文章编号:1003 - 207(2008)02 - 0007 - 07

中国股市超高频持续期序列长记忆性研究

耿克红,张世英

(天津大学管理学院,天津 300072)

摘 要:针对股市超高频持续期序列,提出了长记忆随机条件持续期模型(LMSCD),并设计了一类基于混沌禁忌遗传算法的谱似然函数模型参数估计方法,通过 Monte Carlo 模拟实验,验证了方法的可行性。然后,利用沪市浦发银行股票的超高频数据,分别建立了交易持续期、价格持续期和交易量持续期的长记忆随机条件持续期模型,验证了中国股票市场超高频持续期序列长记忆性的存在。

关键词:长记忆性;长记忆随机条件持续期模型;混沌禁忌遗传算法;谱似然估计

中图分类号:F830.91 **文献标识码**:A

1 引言

在过去,人们对金融时间序列分析是在等时间间隔的低频数据基础上进行的,比如说以年、月、周,甚至是以日为时间间隔采集的数据为基础。近年来,随着对金融市场微观结构研究的深入,人们对日内金融时间序列数据研究产生的极大的兴趣,日内时间序列数据通常分为两类,一类是高频数据,该类数据是在某交易日内以固定的时间间隔采集的数据,而另一类数据是根据市场事件(比如:发生一次交易,价格变化一个给定的值或交易量变化一个给定的值等)到达的时间逐笔(transaction by transaction)记录下来的数据,2003 年诺贝尔经济学奖获得者 Engle 将此类数据称为超高频(UHF)数据[1],此类数据与传统上的时间序列数据的最大不同是认为市场事件的到达是一个随机过程,因此记录数据的时间间隔也是随机的。

由于超高频时间序列数据时间间隔的随机性,传统的时间序列建模方法显然已经不再适合于超高频时间序列数据,因此人们试图探讨新的计量经济模型来刻画此类数据。正如 Engle 所说的,对超高频时间序列的研究应该建立在对持续期建模的基础上进行[1],近年来,国外学者逐渐展开了对持续期建模的研究,但国内学者在此方面的研究并不多见。Bauwens 和 Veredas(2004)提出了随机条件持续期模型(SCD)[2],本文对 SCD 模型进行了扩展,提出了能刻画超高频数据长记忆性的长记忆随机条件持续期模型(LMSCD),考察了市场事件到达时间间隔序列,即持续期(durations)的长记忆性,并针对在参数估计时传统似然函数优化方法存在的问题,设计出了一类基于混沌禁忌遗传算法的谱似然函数优化方法,对模型的参数进行了估计。利用沪市浦发银行实时交易数据,建立了三类,即交易持续期,价格持续期和交易量持续期的长记忆持续性模型,并进行了微观股市的记忆性分析。

收稿日期:2007 - 02 - 05; **修订日期**:2008 - 03 - 06

基金项目:国家自然科学基金资助项目(70471050)

作者简介:耿克红(1973 -),男(汉族),河南襄城人,天津大学管理学院博士研究生,研究方向:金融计量.

2 长记忆随机条件持续期模型

Bauwens 和 Veredas(2004)提出了 SCD 模型,该模型假设存在一个产生持续期的潜在随机变量。他认为该随机潜在变量能捕捉到金融市场上的随机信息流,这种信息流通常是很难直接观察到的。该模型与 Engle(1998)提出的 ACD 模型最主要的区别在于 SCD 模型是双随机过程,也就是说该模型具有两个随机新息:一个针对持续期本身而言,即随机扰动项,另一个是针对潜在变量,即假设潜在变量是随机的。换句话说,ACD 模型的条件期望持续期在 SCD 模型中变成了随机变量。

如果 τ_t 代表市场事件发生时间,d_t 代表时间 τ_{t-1} 和 τ_t 两次市场事件之间的时间差,即持续期,则

$$d_t = \tau_t - \tau_{t-1} \quad (1)$$

Bauwens 和 Veredas(2004)将 SCD 模型定义为

$$\begin{cases} d_t = H_t\varepsilon_t \\ H_t = \exp(h_t) \\ h_t = \alpha + \beta h_{t-1} + \mu_t \end{cases} \tag{2}$$

