\tau_r^k = \frac{r}{m} + \frac{k-1}{n}, r=1, \dots, m-1; k=1, \dots, K$. 选取一个 l , 令 $l \rightarrow \infty$, 当 $\frac{l}{n} \rightarrow 0$, $n \rightarrow \infty$ 时,对每个 $r \geq l$, 令 $Y_{\tau_r^k} = \frac{1}{l} \sum_{i=1}^l Y_{\tau_{r-i}^k}$. 区间 $[0, 1]$ 上观测到的数据记为 n 个,其观测值为 $Y_{\tau_i}, i=1, 2, \dots, n$.

令 $[Y, Y]^{\text{all}} = \sum_{r=1}^m \sum_{k=1}^K f(Y_{\tau_r^k}) (Y_{\tau_r^{k+1}} - Y_{\tau_r^k})^2$, 再令 $[Y, Y]^{\text{avg}} = \frac{1}{K} \sum_{r=2}^m \sum_{k=1}^K f(Y_{\tau_r^k}) (Y_{\tau_r^{k+1}} - Y_{\tau_r^k})^2$, 则 ISV 的估计可以定义为: $\hat{ISV} = [Y, Y]^{\text{avg}} - \frac{m-1}{n-K+1} [Y, Y]^{\text{all}}$, 其中 $\frac{1}{2(n-k+1)} [Y, Y]^{\text{all}}$ 是微观结构噪音的估计.

2 引理介绍

引理 1 假设 X_t, U_t, σ_t 对过滤 \mathcal{R}_t 是适应的 [4], Z_n 是一系列 \mathcal{R}_t -可测的随机变量, Z_n 依概率收敛到 z . 当 $n \rightarrow \infty$ 时,如果 z 在 \mathcal{R}_t 的扩张下可测,那么对于所有 $A \in \mathcal{R}_t$ 对所有有界连续 g ,

$$E I_A g(Z_n) \rightarrow E I_A g(Z).$$

引理 2 假设 x_t, μ_t, σ_t^2 对过滤 \mathcal{R}_t 是适应的、可积的并且局部有界的, $\sigma_t^2 \geq c > 0$ 不随机,同样假设 (所有的 $\varepsilon > 0$) $\max |t_{n,i+1} - t_{n,i}| = O_p \left(n^{-\left(\frac{2}{3} + \varepsilon\right)} \right)$ [5]. 同样假设对所有 t , $n[X, X, X, X]_t \xrightarrow{p} \int_0^t u_s ds$, $n^{\frac{1}{2}}[X, X, X]_t \xrightarrow{p} \int_0^t v_s ds$, $u_s, |v_s|$ 是可积的,最后假设 \mathcal{R}_t 是由有限的连续鞅生成 [6], 有以下结果成立:

$$n^{\frac{1}{2}}([X, X]_t - \langle X, X \rangle_t) \rightarrow \frac{2}{3} \int_0^t \frac{v_s}{\sigma_s^2} dX_s + \int_0^t \sqrt{\frac{2}{3} u_s - \frac{4}{9} \frac{v_s^2}{\sigma_s^2}} dB_s,$$

其中 B_s 是 Brown 运动.

3 定理及证明

令 b_s 与 σ_s 适应于滤子 \mathcal{R}_s , 并且局部有界和可积, 我们有以下定理成立:

$$[Y, Y]^{\text{avg}} - \frac{m-1}{n-K+1} [Y, Y]^{\text{all}} \xrightarrow{p} \int_0^t \sigma_s^2 ds, \quad (1)$$

$$n^{\frac{1}{2}} \left([Y, Y, Y]^{\text{avg}} - \frac{m-1}{n-K+1} [Y, Y, Y]^{\text{all}} \right) \xrightarrow{p} \int_0^t u_s ds, \quad (2)$$

$$n \left([Y, Y, Y, Y]^{\text{avg}} - \frac{m-1}{n-K+1} [Y, Y, Y, Y]^{\text{all}} \right) \xrightarrow{p} \int_0^t v_s ds. \quad (3)$$

