# 金融高频数据挖掘研究评述与展望

朱建平　魏　瑾　谢邦昌

内容提要: 金融高频数据构成海量数据集, 属于数据挖掘的研究范畴, 然而在金融高频数据的研究中, 数据挖掘技术尚未得到足够的重视。金融高频数据的研究目前主要集中于对波动率、交易间隔等特征的建模, 最优抽样间隔的选择等应用领域, 国内鲜有方法论框架下直接将金融高频数据作为研究对象的理论讨论与分析, 这不可避免导致对高频数据认识上的一些误区和不一致。为此, 本文对国内外金融高频数据的研究现状进行了剖析, 澄清了金融高频数据的概念与特征, 并从统计的视角重新审视了金融高频数据研究。在此基础上, 提出了金融高频数据挖掘进一步的研究思路。

关键词: 金融高频数据　数据挖掘　统计分析

## 一、引言

高频数据并不是新事物, 地质、气象、工厂生产线、实验观测等各领域的高频数据俯拾皆是, 然而随着计算机存储技术的飞速发展, 记录高频数据日趋便捷, 且处理大规模数据的数据挖掘技术也越来越成熟, 因而高频数据问题日益受到学界广泛关注。特别在金融领域, 鉴于中国证券市场历史短暂且发展迅速, 跨期的观测数据往往在可比性上不能令人满意, 如果采用高频数据, 那么就可以在较窄的观测区间内产生满足分析所需要的数据量, 同时可以对市场微结构模型做出恰当的验证。

对金融高频数据的研究至少可以追溯到上世纪 80 年代, 如 Harris(1986)等发现交易量波动率在日内呈“U”型。随后有 Baillia Bollerslev(1989, 1990)、Andersen & Bollerslev(1994)、Goodhart & Maureen(1997)、Granger(1998)、Bauwens(2008)、Andersen(2001)、Nielsen & Frederiksen(2008)等多位学者在波动率和交易间隔建模等方面的跟踪研究。

国内关于高频数据的研究主要有: 来升强等(2010)针对粗集分类方法因离散化而损失数值型变量提供的高质量信息, 提出一种基于 Bayes 概率边界域的粗集分类方法, 并将其应用于高频数据。然而该文只是把高频数据作为方法的验证, 并没有正面讨论高频数据。徐国祥等(2007)通过衡量残差密度函数的参数和非参数估计值之间的紧密程度对 ACD 模型的设定进行了检验。唐勇等(2006)研究了针对高频金融时间序列而开发的波动率测量方法——已实现极差波动(realized range-based volatility, RRV)的加权形式。韩冬等(2006)研究了流动性的“周内效应”和“日内效应”后发现, 当控制波动性、交易量和股价等对流动性有重要影响的变量时, 效应依然存在。凌士勤等(2005)提出基于高频数据的分类信息混合分布 GRACH 模型。

事实上, (超)高频数据这类大规模数据集本身是数据挖掘的对象, 然而在金融高频数据的研究中, 数据挖掘技术尚未得到足够的重视, 目前的研究仍多仰赖于经典统计方法和计量经济模型的修正。另一方面, 统计分析的对象是数据, 而国内鲜有方法论框架下直接将金融高频数据作为一类研究对象进行的讨论与分析, 这不可避免对金融高频数据产生一些认识上的误区。本文从统计视角对高频数据挖掘研究过程中的一些误区和被忽略的问题展开讨论, 并对金融高频数据挖掘进一步研究的思路做了展望。

## 二、金融高频数据的概念及特征澄清

经济金融领域研究的“高频数据”、“超高频数据”、“低频数据”主要是根据计量单位来做区分的。金融“高频数据”(high-frequency data)特指日内数据(high-frequency intra-daily data), 即主要以小时、分钟或秒为采集频率的数据; “低频数据”通常指以天、周、月、年作为计量单位的数据; 而金融“超高频数据”则是对交易过程实时采集的数据(tick-by-tick data), 即按照每笔交易的发生逐笔记录的

数据。这里需要注意的是,超高频数据并不是抽样数据,而是全样本数据;不是等间隔数据,而是不等间隔且间隔随机的数据。

从函数的观点来看,金融高频数据以时间 t 为自变量,$t_i$ 指时刻 i,并假定 $\Delta t_i = t_i - t_{i-1}$ 等间隔;而超高频数据则以交易为自变量,$t_i = t(i)$ 指第 i 次交易的时刻,$\Delta t_i = t_i - t_{i-1}$ 是两次交易的时间间隔,往往是不等间隔的。

