doi: 10.7690/bgzdh.2024.12.009

基于数据挖掘的金融时序数据分析

李慧玲
（国网河北省电力有限公司信息通信分公司，石家庄 050020）

摘要：为提高金融时序数据分析评估及预测性能，在研究数据挖掘、极大似然估计、序贯参数学习等模型基础上，设计一种序贯贝叶斯学习方法来估计非对称广义自回归条件异方差（autoregressive moving average，GARCH）模型。考虑杠杆效应，描述收益率和波动率之间的负相关关系，从而解决股票模拟模型估计中的复杂数值问题。通过仿真分析，结果表明：该模型能较好地模拟股票波动及价格趋势，具备有效性。

关键词：金融大数据；数据挖掘；时序数据分析；序贯贝叶斯；股票模拟
中图分类号：TP393　　**文献标志码**：A

Analysis of Financial Time Series Data Based on Data Mining

Li Huiling
(*Information Communication Branch*, *State Grid Hebei Electric Power Co*., *Ltd*., *Shijiazhuang* 050020, *China*)

Abstract: In order to improve the performance of financial time series data analysis, evaluation and prediction, a sequential Bayesian learning method is designed to estimate the asymmetric generalized autoregressive moving average (GARCH) model based on the study of data mining, maximum likelihood estimation and sequential parameter learning. The leverage effect is considered to describe the negative correlation between return and volatility, thus solving the complex numerical problems in the estimation of stock simulation models. Through the simulation analysis, the results show that the model can better simulate the stock volatility and price trends, and is effective.

Keywords: financial big data; data mining; time series data analysis; sequential Bayesian; stock simulation

0　引言

随着计算机、通信技术不断发展，金融领域时间序列数据规模越来越大，金融数据[1-3]已符合大数据特征。随着人们对规避金融风险、金融规律预测等需求增加，不同的研究者基于大数据、数据挖掘技术探索了许多金融时序数据序列分析方法。

时间序列分析[4-6]是通过时间序列数据揭示现象随时间变化规律，预测现象未来发展的一种方法。最常用的时间序列分析模型有自回归滑动平均[7-8]（autoregressive moving average，ARMA）模型和自回归条件异方差（GARCH）模型[9-10]。ARMA 模型不能描述异方差性质。GARCH 模型在金融学中常用于资产定价、风险管理和波动率预测。人们一致认为，在考虑股价动态、利率和波动性时，应考虑到经常观察到的典型事实。这些风格化的事实包括非正态性、跳跃行为、波动性聚集和杠杆效应。由于估算的复杂性，在实施过程中仍然存在一些困难和挑战。首先，GARCH 模型属于一类明显增加优化复杂度的非线性动力学；其次，股票市场的波动性是不可观测的，股票价格是在大量噪声中采集的，这使得收益率表现为一个状态变量。例如，在应用时间序列模型时，传统的频率统计方法往往遇到许多困难。

为此，有学者将贝叶斯方法[11-13]引入时间序列分析。贝叶斯方法能够综合统计模型和分析各种信息，因此广泛应用于财务数据风险分析的研究。在数据分析中，通常将频率分析和主观（经验）相结合来建立合理的先验信息，对利用先验变化做出的统计推断进行敏感性分析，以确认所得到的统计推断的合理性。序贯贝叶斯方法[14-15]的系统研究是动态的、时间相关的，可以周期性地观察，根据每次观察到的状态和以前状态的记录，从一组可行解中选择一个最优决策，然后观察下一步可能的状态，收集新的信息，做出一个新的最优决策，重复操作，称为序贯学习，系统的下一个可能状态是随机的或不确定的、自简的；因此，序贯参数学习方法比数值极大似然估计[16]（maximum likelihood estimation，NMLE）具有更高的鲁棒性和准确性。