这里 H_t 代表潜在随机变量，$\frac{\varepsilon}{I_{t-1}}$ 服从 *i. i. d*，具有正的支撑，I_{t-1} 代表持续期 d_{t-1} 末的信息，Bauwens 和 Veredas(2004) 假设 $\frac{\varepsilon}{I_{t-1}}$ 服从 Gamma(或 Weibull) 分布，即 $\frac{\varepsilon}{I_{t-1}} \sim$ Gamma $(\nu,1)$ (或 Weibull $(r,1)$)，$\nu > 0, r > 0$，当 ν(或 $r = 1$) 时，它服从指数分布。$\frac{\mu_t}{I_{t-1}} \sim iid.\ N(0,\sigma_\mu^2)$，且 $\frac{\varepsilon}{I_{t-1}}$ 和 $\frac{\mu}{I_{t-1}}$ 被认为是独立的。$|\beta| < 1$ 以保证是一个平稳过程，$|\beta|$ 的值越接近于 1 说明具有越高的集聚性和越强的持续性。

模型(2)存在两个不足之处：一是仅考虑了最简单情况下的 SCD 模型形式；二是仅考虑了期望持续期的短期相关，认为模型的自相关函数呈指数模式衰减。但是，Golia (2001) 通过研究表明，发现持续期也存在长记忆性，有必要对持续期建立能刻画长记忆性的模型[3]。所谓长记忆性是指：如果平稳序列 $\{X_t\}$ 的自相关系数((依负幂指数率(双曲率)随间隔阶数(增大而缓慢下降，即：$\rho_\tau \sim C\tau^{2d-1}$，$\tau \to \infty$(其中：$C$ 为常数，d 为长记忆参数，$|d| < 0.5$，d 的绝对值越接近 0.5 说明长记忆性越强，"$\sim$"表示收敛速度相同)，则称 $\{X_t\}$ 为长记忆时间序列[4]。

为了描述时间序列中的长记忆成分，Granger 和 Joyeux(1980)[5] 以及 Hosking(1981)[6] 独立的提出了分整 ARMA 过程，即 ARFIMA 过程。ARFIMA 过程表示为

$$(1-B)^d\phi(B)h_t = \theta(B)\mu_t \quad \mu \sim iid.\ N(0,\sigma_\mu^2) \tag{3}$$

其中 $\phi(B) = 1 - \sum_j^p \phi_j B^j$，$\theta(B) = 1 + \sum_j^q \theta_j B^j$，两者均为根在单位园以外的滞后多项式。

$(1-B)^d$ 是分数差分算子，将它按二项式展开，可得到展开式：

$$(1-B)^d = \sum_k^\infty \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)} B^k = 1 - \sum_k^\infty c_k(d) B^k \tag{4}$$

其中 $c_k(d) = \frac{1}{k!}\prod_1^k (i-1-d)$。通过构建，对于 $d(|d| < 0.5)$ 的任何值，$\sum_k^\infty c_k(d) = 1$，$\Gamma(.)$ 代表 Gamma 函数。

在本文中为了刻画 SCD 模型(3)中潜在随机变量的长记忆性，我们提出了一种长记忆随机条件持续期(LMSCD)模型，LMSCD 模型就是把 ARFIMA 过程纳入上述的 SCD 模型的框架. 那么，本文所提出的 LMSCD(p，d，q)模型的结构如下：

$$\begin{cases} d_t = H_t\varepsilon_t \\ H_t = \exp(h_t) \\ (1-B)^d\phi(B)h_t = \theta(B)\mu_t \end{cases} \tag{5}$$

同样，这里 $\frac{\varepsilon}{I_{t-1}}$ 服从 *i. i. d.*，具有正的支撑，I_{t-1} 代表持续期 d_{t-1} 末的信息。可以假设 $\frac{\varepsilon}{I_{t-1}}$ 服从不同的分布，这样得到不同分布情况下的 LMSCD 模型，本文采用 Bauwens 和 Veredas(2004) 的经验，假设 $\frac{\varepsilon}{I_{t-1}}$ 服从 Gamma(或 Weibull) 分布。$\frac{\mu_t}{I_{t-1}} \sim iid.\ N(0,\sigma_\mu^2)$，且 $\frac{\varepsilon}{I_{t-1}}$ 和 $\frac{\mu}{I_{t-1}}$ 被认为是独立的。

一阶自回归长记忆随机持续期模型是较为简单且常用的 LMSCD(1，d，0)模型，其模型形式如下：

$$\begin{cases} d_t = H_t\varepsilon_t \\ H_t = \exp(h_t) \\ h_t = \phi h_{t-1} + (1-B)^{-d}\mu_t \end{cases} \tag{6}$$