(1)式的证明

$$\begin{aligned} [Y, Y]^{\text{avg}} &= \frac{1}{K} \sum_{r=2}^m \sum_{k=1}^K (Y_{\tau_r}^{k+1} - Y_{\tau_r}^k)^2 = \frac{1}{K} \sum_{r=2}^m \sum_{k=1}^K f(Y_{\tau_{r-1}}^k) (X_{\tau_{r-1}}^{k+1} - X_{\tau_r}^k)^2 + \frac{2}{K} \sum_{r=2}^m \sum_{k=1}^K f(Y_{\tau_{r-1}}^k) (X_{\tau_r}^{k+1} - X_{\tau_r}^k) (\varepsilon_{\tau_r}^k - \varepsilon_{\tau_{r-1}}^k) + \\ &\quad \frac{1}{K} \sum_{r=2}^m \sum_{k=1}^K f(Y_{\tau_{r-1}}^k) [Y, Y]^{\text{all}} = \sum_{r=2}^m \sum_{k=1}^K (Y_{\tau_{r-1}}^{k+1} - Y_{\tau_r}^k)^2 = \sum_{r=2}^m \sum_{k=1}^K f(Y_{\tau_{r-1}}^k) (X_{\tau_{r-1}}^{k+1} - X_{\tau_{r-1}}^k)^2 + \\ &\quad \sum_{r=2}^m \sum_{k=1}^K f(Y_{\tau_{r-1}}^k) (X_{\tau_{r-1}}^{k+1} - X_{\tau_r}^k) (\varepsilon_{\tau_{r-1}}^{k+1} - \varepsilon_{\tau_r}^k) + \sum_{r=2}^m \sum_{k=1}^K f(\varepsilon_{\tau_{r-1}}^{k+1} - \varepsilon_{\tau_r}^k)^2. \end{aligned}$$

我们有

$$E \left[\frac{1}{K} \sum_{r=2}^m \sum_{k=1}^K f(Y_{\tau_{r-1}}^k) (\varepsilon_{\tau_r}^{k+1} - \varepsilon_{\tau_r}^k)^2 - \frac{m-1}{n-K+1} \sum_{r=2}^m \sum_{k=1}^K f(Y_{\tau_{r-1}}^k) (\varepsilon_{\tau_r}^{k+1} - \varepsilon_{\tau_r}^k) | X \right] = 0,$$

并且

$$\begin{aligned} E \left[\frac{2}{K} \sum_{r=2}^m \sum_{k=1}^K f(Y_{\tau_{r-1}}^k) (X_{\tau_{r-1}}^{k+1} - X_{\tau_r}^k) (\varepsilon_{\tau_{r-1}}^{k+1} - \varepsilon_{\tau_r}^k) \right] &\xrightarrow{p} 0, \\ E \left[\frac{m-1}{n-K+1} \sum_{r=2}^m \sum_{k=1}^K f(Y_{\tau_{r-1}}^k) (X_{\tau_{r-1}}^{k+1} - X_{\tau_r}^k) (\varepsilon_{\tau_{r-1}}^{k+1} - \varepsilon_{\tau_r}^k) \right] &\xrightarrow{p} 0, \end{aligned}$$

则可得到:

$$[Y, Y]^{\text{avg}} - \frac{m-1}{n-K+1} [Y, Y]^{\text{all}} \xrightarrow{p} \int_0^t \sigma_s^2 ds.$$

这样,我们就证明了 ISV^{\wedge} 依概率收敛于 ISV .

(2)式的证明

$$\frac{1}{k} \sum_{r=2}^m \sum_{k=1}^K (X_{\tau_r}^k - X_{\tau_{r-1}}^k + \varepsilon_{\tau_r}^k - \varepsilon_{\tau_{r-1}}^k)^3 = \frac{1}{K} \sum_{r=2}^m \sum_{k=1}^K (X_{\tau_{r-1}}^{k+1} - X_{\tau_r}^k)^3 + \frac{3}{K} \sum_{r=2}^m \sum_{k=1}^K (X_{\tau_r}^k - X_{\tau_{r-1}}^k) (\varepsilon_{\tau_r}^k - \varepsilon_{\tau_{r-1}}^k),$$