笔者提出一种序贯贝叶斯学习方法来估计非对称 GARCH 模型，该模型考虑了杠杆效应，描述了收益率和波动率之间的负相关关系，从而解决股票

收稿日期：2024-06-15；**修回日期**：2024-07-20
第一作者：李慧玲（1968—），女，河北人。

模拟模型估计中的复杂数值问题，方便得到参数的后验分布。

1 模型介绍

一般情况下，在非对称 GARCH 动力学假设下，股票价格可描述如下：

$$S_t = S_{t-1}e^{r_t+\lambda\sigma_t+\varepsilon_t-\varphi_z(\sigma_t)};\ \ \varepsilon_t=\sigma_t z_t;$$
$$\sigma_t^2=\alpha_0+\alpha_1\sigma_{t-1}^2(z_{t-1}-\gamma)^2+\beta\sigma_{t-1}^2。\tag{1}$$

式中：r_t 为无风险资产的无风险利率；λ 为风险的市场价格；σ_t 为条件波动率；ε_t 为残差；z 为观测值；S 为股票价格；$\varphi_z(\sigma_t)$为平均校正项，由带参数 $\beta, \alpha_1, \alpha_0, \gamma$ 的矩母函数计算得出，则有：

$$\varphi_z(\sigma_t)=\log E[e_{\varepsilon t}\mid\sigma_t]。\tag{2}$$

式中 $E[\bullet]$为期望。该式使得价格过程服从指数鞅。

此外，本研究中假设非对称 GARCH 动力学的初始参数如表 1 所示。这些参数值与股市估算值最接近。为了比较 NMLE 和序贯贝叶斯学习方法(SBLA)的估计结果，本研究模拟了 10 000 条股票价格路径(L=10 000)，每条路径包含 1 000 个观测值(T=1 000)。

表 1 非对称 GARCH 动力学的初始参数

参数	β	α_1	α_0	λ	γ	σ_0	z_0	s_0
值	0.85	2 000	0.1	16	400	0.012	0	1 000

2 算法实现

2.1 极大似然估计

首先，利用 NMLE 方法进行评估。在非高斯情况下，可以使用基于给定特征函数的快速傅里叶变换，则残差 ε_t 满足：

$$E[\varepsilon_t\mid F_{t-1}]=0;\ \ E[\varepsilon_t^2\mid F_{t-1}]=\sigma_t^2。\tag{3}$$

式中 F 为傅里叶变换。

在此基础上，可以利用高斯核密度函数(例如正态分布)，迭代地获得收益率的拟似然函数；因此，对数似然函数可描述如下：

$$\log f(\varepsilon_t;\theta)=-(y_t-r_t-\lambda\sigma_t+\varphi_z(\sigma_t))^2/2\sigma_t^2+\log(1/\sqrt{2\pi}\sigma_t);\tag{4}$$

$$y_t=\log\left(S_t/S_{t-1}\right),\ \ t=1,2,\cdots,T。\tag{5}$$

式中：T 为时序时间步长；θ 为模型参数。进一步，评估的目标函数定义如下：

$$\hat{\theta}=\arg\max\sum_{t=1}^{T}\log f(\varepsilon_t;\theta)。\tag{6}$$

当 z_t 遵循离散时间 lévy 跳过程，其为非高斯分布的。假设 $\phi_z(u)$为特征函数，因此可通过快速傅里叶变换估计密度函数。进一步，笔者研究了一种序贯贝叶斯学习方法，通过过滤非正态分布的 GARCH 模型的历史噪声对其进行联合估计，并给出了新息的条件密度 $f(\varepsilon_t)$，则有：

$$f(\varepsilon_t)=\int_{-\infty}^{\infty}e^{-iuz_t\sigma_t}\phi_z(\sigma_t u)\mathrm{d}u\Big/2\pi。\tag{7}$$

此外，条件密度函数计算如下：

$$f(\varepsilon_t\mid\sigma_t)=\frac{1}{2\pi\sigma_t}\int_{-\infty}^{\infty}e^{-iuz_t\sigma_t}\phi_z(\sigma_t u)\mathrm{d}(\sigma_t u)=f(z_t)/\sigma_t。\tag{8}$$