$|\phi| < 1$ 以保证是一个平稳过程，$|\phi|$ 的值越接近于 1 说明具有越高的集聚性和越强的持续性。

3 基于混沌禁忌遗传算法的谱极大似然模型参数估计

3.1 模型的变换

为了方便参数估计，可以把上述 LMACD 模型(5)转换为线性形式。令 $x_t = \ln d_t$，则 LMSCD 模型变为如下形式。

$$\begin{cases} x_t = u + h_t + \xi_t \\ (1-B)^d\phi(B)h_t = \theta(B)\mu_t \end{cases} \tag{7}$$

这里 $\xi_t = \ln\varepsilon_t - u$，$u = E[\ln\varepsilon_i]$，通过该变换就可以使误差项 ξ_t 为零均值的随机变量，$\xi_t \sim iid(0,\sigma_\xi^2)$。如果假设 ε 服从 $W(\gamma,1)$，那么 $\ln\varepsilon$ 的概率密度函数为 $f((\ln\varepsilon)) = \gamma e^{(\ln\varepsilon)\gamma}e^{-(\ln\varepsilon)\gamma}$，分布的均值为 $-0.57722/\gamma$，方差为 $\pi^2/6\gamma^2$；如果假设 ε 服从 $G(\nu,1)$，那么 $\ln\varepsilon$ 的概率密度函数为 $f((\ln\varepsilon)) = e^{(\ln\varepsilon)\nu}e^{-(\ln\varepsilon)}/\Gamma(\nu)$，分布的均值为 digamma 函数 $\psi(\nu)$，方差为 trigamma 函数 $\psi(\nu)$。

3.2　模型参数的估计方法

3.2.1　谱似然函数

根据 hosking(1981)[6],ARFIMA 模型的谱密度为

$$f(\omega) = \frac{\sigma_u^2}{2\pi}\left|\frac{\theta(e^{-i\omega})}{\phi(e^{-i\omega})}\right|^2 |1-e^{-i\omega}|^{-2d} \tag{8}$$

假设 $\xi_t = (\ln\varepsilon_t - u)$ 服从正态分布(ξ 实际上是服从对数 Weibull 分布(或对数 GAMMA 分布),所以有时也把谱似然估计称为频域伪极大似然估计),那么 ξ_t 的谱密度函数为 $\sigma_\xi^2/2\pi$。则 LMSCD 模型的谱密度函数为

$$f_\delta(\omega) = \frac{\sigma_u^2}{2\pi}\left|\frac{\theta(e^{-i\omega})}{\phi(e^{-i\omega})}\right|^2 |1-e^{-i\omega}|^{-2d} + \frac{\sigma_\xi^2}{2\pi} \tag{9}$$

这里 $\sigma_\xi^2 = \begin{cases} \dfrac{\pi^2}{6} & \text{当 } \varepsilon \sim W(\gamma,1) \text{ 时} \\ \psi(\nu) & \text{当 } \varepsilon \sim G(\nu,1) \text{ 时} \end{cases}$

那么,对数谱似然函数为

$$L_n(\delta) = -2\pi n^{-1}\sum_{k=1}^{[n/2]}\left\{\log f_\delta(\omega_k) + \frac{I_n(\omega_k)}{f_\delta(\omega_k)}\right\} \tag{10}$$

这里 $\delta = (d,\sigma_u^2,\sigma_\xi^2,\phi_1,\cdots,\phi_p,\theta_1,\cdots\theta_q)$,[.]表示数值取整,$p,q$ 分别为时间序列的滞后阶数,$\omega_k = 2\pi k n^{-1}$ 是第 k 阶的傅立叶频率,n 是最大时间间隔。

$$I_n(\omega) = \frac{1}{2\pi n}\left(\sum_{t=1}^{n} x_t\cos\omega_k t\right)^2 + \frac{1}{2\pi n}\left(\sum_{t=1}^{n} x_t\sin\omega_k t\right)^2$$

是第 k 个正规化周期图。

对上述的对数谱似然函数最优化,就可以得到模型的参数

$\delta = (d,\sigma_u^2,\gamma,\phi_1,\cdots,\phi_p,\theta_1,\cdots\theta_q)$　当 $\varepsilon \sim W(\gamma,1)$

$\delta = (d,\sigma_u^2,\nu,\phi_1,\cdots,\phi_p,\theta_1,\cdots\theta_q)$　当 $\varepsilon \sim G(\nu,1)$

对于对数谱似然函数的最大化一般可使用梯度信息进行优化,如最优下降法,牛顿法,共扼梯度法,变尺度法,但这类方法在搜索中往往存在振荡,容易陷入局部最优点,且结果的好坏对起始点的选择依赖性较大。鉴于此,本文设计出了一类基于混沌 - 禁忌遗传算法的谱似然函数优化方法。

