金融高频数据分析的现状与问题研究

常　宁[1]，徐国祥[2]
（1. 上海财经大学 统计学系，上海 200433；
2. 上海财经大学 应用统计研究中心，上海 200433）

摘　要：近年来，在西方国家对金融高频数据的分析已成为实业界和学术界的热点问题和难点问题。本文讨论了金融高频数据的概念和特征，分析了对高频数据分析的基本动因，阐述了金融高频数据分析已涉及的主要领域，探讨了金融高频数据分析中遇到的问题。最后，还对金融高频数据分析的发展趋势作出了展望并探讨了我国在这一领域应用研究的重点。

关键词：金融市场；证券市场；金融高频数据分析；市场微观结构
中图分类号：F830.91　**文献标识码**：A　**文章编号**：1001-9952(2004)03-0031-09

一、金融高频数据及其特征分析

1. 什么是金融高频数据

近年来，计算工具与计算方法的发展，极大地降低了数据记录和存储的成本，使得对大规模数据库的分析成为可能。所以，许多科学领域的数据都开始以越来越精细的时间刻度来收集，这样的数据被称为高频数据(high-frequency data)。金融市场中，逐笔交易数据(transaction-by-transaction data)或逐秒记录数据(tick-by-tick data)就是高频数据的例子，值得注意的是这里的时间通常是以"秒"来计量的，具体如 NYSE(New York Stock Exchange)的交易与报价数据库(Trades and Quotes)所记录的从 1992 年至今的 NYSE、NASDAQ 和 AMEX(American Exchange)的全部证券的日内交易和报价数据、Berkeley 期权数据库所提供的 1976 年 8 月至 1996 年 12 月的期权交易数据、以及美国外汇交易 HFDF93 数据库中德国马克－美元的现汇交易报价数据等，都是金融高频数据。

2. 金融高频数据的主要特征

收稿日期：2003-12-08

作者简介：常　宁(1973－)，女，陕西西安人，上海财经大学统计学系教师，经济学博士；
徐国祥(1960－)，男，上海人，上海财经大学应用统计研究中心主任，统计学系教授，博士生导师。

与传统的低频率观测数据(如周数据、月度数据等)相比,按照更短时间间隔所取得的金融高频数据呈现出了一些独有的特征,正是这些特征,诱发了人们对金融高频数据分析的日益浓厚的兴趣。以 NYSE 的交易数据为例,金融高频数据主要有四个特征:

一是数据的记录间隔不相等,因为市场上某只股票的交易并不一定以相同的时间间隔发生,这样所观测到的交易价格等变量的时间间隔就不相等;

二是所记录的价格数据是离散变量,如在 NYSE 中,某项资产的价格变动只以计量单位 tick size[①]的若干倍而发生,这样所记录的逐项交易价格就变成了一个离散取值的变量;

三是数据存在日内周期模式,在正常交易条件下,NYSE 的交易量往往在每一天的开盘时间和收盘时间附近较大,而在午饭时间左右较小,形成了一个"U"型的模式,随之而来的,是交易与交易之间的时间间隔在一天内也呈现出了循环模式的特征;

四是多笔交易同时(甚至是以不同的价格)发生,这种现象部分归因于在每天交易量较大的时候,以秒来计量时间都成为一个太长的时间刻度了。

二、金融高频数据分析研究的现状

1. 金融高频数据分析的基本动因

从金融高频数据产生至今,对金融高频数据的分析一直是金融研究领域中一个倍受瞩目的焦点。这可以归结为两个原因:一个是由于对金融高频数据本身所具有的特征值的关注。通常所指的交易数据,除了交易价格外,还包括与交易相连的询价和报价、交易数量、交易之间的时间间隔、相似资产的现价等等,因此,对于金融高频数据的分析,实质上是一个关于"以不同时间间隔观察到的、具有不规则强度、既有离散变量又有连续变量的"复杂多变量问题。这样如何从总体上来分析金融高频数据、又如何处理具体金融交易中高频数据的特殊性,便成为众多金融领域的从业者和研究者所面临的一个有趣而又富有挑战性的课题。