需注意，条件波动率 σ_t 可以从动力学假设式(1)中迭代计算。

2.2 序列贝叶斯学习

对于金融时序数据，通常使用数值最大似然法来估计模型。然而，该方法通常需要大样本周期来实现似然函数的优化，且投资者需要在估计过程中纳入先前的信息。该过程可以应用贝叶斯方法。基于贝叶斯规则，后验密度 $p(\theta|y_{1:T})$可通过式(9)计算：

$$p(\theta\mid y_{1:T})=p(y_{1:T}\mid\theta)p(\theta)。\tag{9}$$

式中：$p(\theta)$为先验概率；$p(y_{1:T}|\theta)$为其似然函数，可通过粒子滤波计算。故有：

$$\log f(y_{1:T};\theta)\approx\sum_{t=1}^{T}\log\left\{\frac{1}{N}\sum_{i=1}^{N}w_t^{(i)}\right\};\tag{10}$$

$$w_t^{(i)}=p\left(y_t\mid\sigma_t^{(i)};\theta\right)=f\left(\varepsilon_t^{(i)}\right)。\tag{11}$$

式中 $w_t^{(i)}$ 为重要性权重。对于每个样本 $\theta^{(m)}$，本研究也近似了其后验权值 $w(\theta^{(m)})$。具体来说，在粒子的有效样本量(efficient sample size，ESS)低于阈值(例如，采样粒子数的一半)时使用 MCMC 算法；因此，参数更新可描述如下：

$$\alpha(\theta^*)=1\wedge\frac{p(\theta^*)p(y_{1:T}\mid\theta^*)N(\theta;\hat{\theta}^*,\hat{\Sigma})}{p(\theta)p(y_{1:T}\mid\theta)N(\theta^*;\hat{\theta},\hat{\Sigma})}。\tag{12}$$

进一步，估计参数及其方差计算如下：

$$\alpha(\theta^*)=1\wedge\frac{p(\theta^*)p(y_{1:T}\mid\theta^*)N(\theta;\hat{\theta}^*,\hat{\Sigma})}{p(\theta)p(y_{1:T}\mid\theta)N(\theta^*;\hat{\theta},\hat{\Sigma})};$$
$$\hat{\theta}=\sum_{m=1}^{M}\theta^{(m)}\tilde{w}(\theta^{(m)});$$
$$\hat{\Sigma}=\sum_{m=1}^{M}(\theta^{(m)}-\hat{\theta})\tilde{w}(\theta^{(m)})(\theta^{(m)}-\hat{\theta})'。\tag{13}$$

因此，在给定模型不确定性的情况下，贝叶斯学习方法可以随着新信息的到来而不断更新。

综上，模型执行过程如图 1 所示。

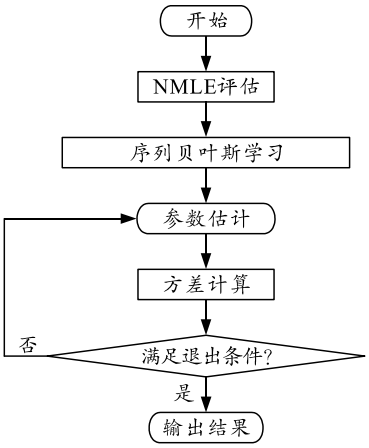


图 1　模型执行过程

3　仿真与分析

3.1　仿真环境

在对所提方法进行测试时，配置的测试环境如表 2 所示。应用服务器和数据库服务器采用 inter L5520 CPU、win7 系统和 MySQL5.5.28 数据库。