3.2.2　混沌禁忌遗传算法

遗传算法是建立在自然选择机理基础上的随机、迭代和进化,具有广泛适用性的搜索方法。遗传算法本身具有并行性,不容易陷入局部最优解,能以概率收敛到全局最优解,且能较好地处理大空间搜索的问题,在搜索过程中,基本不利用搜索空间的知识,也不利用目标函数的梯度信息,而仅用适应值函数来评估个体,这样就避免了梯度算法中由于搜索的震荡或梯度的不存在而使似然函数值陷入局部最优解[9]。但由于算法结构、算法复杂度和编码长度的限制,其局部搜索速度和精度并不能得到很好的保证,存在"早熟"现象。为了避免早熟现象,提高算法的爬山能力,可以在遗传算子的设计时引入禁忌搜索技术[10]。所谓禁忌搜索技术是指为了避免局部邻域搜索局部最优的不足,用一个禁忌表记录已经达到过的局部最优点,在下一次搜索中,利用禁忌表中的信息不再或有选择地搜索这些点,以此来跳出局部最优点,这样大大提高了搜索速度。将禁忌搜索技术引入遗传算法的思路是:把禁忌技术引入到遗传算法的进化搜索过程之中,构造了新的交叉算子,即禁忌交叉算子;针对遗传算法爬山能力比较差的不足,根据禁忌技术来改造遗传算法的变异算子,即禁忌变异算子[11]。

种群初始化是遗传算法的重要步骤,其结果直接影响遗传算法的收敛速度,传统遗传算法一般采用随机的方法产生初始种群。由于混沌作为自然界非常广泛存在现象,它看似随机,却隐含着精致的内在结构,具有遍历性、随机性和对初始条件的极度敏感性,能在一定范围内按其自身规律不重复地遍历所有状态,在局部寻优领域具有极为优越的性能[12]。将禁忌遗传算法和混沌优化的结合,可以使禁忌遗传算法的全局寻优能力,搜索精度,搜索速度等几方面得到较为明显的改进,可以将此优化算法用于对本文的对数谱似然函数(10)进行优化。本文方法采用混沌映射 $z_{n+1} = \sin(2/z_n)$ 迭代方程产生初始种群。

3.2.3　混沌禁忌遗传算法的设计

混沌禁忌遗传算法的设计包括:

编码:遗传算法的编码通常采取实值编码和二进制编码两种方法,为了避免采用二进制位串编码导致染色体过长,计算精度降低的问题,可以采用实值编码。可将染色体 A 表示为:$A = m_1 m_2 \cdots m_K$。其中:$k = 2 + p + q$,是模型中参数个数,$m_k \in [k_{\min}, k_{\max}]$,$[k_{\min}, k_{\max}]$ 表示参数的取值区间,根据文献[4],参数 $d,\sigma_u^2 \in [0,0.5]$; $\phi_i,\theta_j \in [0,1]$; $i = 1,\cdots,p$; $j = 1,\cdots,q$。对于参数 γ 取值区间的选取(这里假设 $\varepsilon \sim W(\gamma,1)$,若假设 $\varepsilon \sim G(\nu,1)$,则 ν 参数区间的选取方法也依此进行),为了降低搜索时间,首先选取较大的取值区间,本文通过实验选定的区间为[0,10],然后将该区间等分为若干个子区间,将这若干个子区间分别与其他参数的区间组合进行混沌禁忌遗传优化运算,在这些区间内选取使谱似然函

数值最大的 γ 为所要求的参数值。

群体规模:设群体规模为 M,则第 i 代群体为:

$$P(i) = \{A(1,i), A(2,i), \cdots, A(M,i)\}$$

群体规模影响禁忌遗传算法优化的最终结果以及遗传算法的执行效率:当群体规模太小时,禁忌遗传算法的性能不会太好;群体规模越大,群体中个体的多样性越高,算法陷入局部最优解的危险性就越小,但计算量会增大。本文选择群体规模 $M = 30$。