其中

$$E \left[\frac{4}{K} \sum_{r=2}^m \sum_{k=1}^K (X_{\tau_{r-1}}^{k+1} - X_{\tau_r}^k)^2 (\varepsilon_{\tau_r}^k - \varepsilon_{\tau_{r-1}}^k) \right] \xrightarrow{p} 0.$$

同理:

$$E \left[\sum_{r=2}^m \sum_{k=1}^K f(Y_{\tau_{r-1}}^k) (X_{\tau_{r-1}}^{k+1} - X_{\tau_r}^k) (\varepsilon_{\tau_{r-1}}^{k+1} - \varepsilon_{\tau_r}^k) \right] \xrightarrow{p} 0.$$

由于 $\varepsilon_{\tau_r}^k$ 独立同分布,

$$\frac{3}{k} \sum_{r=2}^m \sum_{k=1}^K ((X_{\tau_r}^k - X_{\tau_{r-1}}^k) + (\varepsilon_{\tau_r}^k - \varepsilon_{\tau_{r-1}}^k))^2 = \frac{6}{K} \sum_{r=2}^m \sum_{k=1}^K (X_{\tau_{r-1}}^{k+1} - X_{\tau_r}^k) (\varepsilon_{\tau_r}^k)^2.$$

同理:

$$\frac{m-1}{n-k+1} \sum_{r=2}^m \sum_{k=1}^K ((X_{\tau_{r-1}}^{k+1} - X_{\tau_r}^k) (\varepsilon_{\tau_r}^k - \varepsilon_{\tau_{r-1}}^k)^2) = \frac{2(m-1)}{n-k+1} \sum_{r=2}^m \sum_{k=1}^K (X_{\tau_{r-1}}^{k+1} - X_{\tau_r}^k) (\varepsilon_{\tau_r}^k)^2.$$

则:

$$E \left[\frac{6}{K} \sum_{r=2}^m \sum_{k=1}^K (X_{\tau_{r-1}}^{k+1} - X_{\tau_r}^k) \varepsilon_{\tau_r}^k - \frac{6(m-1)}{n-k+1} \sum_{r=2}^m \sum_{k=1}^K (X_{\tau_{r-1}}^{k+1} - X_{\tau_r}^k) (\varepsilon_{\tau_r}^k)^2 \right] \xrightarrow{p} 0,$$

所以有:

$$n^{\frac{1}{2}} [Y, Y, Y]_t \xrightarrow{p} \int_0^t u_s ds.$$

(3)式的证明

$$\begin{aligned} [Y, Y, Y] &= \frac{1}{K} \sum_{r=2}^m \sum_{k=1}^K \left(Y_{\tau_r}^{k+1} - Y_{\tau_{r-1}}^k - \frac{m-1}{n-k+1} \sum_{r=2}^m \sum_{k=1}^K (Y_{\tau_{r-1}}^{k+1} - Y_{\tau_r}^k) \right)^4 - \frac{24(m-1)}{m(n-k+1)} \times \\ &\quad \left[\frac{1}{K} \sum_{r=2}^m \sum_{k=1}^K ((Y_{\tau_r}^{k+1} - Y_{\tau_{r-1}}^k)^2 - \frac{m-1}{n-k+1} \sum_{r=2}^m \sum_{k=1}^K (Y_{\tau_{r-1}}^{k+1} - Y_{\tau_r}^k)^2) \right]. \end{aligned}$$

我们有:

$$E \left[\frac{4}{K} \sum_{r=2}^m \sum_{k=1}^K (X_{\tau_r^k} - X_{\tau_{r-1}^k})^3 (\varepsilon_{\tau_r^k} - \varepsilon_{\tau_{r-1}^k}) \right] \xrightarrow{p} 0.$$

同理:

$$E \left[\sum_{r=2}^m \sum_{k=1}^K (X_{\tau_{r-1}^{k+1}} - X_{\tau_{r-1}^k})^3 (\varepsilon_{\tau_r^{k+1}} - \varepsilon_{\tau_{r-1}^k}) \right] \xrightarrow{p} 0.$$