另一个是因为金融高频数据对理解市场的微观结构来说相当重要。对金融高频数据的逐步积累和了解,不仅转变了一些陈旧的研究理念,如以前认为短期的价格波动是不相关的噪音并且不值得去搜集,但现在我们知道高频数据中的这种波动恰恰包含着理解市场微观结构的重要信息;而且随着对金融高频数据统计特征认识的深化,也使先前一些关于如金融市场同类性(homogeneous)、短期价格波动服从高斯随机游程(gaussian random walk)的古典经济假定受到了质疑。不难看出,在探寻金融市场微观结构的过程中,需要对基础经济理论、研究方法和计量模型等进行不断地创新和完善,而金融高频数据及其分析的出现则正好为这些转变的实践提供了条件。

2. 金融高频数据分析已涉及的主要领域

尽管人们对金融高频数据分析研究的历史并不长，但是目前的发展状况却着实令人鼓舞。众多学科的研究者对此都表现出了极大的兴趣，分别从各自不同的角度对金融高频数据进行了探索和研究。已有研究所涉及的内容之广令人无法一一穷尽，所以我们在此以金融高频数据研究的四个主要分支为脉络，有所侧重地阐述一些具有代表性的研究内容。

第一个分支是关于金融高频数据库的研究。其中 Robert Wood 是创建研究市场微观机构（金融高频）数据库的先驱。在他的文章（2000）中，Wood 不仅从对金融市场微观结构研究的初衷、对结构数据的基础检验、TAQ 数据库的组织形式和特征等角度对金融高频数据库的发展历程做了介绍，而且还讨论了金融高频数据量（如 NASDAQ 报价数据等）的快速增长趋势以及这种数据量的增长趋势在市场结构研究中的应用问题。这些内容对于了解金融高频数据库的组织结构、形式和数据特征来说都是非常必要的。

第二个分支是关于金融高频数据分析应用于对市场微观结构分析的研究。在这个领域中，最初的文献是关于日内（intra-day）收益与波动性时间序列的模式的研究，如 Wood（1985）、Harris（1986）、Lockwood，Linn（1990）和 McNish（1993）等是最早一批对 NYSE 高频交易数据进行研究的人，而 Goodhart、Figliouli（1991）和 Guillaume（1994）等人则是最早对外汇市场的高频交易数据进行研究的先驱。此后，便陆续不断地有许多文章对日内金融市场数据的行为特征作了更深入的研究。从 Goodhart 和 O'Hara（1997）所做的有关研究文献纵览中可以看出，基于金融高频数据对市场微观结构所作的实证研究主要集中于以下几个方面：

（1）对金融市场交易数据观测时间间隔特征的研究；

（2）对交易数据如波动性、交易量与价格差额之间交互作用的研究；

（3）对价格差额的决定因素的研究；

（4）对金融高频数据的波动性及其记忆的研究；

（5）对促使价格变动的交易的研究；

（6）对收益、报价等交易数据中的自相关性以及收益、报价、交易与交易之间的横向相关关系的研究；

（7）对金融高频数据的季节性与非线性特征的研究；

（8）对金融市场的技术分析和市场效率的研究；

（9）对不同金融市场（如证券市场与衍生证券市场）之间联系的研究等等。

最近几年，关于对市场微观结构的实证研究在深度和广度方面又有了新的进展，其中尤其以对股票市场高频数据的分析最具代表性。主要有用高频交易数据对不同交易系统（如 NYSE 的公开喊价系统与 NASDAQ 的计算机交易系统）在价格发现中的效率进行比较；用高频交易数据对某一个特殊股票

的报价与询价的动态性进行研究(如 Hasbrouk,1999;Zhang,Russell 和 Tsay,2001);在一个订单驱动的股票市场(如台湾股票市场)中,高频交易数据被用于研究订单的动态性以及回答"是谁提供了市场的流动性"问题。此外还有 Hol 和 Koopman(2002)用 S&P500 的高频数据对股票指数的波动性进行了预测研究;Bollerslev、Zhang(2003)将股票市场的高频交易数据应用于对因素定价模型(factor pricing models)中系统风险因素的计量和建模等一系列的相关研究。