混沌种群初始化:初始种群选取时采取文献[13]提供的一类在有限区域范围内折叠次数无限的一维迭代混沌自映射:

$$z_{n+1} = \sin(2/z_n) \qquad n = 0,1,2,\cdots\cdots \tag{11}$$

式中,$z \in [-1,1]$,且 $z_n \neq 0$。对式(11)中的 z_n 分别赋予 l 个初值 $z_{0,i}$, $z_{0,i} \neq 0, i = 1,2,3\cdots$,产生 l 个不同轨迹的混沌变量 $\{z_{n,i}, i = 1,2,\cdots\}$[12],将 l 个混沌变量按式(12)作相应的线性变换,分别载波到优化变量的取值范围使其变为解空间内的混沌变量

$$y_{n,i} = \begin{cases} round\left(\dfrac{b_i + a_i}{2} + \dfrac{b_i - a_i}{2} z_{n,i}\right) & i = 1,2,\cdots,l_1 \\ \dfrac{b_i + a_i}{2} + \dfrac{b_i - a_i}{2} z_{n,i} & i = l_1 + 1, l_2 + 2, \cdots, l \end{cases} \tag{12}$$

式中,round 为 Matlab 环境下的舍入取整函数。对于某个固定的 n,便得到了原问题解空间中一个可行解。为使混沌变量在解空间内充分遍历,混沌序列点数应取得足够大,本文通过实验,选取 4000 作为混沌序列点数。对每个可行解分别计算其适应值,选择前 30 个适应值较大的可行解进行混合编码组成初始种群。

适应值函数:混沌禁忌遗传算法在进化搜索中基本上不用外部信息,仅以目标函数,也就是适应值函数为依据。在一般基于梯度信息的优化算法中要求目标函数连续可微,但遗传算法不受此约束。由于本问题就是优化对数谱似然函数(10),所以直接把(10)作为适应值函数。

选择操作算子:选择的目的是把较优的个体直接遗传到下一代,或通过配对交叉再遗传给下一代。选择操作建立在对群体中个体适应值的评估基础上,目前常用的选择策略有:适应值比例方法,最佳个体保存方法,期望值方法等。本文采用正规化几何选择方法,它对每代中的个体赋予正规化的几何分布,而后依分布概率选择。

禁忌交叉算子:交叉是指把两个父代个体的部分结构加以替换重组而生成新个体的操作。禁忌遗传算法在交叉操作中使用禁忌交叉算子。在具体的操作中,使用概率交叉,并设定一张不断更新的禁忌表,只有交叉操作得到的个体在禁忌表以外,那么这次交叉才操作是有效的,否则重新交叉。

禁忌变异算子:变异算子对群体中个体串的某些基因值进行变动。禁忌遗传算法在变异操作中使用禁忌变异算子。在具体的操作中,使每个点产生变异的概率相等,并设定一张不断更新的禁忌表,只有变异操作得到的个体在禁忌表以外,那么这次变异操作才是有效的,否则重新交叉。

停止准则:本文以最大试验代数作为停止准则,本文通过试验,发现 200 代之内基本可以得到一个满意的解。

4 Monte Carlo 模拟实验

为了检验基于混沌禁忌遗传算法的谱似然函数优化的参数估计方法的可靠性,本文针对较为简单的模型(6)进行了 Monte Carlo 模拟实验。

首先我们根据不同的分布假设,设计了两类参数:

$\varepsilon \sim W(\gamma,1)$: $(\phi, d, \sigma_\mu^2, \nu) = (0.900, 0.450, 0.004, 0.900)$

$\varepsilon \sim G(\nu,1)$: $(\phi, d, \sigma_\mu^2, \nu) = (0.900, 0.450, 0.004, 0.900)$

然后针对两种不同假设,分别产生样本个数分别为 6000 个和 60000 的两组数据。

最后,针对每组样本分别利用前文给出的基于混沌禁忌遗传算法的谱似然估计方法来估计式(6)的参数,估计结果见表 1。从表 1 可以看出,估计值和真实值相差很小,且随着样本数据的增加,估计值与真实值之间的差距变小,渐进标准误差也随着样本容量的增加而减小。因此,在实证研究中我们应该尽可能多的获取样本数据,增大样本容量。

表 1 Monte Carlo 模拟实验及估计结果

	真实值	N = 6000		N = 60000	
		估计值	S. E.	估计值	S. E.
Weibull 分布					
ϕ	0.900	0.9045	0.092	0.9033	0.087
d	0.450	0.4562	0.007	0.4534	0.004
σ_μ^2	0.040	0.0406	0.053	0.0402	0.027
γ	0.900	0.9499	0.084	0.9328	0.069
Gamma 分布					
ϕ	0.900	0.9065	0.066	0.9012	0.060
d	0.450	0.4691	0.009	0.4552	0.005
σ_μ^2	0.040	0.0411	0.071	0.0406	0.049
ν	0.900	0.9309	0.061	0.9237	0.055