则:

$$E \left[\frac{12}{K} \sum_{r=2}^m \sum_{k=1}^K (X_{\tau_r^k} - X_{\tau_{r-1}^k})^2 (\varepsilon_{\tau_r^k} - \varepsilon_{\tau_{r-1}^k})^2 \right] \xrightarrow{p} \frac{24(m-1)}{m(n-k+1)} \int_{\tau_0^1}^{\tau_m^k} \sigma_t^2 dt,$$

所以有:

$$n [Y, Y, Y, Y]_t \xrightarrow{p} \int_0^t v_s ds.$$

说明 我们知道,金融数据中的确存在时间内生性,因此,一致的估计引理中提到的偏和方差将会允许我们利用时间的内生性来提高波动率估计的精确性.截止到目前为止,大部分的文章都假定观察的资产价格的时间间隔是等速的,而本文假定资产价格的观测时间是随机的.已知二阶变差的收敛结果在很弱的条件下都是成立的,因此我们可以将观测时间拓展到随机观测时间和有内生性的观测时间,在没有时间和基本半鞅过程独立的假定下,我们得到了随机速度观测下的二阶、三阶、四阶情况下的中心极限定理,这就使得结果极其重要了.

4 实验

本节利用计算机模拟与实际数据对之前的理论结果进行验证说明.

首先,在完全独立的情况下,利用双时间序列的方法对检测量进行估计,得到了估计结果.然后,以深市股票中信证券(SH600030)、弘业股份(SH600128)与沪市南大光电(SZ300346)的日内价格作为高频数据,通过对其检测量进行计算检验,从数值结果来看,实际上股票高频数据并非完全独立地,具有很强的规律性,考虑样本容量较大,而实际数据采用的是即时数据,其数值并不完全相同,大约在2000~3000之间,在此我们统一取为 $n=3000$.而股票价格其真值由ornstein-uhlenbeck^[7]过程刻画,即: $X_t = \int_0^t \cos(s) ds + \int_0^t e^{-2(t-s)} dw_s$, 其中 w_s 是标准布朗运动.微观结构噪音 ε_i 关于 $N(0, \omega)$ ($\omega=0.05$) 独立同分布.图1为在模拟假设下,一模拟数据的样本路径.不论是其波动大小或波动时间都表现出了波动的随机性和独立性^[8].

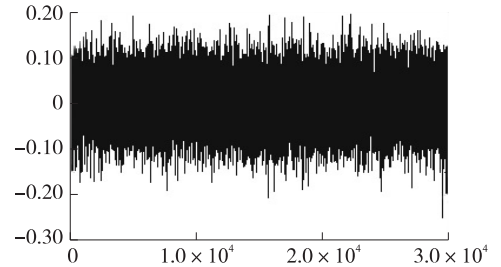


图1 随机模拟的样本路径

Fig.1 The sample path under stochastic simulation

实验1 在零假设 (H_0)^[9]下,过程独立.假设数据被分成 M 大小的 J 块,各自编号为 j .周期内的统计值为:

$$R_j^2 = \frac{\left(\sum_{i=M_{j-1}}^{M_j-1} (\Delta X_{t_i})^3 \right)^2}{\left(\sum_{i=M_{j-1}}^{M_j-1} (\Delta X_{t_i})^2 \right) \left(\sum_{i=M_{j-1}}^{M_j-1} (\Delta X_{t_i})^4 \right)},$$

这时整体检测统计量为:

$$T_1 = \sum_{j=1}^J R_j^2 \Delta T_j.$$

实验2 再次假设数据分成 M 大小的 J 块,编号为 j , $(t_{M_{j-1}}, t_{M_j})$ 区间内,定义

$$A_j = \Delta T_j \cdot \frac{\left(\sqrt{n} \sum_{i=M_{j-1}}^{M_j-1} (\Delta X_{t_i})^3 \right)^2}{\left(\sum_{i=M_{j-1}}^{M_j-1} (\Delta X_{t_i})^2 \right)^3},$$