第三个分支是关于金融高频数据分析中所使用的计量模型的研究。随着金融高频数据的不断增加,如何使用模型来恰当地描述这些数据就成为一个重要的问题。从计量经济学角度来看,金融高频数据的一个最显著特征是观测值以变动的、随机的时间间隔取得。该特征隐含着对我们所熟悉的、固定的、等值的时间间隔数据的偏离,也意味着原有的一些深受喜爱的模型,如关于波动性研究的 GARCH(Generalized Autoregressive Conditional Heterosce Dasticity)模型、SV(Stochastic Volatility)模型等将不再适用。与以往大多数的理论模型不同[②],近来计量模型研究的核心内容是交易间隔(intratrade duration)与交易特征值,如收益、询报价差额、交易量等之间的 Granger 因果关系。这些模型可以分为两大类:一类是关于交易间隔的模型,它们认为较长的时间间隔意味着缺少交易活动,也代表着一个没有新信息产生的时期,因此时间间隔行为的动态性中含有关于日内市场活动的有用信息。基于这种观念,Russell 和 Engle(1998)使用了与分析波动性的 ARCH 模型相似的概念,提出了一个 ACD(Autoregressive Conditional Duration)模型来描述(交易活跃的)股票交易间隔的发展过程。随后,Zhang、Russell 和 Tsay(2001)对 ACD 模型作了扩展,用于分析金融高频数据中的非线性和结构性间隙问题。

另一类是关于交易间隔对交易价格变化的影响的模型,被研究对象的离散性和研究者对于"无变化"的关注,使得对日内价格变化的建模变得困难了。Campbell、Lo 和 MacKinlay(1997)曾对相关文献中所提及的若干计量模型进行了讨论,其中有两个在选择解释变量方面具有优势的模型值得关注。一个是 Hauseman、Lo 和 MacKinlay(1992)使用的规则概率模型(ordered probit model),它将交易的间隔作为一个影响逐秒价格变动概率的回归量,但是这个模型有其他的一些缺陷;第二个是 Rydberg、Shephard(1998)和 MacKinlay、Tsay(2000)的分解模型(decomposition model),作为一种替代方法,它将价格的变动分解为价格变动指数、价格运动方向和价格变动幅度(如果有价格变化)三个部分进行研究;这两个模型的主要区别是后者不需要对价格变化幅度作任何划分。相关的研究还有 Ghysels 和 Jasiak(1998)使用了一个关于不定期取值的金融数据的 ACD-CARCH 模型,发现在交易间隔的时间序列与收益波动的时间序列的变动中存在因果关系,尤其是日内交易间隔会对收益

波动中的意外事件有所反应。

第四个分支是关于金融高频数据统计特征的研究。在讨论金融高频数据如何应用的同时，对数据本身的统计特征也不能忽视。因为统计特征不仅是认识数据的基本依据，也是正确使用数据的首要前提。早期的研究表明，与低频金融数据（如月度数据）服从高斯分布的特征不同，金融高频数据是不稳定的，在较短期间内有着增长性的拖尾趋势（heavy-tailed），并且数值具有离散性的特点。相比较而言，近期对金融高频数据的统计分析则更为深入和具体。如 Jacquier、Polson 和 Rossi（1994，1995）的研究发现 S&P500 指数的日收益数据具有非正态性；Jobson 和 Korkie（1980）的研究表明在决定最优证券组合的输入变量的均值—方差模型中，方差—协方差/期望收益与最优组合的权重之间的映射是高度非线性的；Chopra 和 Ziemha（1993）对同样问题所作的研究指出期望收益的估计误差所带来的危害通常是方差估计中同样误差的 10 倍，是协方差估计中同样误差的 100 倍。在这些研究的基础上，Nich Pplson 和 Bernard Tew（2000）基于 S&P500 指数的数据建立了可变参数的证券组合框架，描述了对收益的多变量分布进行建模的几种方法，其中不仅利用了期望收益和方差—协方差矩阵的先验信息；并且在使用日收益数据进行估计时，给出了单个证券收益估计的上限和下限。还有 Thomas 和 Patnaik（2002）的研究，他们用 VR（Variance Ratio）检验对印度证券市场的证券价格之间的连续相关性作了分析，得出了在以 5 分钟为间隔的高频率数据基础上，所有股票都显示出了均值回归（mean-reversion）趋势的结论。