注:S. E. 为渐进标准误差(HCSE),它是标准误差的无偏估计量。各个参数在 5 %显著性水平下显著

5　实证研究

5.1　样本的选取

本文所采用的超高频数据是在上海证券交易所交易的股票浦发银行的 2005 年 7 月 4 日～2005 年 9 月 31 日共 65 个交易日的实时交易的数据,包括交易时间(以秒为单位),竞 - 叫价格和交易量等,在这期间共有 68185 笔交易,平均每天有 1049 笔交易,可见该股票交易活跃,具有代表性。上海证券交易所每天上午 9:15～9:25 是集合竞价时间,上午 9:30～11:30 和下午 13:00～15:00 是连续竞价时间,研究中剔除集合竞价的交易,而且为了避免开盘和收盘的影响,研究中还要剔除每天上午 9:30～9:50 和每天下午 14:40 - 15:00 的交易数据,其他在连续竞价时间以外的交易数据都要剔除。

Luc Bauwens 和 Giot(2001)指出不同市场事件的持续期反映了市场微观结构的不同方面[14],因此可以分别定义几种不同的持续期:1、交易持续期,所谓交易持续期是指,发生一笔交易的时间间隔;2、交易量持续期,所谓交易量持续期是指,发生的交易量不小于某一给定的值时的时间间隔,本文选取的给定的交易量为 5000 手;3、价格持续期,所谓价格持续期是指价格发生变化不小于某一给定的值时的时间间隔,本文选取的给定的价格变化为 0.01 元。为了避免竞 - 叫价跃动的影响,可以选取竞价和要价的均值。根据定义,本文得到的三类持续期见表 2,从上表的 Ljung - Box Q 统计量可以初步看出,三类持续期都具有比较强的集聚性,即短的持续期后面往往跟随着短的持续期,长的持续期后面也往往跟随着长的持续期。

表 2　三类持续期序列数据基本统计

	交易持续期	交易量持续期	价格持续期
样本数(N)	59468	50549	32375
最大值(秒)	784	784	893
最小值(秒)	1	1	1
平均值(秒)	12.31	14.49	22.42
中位数	12	12	13
标准差	11.12	14.562	28.157
Ljung-Box Q 统计量	64127	58869	14204.50
Ljung-Box Q 统计量临界值:24.996			

5.2　剔除样本数据的“日历效应”

由于持续期均存在明显的“日历效应”,在超高频时间序列建模分析之前,首先必须消除“日历效应”[1]。“日历效应”本质上是时间的函数,所以可以采用以时间为自变量的线性样条函数来刻画它。将对数持续期($\ln D_i$)对线性样条函数 $c_0 + \sum_{j=1}^{K} c_j \cdot I_j \cdot (\tau_i - k_j)$ 进行回归[15],其中 D_i 是未剔除“日历效应”的持续期,k_j 是样条函数的结点,本文中在每天中选择 7 个结点,即 $K = 7$,I_i 是虚拟变量,当 $t_i \geq k_j$ 时,$I_i = 1$,否则 $I_i = 0$。得到参数的估计值之后,代入样条函数 $c_0 + \sum_{j=1}^{K} c_j \cdot I_j \cdot (t_i - k_j)$ 中,就得到 D_i 随时间周期性变化的“日历效应”部分

$$\ln D_t = \hat{c}_0 + \sum_{j=1}^{K} \hat{c}_j \cdot I_j \cdot (t_i - k_j)$$

那么,调整后的持续期为

$$d_i = D_i / \exp(\hat{c}_0 + \sum_{j=1}^{K} \hat{c}_j \cdot I_j \cdot (t_i - k_j))$$

利用上述方法,分别对三类持续期剔除“日历效应”,剔除“日历效应”后持续期序列及其自相关函数图见图 1 所示。从最左边一列图中可以看出,剔除“日历效应”的三类持续期时间序列均具有很强的集聚性。调整后,价格持续期,交易持续期,交易量持续期的 Ljung-Box Q(15)统计量分别为 14092,67052,58451,可以初步得出,调整后的持续期虽然集聚性有所降低,即“日历效应”对序列的集聚性有影响,但影响不大,调整后的持续期仍然具有很强的集聚性,远远大于 Ljung-Box Q(15)统计量的临界值 24.996。且所有序列的自相关函数图呈缓慢衰减的趋势,这说明三类持续期均具有长记忆性。通过后面的建模,更能清晰的看出三类持续期长记忆性的强弱。另外从调整和未调整持续期的自相关函数图比较来看,“日历效应”剔除与否对序列长记忆性并没明显影响。