这时整体检测统计量为:

$$T_2 = \sum_{j=1}^J A_j \Delta T_j.$$

实验3 这里检测统计量为

$$T_3 = \frac{\sum_{j=1}^J \left(\frac{2}{3} n \sum_{i=M_{j-1}}^{M_j} (\Delta X_{t_i})^4 - \frac{4}{9} \frac{\left(\sqrt{n} \sum_{i=M_{j-1}}^{M_j} (\Delta X_{t_i})^3 \right)^2}{\sum_{i=M_{j-1}}^{M_j} (\Delta X_{t_i}^2)} \right)}{2 \sum_{j=1}^J \frac{\left(\sum_{i=M_{j-1}}^{M_j} (\Delta X_{t_i})^2 \right)^2}{(\Delta t_j)^2} - n \sum_{i=M_{j-1}}^{M_j} (\Delta t_i)^2}.$$

我们对 T_1 、 T_2 、 T_3 的值进行了连续模拟,结果如图2、3、4、5、6.

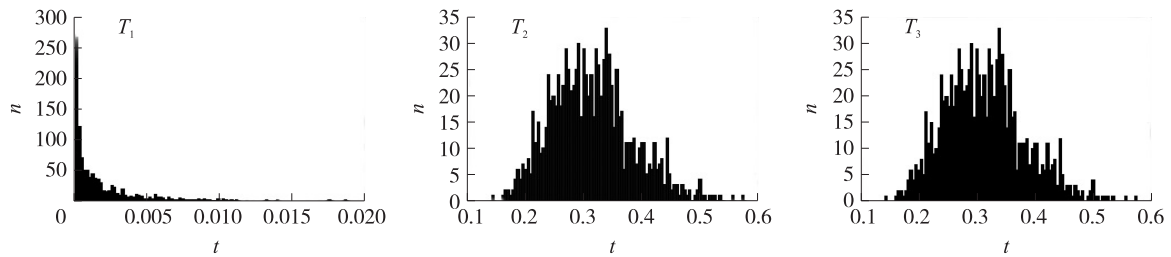


图2 模拟1000次的检测量方图

Fig.2 The bar graph by T test when simulation times is 1000

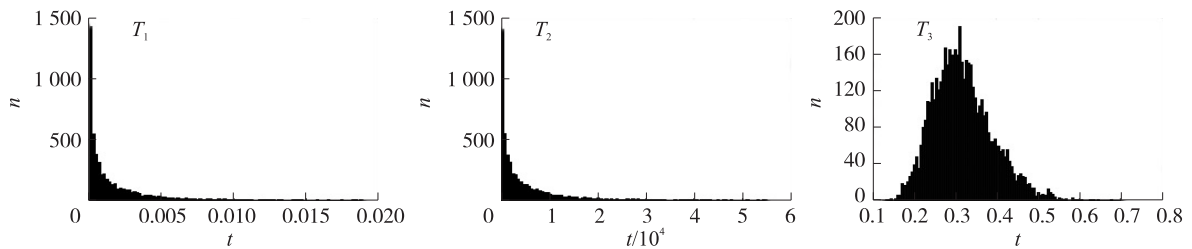


图3 模拟3000次的检测量方图

Fig.3 The bar graph by T test when simulation times is 3000

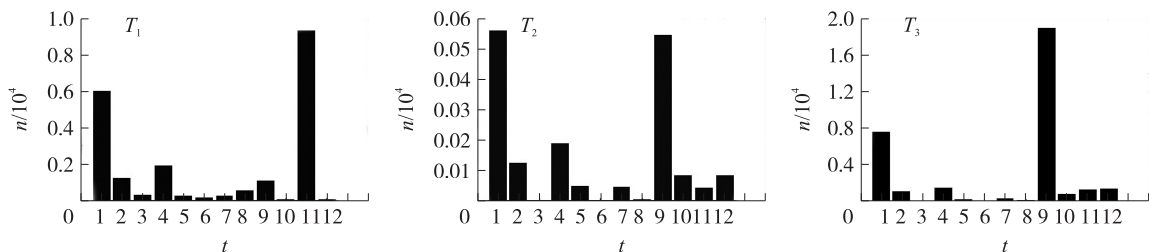


图4 中信证券的检测量方图

Fig.4 The bar graph by T test of ZhongXin company

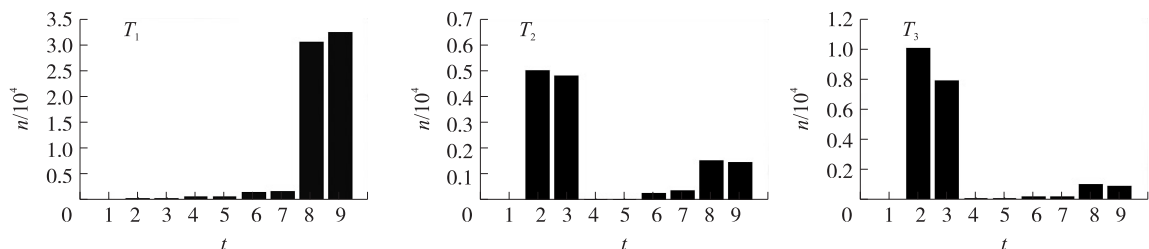


图5 弘业股份的检测量方图

Fig.5 The bar graph by T test of HongYe company

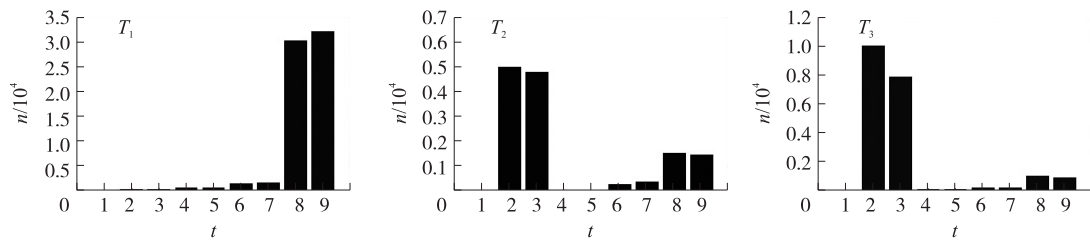


图6 沪市南大光电的检测量方图

Fig.6 The bar graph by T test of NanDaGuangDian company

5 实验分析

通过对比发现,模拟数据 T_1 、 T_2 、 T_3 的值都在 0.02×10^4 以下,甚至大都趋近于 0,而 T_3 更趋近于正态分布,但拥有“后尾”^[10] 的特征,在套用真实数据的时候,这些值平均都在 0.5 以上,甚至更高,与以往研究不同的是,我们对噪音影响进行了处理,通过运用前面得出的中心极限定理,在二阶、三阶、甚至四阶的情况下计算出了检测量的实值,并用直方图清晰地展现了数据的变化过程,所以完全有足够的理由拒绝原假设,即本文通过实验呈现了这种内生性.但是这种影响的大小如何估计和控制^[10],还需要进一步关注.