应该注意到,首先,实际交易时间与模型时间(钟表时间)的这种不一致性在超高频数据中还表现为,在同一市场上,多笔交易同时发生,甚至可以同时以不同的价格成交,即"同一时刻的交易可能会因为交易系统或数据传输等原因从而在不同的时刻发布出去;而不同时刻的交易也可能在同一时刻被合并称同一数据被发布"。从这个角度来讲,以秒来计量时间都已经是非常大的尺度了。其次,金融高频数据和超高频数据的价格都是离散的(price discreteness)。这是因为交易所对最小交易价格单位有限制,所以每笔成交价格只能是最小交易价格(tick size)的整数倍。我们通常遇到的时间序列、连续时间金融,其区别主要是自变量(时间)是否离散,因变量(x(t))的取值都是在整个实数域或大于 0 的部分,而这里的离散价格意味着因变量的取值是离散的。第三,与低频数据相比,金融高频数据的质量往往并不高(大规模数据的基本特征),因为交易数据会因种种原因而缺失,某些交易的确切时间也不见得准确,而且还有微结构噪音等因素干扰。所以,在进行金融高频数据挖掘之前,数据预处理工作仍然是非常重要的环节。而了解清楚市场是如何运作的、数据是如何产生的,无疑对数据预处理有非常大的助益。

## 三、对金融高频数据研究的思考

近年来,国内外学术界对金融高频数据、超高频数据展开了广泛的研究,这也为我们进行高频数据挖掘提出了新的思考。

1. 澄清金融高频数据认识上的误区。在明晰金融高频数据概念的同时,我们发现很多文献对高频与超高频这两个概念混同使用,高频、超高频与低频之间的界限也较为随意。事实上,根据数字信号处理的相关理论,设若频率小于某个临界值,会出现混叠现象,进而无法真实还原序列所要传达的信息。为此,我们需要从更严格的意义上对低频数据、高频数据与超高频数据做出界定和辨析,进而从统计学理论和方法的角度来审视金融高频数据挖掘的内容和方法,这一方面有利于明确统计方法的应用现状和所面临的困难;另一方面可以引起统计学界对金融高频数据挖掘的广泛关注,也有利于统计学方法研究的进一步拓展和深入。

此外,不少文献认为金融高频数据仅仅是加细了取样间隔,增加了样本容量,因而包含了比以往更多的信息。然而事实上并非取样频率越高就越精确,因为取样频率越高也越容易受到微结构噪声(microstructure noise)的影响。需要注意,对金融高频数据的建模方法不同于低频,比如 ARCH 模型族在金融高频数据中基本无法使用;超高频数据与高频数据的研究方法也有质的区别,比如超高频数据取样间隔不等距且随机,而多数统计计量方法都是针对固定等距情形而设计的。但是目前国内对金融(超)高频数据的研究多集中在引入国外模型做应用实证分析,对研究方法的探讨并不多。

2. 探索金融高频数据挖掘的统计方法。单从数据处理的角度来看,低频数据似乎可以看作是对高频数据的抽样。在抽样理论中,用一个点代表它所属的"层"是可以接受的,而事实上日内高频数据似乎更应该理解为"群",因为群间有相似的统计特征(如"U"型分布),群内异质性较大(如开盘和收盘交易量较大,而中间时段交易量小)。所以需要对高频数据的日内效应进行更为细致的统计观察和分析,进而探索其中的微结构。

以波动率的研究为例,金融研究领域的很多模型都是为刻画波动的时变性、聚集性、非对称性和长记忆性等特征提出的,然而这些模型大都无法直接应用于高频数据,与低频数据采用 ARCH 模型族讨论波动不同的是,高频数据主要采用已实现波动率(realized volatility)来对波动率进行测量,通过波动率来深入分析和研究交易的内在机制。这方面主要集中在对市场微观结构理论的探讨。与时间序列模型强调数据的统计性质所不同的是,微结构模型(market microstructure)更多地关注市场行为,着意于交易的细节,如交易价格的形成过程、代理人的行为、交易成本、交易机制等。狭义地来讲,微结构模型旨在考察市场参与者的潜在需求如何转化为交易价格和交易量的过程。尽管这部分内容与金融高频数据分析紧密相关,但从数据挖掘角度的深入研究并不多。这样就有必要从统计学理论和方法的角度来审视金融高频数据挖掘的内容和方法。