5.3　模型参数估计的结果

利用上节得到的三类经过调整后持续期数据,假设随机扰动项服从 Weibull 分布或 Gamma 分布,分别利用基于混沌禁忌遗传算法的谱似然函数估计方法估计模型(5)的参数。根据 AIC 准则,经过反复试验发现,LMSCD(1,d,0)模型可以很好的拟合原始数据,利用前述的参数估计方法得到的 LMSCD(1,d,0)模型估计结果见表 3。

从表中可以看出,三类持续期具有不同的模型方程式;各持续期的长记忆性参数 d 均接近于 0.5,这说明三类持续期具有很强的长记忆性(其中,交易持续期的长记忆性最强,交易量持续期的长记忆性次之,价格持续期的长记忆性较低),说明了在对超高频持续期建模时有必要考虑长记忆性对持续期的影响;三类持续期的持续性参数(均接近但不等于 1,这说明三类持续期均具有很强的集聚性,即短的

持续期后面往往跟随着短的持续期,长的持续期后面也往往跟随着长的持续期,其中交易量持续期的集聚性最强,交易持续期的集聚性次之,价格持续期的集聚性较低;三类持续期的潜在随机变量的误差项的方差虽然接近于 0,但在 5%显著水平下不等于 0,这证明了建模时考虑随机潜在变量对条件持续期的影响是很有必要的;另外,我们知道当(ν) = 1 时,随机扰动项服从指数分布,但这里我们发现三类持续期假设服从 Weibull(Gamma)分布时(ν)的值都明显不等于 1,且远远大于 1,这说明在建立模型时不能简单假设随机扰动项服从指数分布。当然,我们可以假设随机扰动项服从包容性更广的 Bull 分布(或广义 Gamma 分布)等。

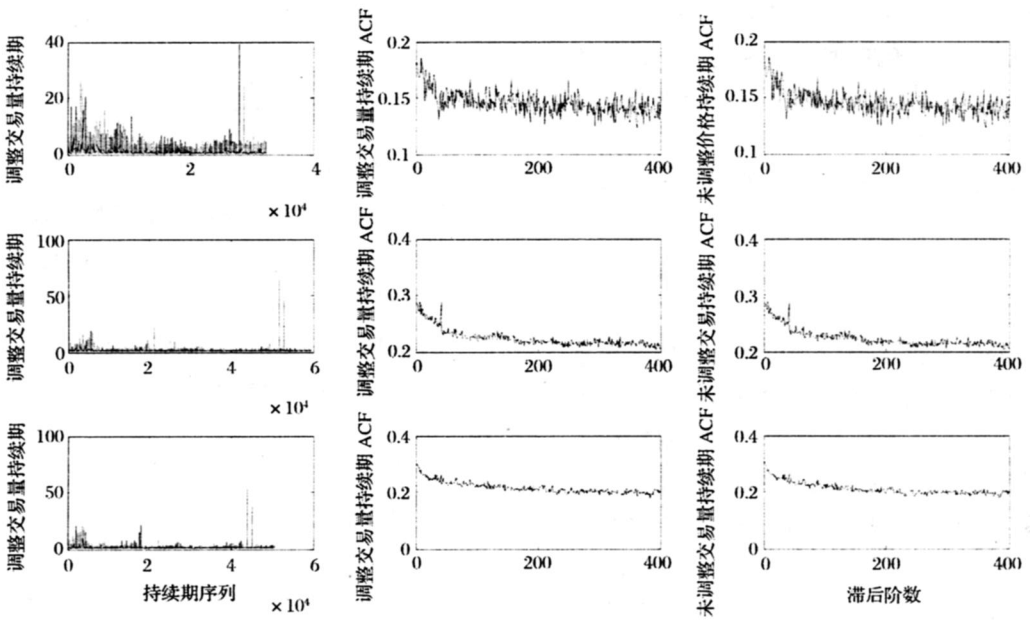

图 1 剔除"日历效应"的三类持续期序列及其自相关函数

表 3 三类持续期序列参数估计结果

	三类持续期					
	交易持续期		价格持续期		交易量持续期	
	参数估计值	S. E.	参数估计值	S. E.	参数估计值	S. E.
Weibull 分布						
ϕ	0.9603	0.0002	0.8602	0.0029	0.9592	0.0288
d	0.4996	0.0015	0.4765	0.0276	0.4932	0.0051
σ_η^2	4.2827e - 4	0.0003	5.3579e - 4	0.0074	5.5074e - 5	0.0016
γ	2.9129	0.0056	1.7733	0.0055	2.517	0.0103
Gamma 分布						
ϕ	0.9511	0.0008	0.8313	0.0044	0.9340	0.0039
d	0.4990	0.0030	0.4904	0.0009	0.4921	0.0281
σ_η^2	3.9806e - 6	0.0046	6.0603e - 4	0.0102	1.1532e - 4	0.0151
ν	5.6033	0.0038	2.3747	0.0290	4.3520	0.0020