[参考文献]

- [1] GRIFFIN J E, RCO O. Covariance measurement in the presence of non-synchronous trading and market microstructure noise[J]. Journal of econometrics, 2013, 160(1): 58-68.
- [2] ADIMYA M, BOUZAHIRB H. Existence for a class of partial functional differential equations with infinite delay[J]. Nonlinear analysis, 2011, 46(2): 91-112.
- [3] AIT S, MYKLAND P A, ZHANG L. Ultra high frequency volatility estimation with dependent microstructure noise[J]. Journal of econometrics, 2011, 80(1): 160-175.
- [4] ZHANG L, MYKLAND P A. Edgeworth expansions for realized volatility and related estimators[J]. Annals of statistics, 2013, 64(1): 190-203.
- [5] 韦来生. 数理统计[M]. 北京: 科学出版社, 2010: 65-77.
- [6] 戴国强, 吴林祥. 金融市场微观结构理论[M]. 上海: 上海财经大学出版社, 1999: 88-92.
- [7] 唐勇, 刘峰涛. 金融市场波动测量方法新进展[J]. 华南农业大学学报(自然科学版), 2005, 44(3): 68-72.
- [8] 张尧庭. 金融市场的统计分析[M]. 桂林: 广西师范大学出版社, 2008: 57-69.
- [9] EAN B. Levy processes[M]. Cambridge: Cambridge University Press, 2011: 61-65.
- [10] AASEK K. Contingent claims valuation when the security price is a combination of an ito process and a random point process[J]. Stochastic and applications, 2008, 28(2): 185-200.

[责任编辑: 丁 蓉]