三、金融高频数据分析中遇到的问题研究

金融高频数据的特征虽然为认识市场微观结构提供了更为详细的信息，但也给相关的实证研究带来了前所未有的问题。目前，理论界虽然对这些问题有了一定的探讨并且提出了若干建议性的对策方案，但离问题的真正解决还相差甚远。因此在未来的研究中，这仍然是一个值得关注的问题。总的来看，金融高频数据的分析中所遇到的问题大致可以归纳为如下三类。

1. 数据问题

（1）不准确的时间（innaccurate times）。对每日数据来说，数据库中对每个观测值（如每日收盘价格）所记录的日历时间通常是准确的。相反，日内交易的记录时间却往往是不准确的。比如在一个采用公开喊价交易机制的金融市场中，交易数据要等到交易者的交易卡片进入计算机系统以后才做时间标记，这当中则可能会有几个小时的时滞。对金融高频数据来说，交易之间的间隔比较短，这种不准确性往往会造成交易或报价被记录到一个错误的间隔中，或者交易或者报价的时间序列不正确等问题。

（2）不正确的交易量（inaccurate volumes）。同样地，在采用公开喊价交

易机制的金融市场中由于单笔交易量较难观察到，在对其所建立的金融高频数据中，往往采用对单笔交易估计而非精确的交易量，从而就意味着用这些数据所作的研究是不可靠的。

(3)失时效的价格(stale prices)。实证研究通常需要现价时间序列，但除非价格形成过程是连续的，否则就无法得到这样的时间序列，而需要使用失时效的价格作为替代。所谓失时效的价格，指的是一段时间之前发生的交易价格。比如说，要得到一个按固定间隔(如每 15 分钟)观察的价格序列，因为在这样短的一段时间内也许不会有交易或报价出现，所以就只能用最近的价格作为替代。可是如果将这样的数据视为固定间隔取值数据的话，就会引起各种各样的偏差。比如，如果把不等间隔的数据视为等间隔的数据的话，就会高估后者的方差，并且收益的时间序列会表现出自相关性。

(4)缺省值(missing value)。用来计算收益的价格必须来自单独的交易或报价，在这里如何处理缺省值问题非常重要，因为它将影响作为结果的时间序列的统计特征。在每月或者每周数据中几乎不可能出现缺省值问题，而且对大多数金融证券来说通常每天至少会有一个交易(或报价)，所以每日数据一般也不会遇到这个问题。然而，在金融高频数据中(如时间间隔缩短为 1 分钟)缺省值却会时常发生，并且成为影响相关研究的一个实质性的问题。

2. 日内数据带来的市场微观结构的影响

(1)离散性(discreteness)。价格的离散性在取值范围很大的低频样本中不是个重要问题，因为它可以用一个连续过程作为很好的近似。但是对日内价格运动来说，离散性却是个严重问题，因为它可能一共只有五、六个观测值。缺少连续性暗示了按照连续间隔状态所建立的模型不能很好地代表数据，并且会导致一系列的统计问题，如有限依赖变量、拒绝随机性检验(因为它可能会带来微弱的负自相关)、增大估计的方差、带来价格变动分布中的峰度问题等等。

(2)季节性(seasonalities)。有关的实证研究已表明，在很多金融市场中都存在交易量、收益波动性、询报价差额的 U 型趋势和收益中的日内模式及自相关关系。由于这些现象会导致周末效应的消失、高估信息对收益波动性的影响以及会隐藏高频数据中的 ARCH 效应等，所以，对它们进行控制是相当重要的。

(3)询报价反弹(bid-ask bounce)。在低频数据中询报价差额对收益计算的影响很小。可是研究表明，在高频数据中，它却会造成收益中的负自相关关系。询报价差额是一个交易成本，它不仅会给基于套利的定价关系带来噪音并且造成算术收益和收益方差的高估；而且还会影响价格时间序列的动态性、价格逆转与延续性检验的效力及增加收益的波动性。