表 2　仿真环境相关参数

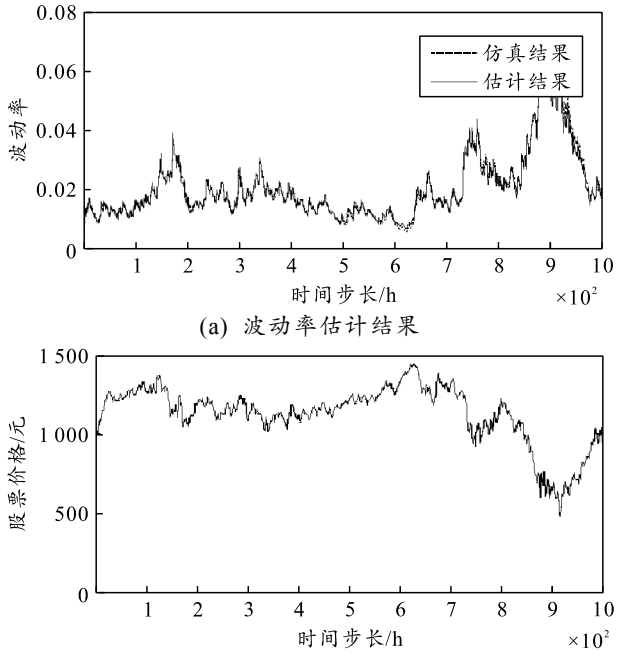
参数	服务器	客户端
CPU	inter L5520	IOS
RAM	16 G	4 G
操作系统	WIN 7	
数据库		MySQL5.5.28

3.2　模型性能测试

仿真阶段首先研究了正态分布的非对称 GARCH 模型，其中新息服从离散时间布朗运动。接着，根据表 1 参数信息对股票价格的路径进行了 10 000 次模拟，并对每条路径分别进行估计，得到了参数的分布。从每条路径分别用极大似然估计和贝叶斯学习方法估计非对称 GARCH 动力学模型。在模拟 10 000 条路径后，对每条路径的参数进行估计，然后对样本参数进行分布分析。表 3 所示为极大似然估计与序列贝叶斯学习评估过程统计结果，可以看出序列贝叶斯学习更加接近实测值。图 2 所示为估计结果，可以看出序贯贝叶斯学习方法能够较好地模拟股票波动及价格趋势。仿真结果进一步验证了所提方法的有效性。

表 3　参数统计

参数	实测值	极大似然估计	标准差	序列贝叶斯学习	标准差
β	0.850 0	0.819 8	0.819 8	0.851 4	0.016 8
α_1	0.060 0	0.055 3	0.020 2	0.059 5	0.006 2
α_0	2.500e−6	4.231e−6	2.112e−6	2.710e−6	0.464e−6
λ	0.040 0	0.038 1	0.032 3	0.048 5	0.011 1
γ	1.200 0	1.570 6	0.808 1	1.172 2	0.136 2
ps	0.996 4	0.989 3	0.011 4	0.992 6	0.004 5



(a) 波动率估计结果

(b) 股票价格曲线

图 2　序贯贝叶斯学习方法估计结果

4　结论

笔者基于大数据中序贯贝叶斯学习方法对时间序列数据进行研究与分析，提出一种基于序贯贝叶斯学习评估模型。该模型可在不确定性情况下，随着新信息的到来而不断更新，从而完成学习过程。在仿真分析时，忽略了数据坏值、无效数据等情况。未来，可对数据清洗、数据特征提取等进行研究，进一步提升模型性能。

参考文献：

[1] 郭志东. 大数据背景下金融统计发展策略探究[J]. 中国市场, 2021(16): 64-66.

[2] 王孟. 大数据时代下的绿色金融前景分析[J]. 商展经济, 2021(10): 69-71.

[3] 刘海洁, 何锡彤. 大数据在商业银行风险管控中的应用[J]. 中国集体经济, 2021(16): 99-100.

[4] 郝宇星. 金融支持文化产业发展的实证研究——基于北京市的时间序列数据分析[J]. 商讯, 2021(15): 87-88.

[5] 高铭甫, 周丹文. 基于 GARCH 模型 VAR 方法的外汇汇率波动性分析[J]. 中国商论, 2021(6): 72-74, 80.

[6] 宋瓷婷. 基于时间序列分析的银行信息系统风险预警平台[J]. 中国金融电脑, 2021(2): 70-76.

[7] 宋华, 胡涛文. 基于 ARMA-GARCH 模型的利率市场风险度量[J/OL]. 宜宾学院学报: 1-7[2021-06-05]. http://kns.cnki.net/kcms/detail/51.1630.Z.20210602.1046.006.html.

（下转第 54 页）