注:S. E. 为渐进标准误差(HCSE),它是标准误差的无偏估计量。各个参数在 5%显著性水平下显著

6 结束语

本文扩展了 Bauwens 和 Veredas(2004)的 SCD 模型,考虑了长记忆性对持续期的影响,提出了长记忆随机条件持续期模型(LMSCD),并设计出了一类基于混沌禁忌遗传算法的谱似然函数优化的参数估计方法,通过 Monte Carlo 模拟实验,验证了参数估计方法的可行性。并用本文提出的模型结合中国的股票市场,选取沪市浦发银行这只股票的超高频数据进行了实证研究,证明了超高频数据下股票市场存在着长记忆性,因此在利用超高频数据进行预测和研究金融市场微观结构时不能忽视长记忆性的影响。当然,本文 LMSCD 模型的提出仅仅是个初步,前面指出超高频数据建模的目的主要是为微观结构实证研究提供服务,因此有必要通过考虑微观结构变量(比如交易强度,平均交易量,等)对 LMSCD 模

型的影响来研究金融市场的微观结构。

参考文献:

[1] Engle R. F., Russell J. R.. Autoregressive conditional durations: a new approach for irregularly spaced transaction data [J]. Econometrica, 1998, 66: 1127 - 1163.

[2] Bauwens L., Veredas D.. The stochastic conditional duration model: a latent variable model for the analysis of financial durations [J]. Journal of econometrics, 2004, 119:381 - 412.

[3] Golia S. .Long memory effects in ultra - high frequency data [J]. Quaderni di Statistica, 2001, 3:43 - 52.

[4] Brockwell P. J., R. A. Davis. Time series: theory and methods [M]. New York: Springer - Verlag, 1991,327 - 331.

[5] Granger C. W. J., Joyeux R.. An introduction to long memory time series and fractional differencing [J]. Journal of time series analysis, 1980, 1:1 - 29.

[6] Hosking J. R. M.. Fractional differencing [J]. Biometrika, 1981, 68: 165 - 176.

[7] Johnson, N.L., Kotz, S., Balakrishnan N. Distributions In Statistics: Continuous Univariate Distributions [M]. Wiley, New York,1994: 383.

[8] Johnson, N.L., Kotz, S., Balakrishnan N. Distributions In Statistics: Continuous Univariate Distributions [M]. Wiley, New York, 1995.

[9] 周明,孙树栋. 遗传算法原理及应用[M]. 北京:国防工业出版社,2002:126 - 128.

[10] 黄继鸿,雷战波,李欣苗. 基于禁忌遗传算法的案例检索策略[J]. 系统工程理论方法应用,2004,13(1):10 - 13.

[11] 张世英,樊智. 协整理论与波动模型 ——金融时间序列分析及应用[M]. 北京:清华大学出版社,2004.

[12] 姚俊峰,梅炽,彭小奇,等. 混沌遗传算法及其应用[J]. 系统工程,2001,19(1):70 - 74.

[13] 尤勇,王孙安,胜万兴. 新型混沌优化方法的研究及应用[J]. 西安交通大学学报,2003,37(1):69 - 23.

[14] Bauwens L., Giot P.. Econometric modeling of stock market intraday activity [M]. Dordrecht: Kluwer Academic Publishers, 2001,46 - 51.

[15] Ghysels E., Gourierous C., Jasiak J.. Stochastic volatility duration models [J]. Journal of Econometrics, 2004, 119: 413 - 433.

The Long Memory for Ultra-High Frequency Durations Series of Chinese Stock Markets

GENG Ke-hong, ZHANG Shi-ying

(School of Management, Tianjin University, Tianjin 300072, China)

Abstract: This paper puts forward a long memory stochastic conditional durations (LMSCD) model for ultra-high frequency (UHF) durations series, and designs a kind of spectrum likelihood estimation method based on chaos-tabu genetic algorithm. Through Monte Carlo simulation experiments, we prove the feasibility of estimation method. Thereafter, making use of the ultra high frequency data in Shanghai stock market, we construct three different LMSCD models, which are for trade durations LMSCD, price durations LMSCD and volume durations LMSCD, respectively. We testify the existence of long memory in ultra-high frequency durations series of Chinese stock market.

Key words: long memory; LMSCD; chaos-tabu genetic algorithm; spectrum likelihood estimation