3. 从观测尺度来理解高频与低频数据的差异。

金融理论通常采用几何布朗运动( the Geometric Brownian Motion) 来刻画价格波动, 但 Zhou( 1992) 的研究发现, 金融高频数据不再像低频数据那样遵循布朗运动。那么二者仅仅是频率上的差别吗? Zhou( 1992) 的研究表明, 高频与低频的区别仅仅是噪声层面的: 在低频数据里, 噪声可以被忽略; 然而在高频数据里, 噪声是显著的。这就好像是在较小的尺度上( 如短期) 可能犯错, 导致出现一个凸点, 但是在较大的尺度上( 如长期) , 这个凸点可能就被“磨圆”了。

所以, 不同尺度下, 可以有截然不同的结论, “横看成岭侧成峰, 远近高低各不同”, 从系统论的角度看, 我们必须承认, 不同层次( 类别) 有不同层次( 类别) 的规律( 除了无特征尺度的“自相似”, 它在不同的尺度上表现出相似或统计相似的性质) 。比如研究了微观个体的行为, 并不可以简单加总去推断群体的行为; 研究了短期的行为, 也不可以妄断长期。应该注意, 这里本身并不涉及推断问题, 不能用这个层次的观察来推断另一个层次, 推断应该是在同一个层面( 尺度) 的, 包括外推和横向比较。比如, 由可获得的样本推断未知总体, 它仅仅是数量上的策略。

4. 抽样并不必然造成信息的损失。大多研究金融高频数据的文献认为, 金融市场上的信息对证券价格变化的影响具有连续性, 而低频数据是离散的, 这必然会造成信息的丢失。而且, 数据频率越低, 则信息丢失就越多。但是, 根据数字信号处理的相关理论, 模拟信号( 连续信号) 首先要经过离散化处理( 抽样) 变成数字信号, 才可以进入下一步分析。

退一步而言, 根据统计抽样理论, 如果采用合适的抽样方法, 那么抽样的效果并不弱于全面调查。所以, 问题并不在于是否采用抽样方法, 而在于如何设计和实施抽样。由于很多金融时序数据在总量观察的尺度上多呈异方差( 异质程度较高) , 所以通过提高抽样频率来挖掘其中所包含的丰富的波动信息是很自然的。另一方面, 根据总体辅助信息设计合理的抽样方法也是值得努力的方向。

事实上, 从统计的视角来看, 过于细致的数据并不利于展现数据的总体特征。因而才会引出分组的重要性, 即分组对数据进行人为的、有目的的离散化梳理, 这有助于问题的发现。模型也正是通过显现本质忽略枝蔓而简化了现实, 使我们专注于要解决的问题。

5. 金融高频数据的本质在于微结构发现。相对于低频数据而言, 高频数据不仅仅是加细了取样间隔, 增加了样本容量, 实现了大样本推断, 更重要的是, 金融高频数据挖掘的目标其实并不是为了改进抽样和样本代表性, 而是为了发现日内的交易行为结构。比如, 原先只是取日收盘价, 以日作为分析单位; 现在则加细日内的间隔, 以发现日内的微结构。我们希望通过这种研究视角的变换——改变了分析的单位或尺度——来发现更多背后的信息, 如宏观经济学转向微观经济基础构建、金融学转向行为金融的研究一样。

## 四、金融高频数据挖掘研究展望

在金融数据挖掘过程中, 统计学的理论和方法越来越受到重视。因此, 基于数据挖掘的视角, 系统地研究金融高频数据的统计分析方法, 对金融机构更好地测度各种统计指标具有重要的现实意义。在此我们有必要进一步明确金融数据挖掘研究的基本思路。

1. 金融高频数据聚类。针对金融高频数据聚类分析主要考虑: ( 1) 如何反映金融高频数据在时间上的动态特征, 如何处理更新数据对已有聚类的影响; ( 2) 在 K- NN 聚类的基础上, 设计出合适的权重函数, 使其既能满足降维的需要, 又能充分反映时间变化的影响; ( 3) 借鉴投影寻踪方法的思想, 在金融高频数据的高维空间中找出最优线性基向量并将其作为降维子空间, 同时把相应的线性变换矩阵作为原维度的权重矩阵。进一步地, 还可以研究如何将这一思想推广到非线性情形, 通过降维得到聚类分析的结果。

2 金融高频数据分类。金融高频数据的分类过程更注重于保证分类模型对于更新数据的适应性和稳定性。可以根据更新数据动态地建立新枝或删除旧枝, 利用分类回归树的改进形式完成对非数值型高频数据的分类任务; 借助动态决策树和广义估计方程解决决策树分类中混合型高频数据的分类; 选择适当的基函数对金融高频数据进行拟合, 可分析大型高频数据的分类问题。在这些方法研究中, 重点是如何设计具有时变特征的权重因子。

