高频交易数据研究的思考

宽客界 2015-10-10 17:59:52

近年来，国内外学术界对金融高频交易数据、超高频交易数据展开了广泛的研究，为此类研究提出了新的思考。

一、澄清金融高频交易数据认识上的误区

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，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。

二、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（realizedvolatility）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。

三、从观测尺度来理解高频与低频数据的差异

金融工程理论通常采用几何布朗运动（the GeometricBrownian Motion）来刻画价格波动，但研究发现，金融高频交易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？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;然而在高频交易数据里，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。

四、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。

五、金融高频交易数据的本质在于微结构发现

相对于低频数据而言，高频交易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、金融工程学转向行为金融的研究一样。（文章来源：宽客界微信公众号quantview）

宽客界 高频数据 波动率 量化投资