3. 金融高频数据波动性分析。波动性是实际结果偏离期望结果的程度, 它能够反映未来收益的不确定性, 这种波动性可以通过规范的统计方法进行量化。更重要的结合时序分析的基本思想, 对金融高频数据中包含的不同性质、不同程度、不同周期的规律性特征进行分离, 用适当的广义可加模型进行描述, 并采用时变参数反映金融高频数据的动态特

征。另外,还可以利用粗糙集等知识推理方法进行约简,将大量不必要的细节信息泛化为若干代表性知识,实现知识泛化。

4. 金融高频数据压缩。指在给定的误差设定下,把历史数据压缩为一个相对较小的概要数据集,同时保证概要数据集对历史数据的代表性。金融高频数据压缩方法和统计模型结合较为紧密,例如可以利用线性拟合、多项式拟合、独立成分分析等统计和数学模型。同时借助函数数据分析的观点,将金融高频数据函数化,不仅可以起到压缩的目的,还可以对金融高频数据的微结构进行分析。

5. 规则发现。相对于其他挖掘方法,规则发现更适合用于非标准金融高频数据的探索性分析。例如对规则的有效性和稳定性,利用抽样误差公式进行抽样并根据抽样频数进行频数估计以及有损频数估计。

6. 随机交易间隔分析。超高频数据是对交易过程的实时记录,因而记录的时间间隔不等且随机;另一方面,钟表时间和交易时间是不同的,更准确地说,这里指的是交易时间间隔随机(stochastic duration),它因交易而变。所以,较短的时间间隔意味着交易频繁,而较长的时间间隔则意味着交易停歇,因此将交易间隔作为随机变量可以考察有关日内市场活动的信息。

7. 离散价格与受限因变量模型。大量实证分析表明,日内相邻两次交易的价格在日内的变动非常小。涨跌停和熔断机制也限制了价格的日内最大波动区间,且价格离散波动。针对以上特征,可以对离散选择模型(discrete choice model)和排序选择模型(ordered choice model)进行扩展来分析离散价格的波动。

8. Copula- VaR 方法的研究。以数据挖掘的角度,从 Monte Carlo 模拟法对 Copula- VaR 的估计进行突破性研究。Monte Carlo 模拟方法是模拟随机变量可能发生的情况,根据变量的模拟变化路径得到大量的模拟数据,进而得到变量未来的分布情况,在此基础上便可以对计算 VaR 值进行深入的研究。

9. 金融高频数据挖掘的应用研究。研究上述问题的主要目的在于应用。我们将利用 Markov 结构转换的 Copula 模型,捕捉不同波动水平下金融市场间非线性、非对称相关关系的变化,进而分析出金融市场间是否存在传染,是否会使它与其他金融市场的相关关系增强,或者说会将危机传染到其他金融市场。另外,对股市(超)高频数据投机交易行为进行探测,并通过股票交易和访问日志数据分析来优化金融网页内容,提高金融网站平均访问率和浏览时间。在研究过程中,将通过计算机软件实现金融高频数据挖掘结果的可视化,并实现人机交互式的数据挖掘过程。

参考文献:


徐国祥 金登贵,2007:《基于金融高频数据的 ACD 模型非参数设定检验》,《统计研究》第 4 期。

唐勇 张世英,2006:《高频数据的加权已实现极差波动及其实证分析》,《系统工程》第 8 期。

来升强 谢邦昌 朱建平,2010:《基于 Bayes 概率边界域的粗集分类方法及其在高频数据中的应用》,《统计研究》第 3 期。

韩冬 王春峰 岳慧煜,2006:《流动性的"周内效应"和"日内效应"——基于指令驱动市场的实证研究》,《北京航空航天大学学报(社会科学版)》第 2 期。

Andersen & Bollerslev(1997), "Intraday periodicity and volatility persistence in financial markets", Journal of Empirical Finance 4: 115- 158.

Baillie & Bollerslev(1990), "Intra- day and inter- market volatility in foreign exchange rates", The Review of Economic Studies 58(3).

Goodhart & O'Hara(1997), "High frequency data in financial markets: Issues and applications", Journal of Empirical Finance 4(2- 3): 73- 114.

Granger, C. W. J. (1998), "Extracting information from mega- panels and high- frequency data", Statistica Neerlandica 52: 258- 272.

Harris, L. (1986), "A transaction data study of weekly and intradaily patterns in stock returns", Journal of Financial Economics 16: 99- 117.

Nielsen & Frederiksen(2008), "Finite, sample accuracy and choice of sampling frequency in integrated volatility estimation", Journal of Empirical Finance 15: 265- 286.



(作者单位:厦门大学经济学院
台湾辅仁大学统计资讯学系)


(责任编辑:白丽健)