3. 统计与计量问题

(1)缺少正态性(lack of normality)。根据中心极限定理所推出的"金融市场的收益数据服从正态分布"的结论是有争议的。对于对数形式的收益来说，每个月的对数收益值等于这个月中每分钟收益值的总和，因而每月收益数据趋于正态分布。但是当交易间隔变得比较短时，正态分布的论点就失去了效力。有实证研究表明，随着交易间隔越来越短，收益的分布也会越来越偏离正态。非正态性之所以重要，不仅因为它会令很多标准统计检验失效，而且它也是建立一些模型如 Black-Scholes 期权定价模型和进行风险价值分析的重要基础。

(2)ARCH 效应。众所周知，在每日或更低频的收益数据中存在 ARCH 效应。关于波动性的建模和预测对金融工具的定价是很重要的。如对期权来说，Engle 和 Bollerslev 的 ARCH 模型就是对波动一致性进行估计的成熟方法。但是研究发现，金融高频数据中的波动一致性远远低于低频数据。如 Andersen 和 Bollerslev 用 1992～1993 年外汇现货交易中 US$-DM 的收益数据所作的研究表明，当交易间隔缩短为 90 分钟时，用 GARCH(1,1)模型所估计的波动一致性就消失了。

四、金融高频数据分析研究的展望及对我国应用的启示

目前，对金融高频数据的研究方兴未艾，在为已取得的成果而感到欣喜的同时，应当看到这些新数据所提出的问题远比它们所解决的问题多，比如说，关于驱动市场和资产价格行为的基本原理还没有解释清楚；对金融高频数据统计特性的认识还不够深入；价格和收益的波动性特征还是一个极大的困扰；还缺少能够对日内交易数据的离散性、季节性进行恰当描述的计量模型；也还没有对交易数据之间、交易市场之间的相互关系得出确定性结论；实证分析中还有若干等待探讨的数据问题、统计问题和计量问题，等等。这些问题有待于众多学科工作者的共同努力，才能得到很好的解决。当然，同任何新生事物的发展一样，关于金融高频数据的研究还有很长的路要走，但前景是光明的。

高频数据分析与市场微观结构理论是紧密联系在一起的。证券市场微观结构即证券市场的交易机制，是指证券交易价格形成与发现的过程与运作机制。市场微观结构的核心是价格发现功能，后者也是整个证券交易市场最核心的环节。市场微观结构理论主要包括两大类内容：一是关于价格发现的模型及其实证研究；二是关于市场结构与设计方面的理论研究与经济研究。证券流通市场的微观结构将影响市场价格波动、流动性以及潜在的投资者数量和交易量。这正是市场微观结构的意义所在。而高频数据分析是理解市场微观结构极为有效的手段。我国加入世界贸易组织后，证券市场的改革步伐必将加快，证券市场的微观结构也将面临重大而深刻的变革，随着市场微观结构理论研究的深入，尤其是对中国证券市场高频数据的实证研究，无疑将为我国

证券市场微观结构的改革提供有益的指导。通过对证券高频数据的分析，积极探索我国证券市场交易机制改进之道，有利于提高我国证券市场的竞争力和国际竞争地位。

如前所述，金融高频数据的取得及相关研究的进展，为更好地理解金融市场的微观结构开辟了一条新的途径。可以预期，国外学者的相关研究结论必将对我国金融市场的发展具有积极的借鉴意义。因此，有必要尽快开展对我国金融高频数据的分析工作。从国外的研究经验来看，以下问题应当是我们未来研究的重点：

(1)分析我国证券市场高频数据的形态特征，并与国际成熟市场高频数据的形态特征作出比较，利用我国证券市场的历史高频数据，对相关理论模型进行检验，研究日内价格模式的异同等。

(2)利用高频数据分析市场价格波动、流动性以及潜在的投资者数量和交易量，以便更好地研究我国证券市场的微观结构。

(3)探寻支配市场交易行为的机制，分析我国金融市场交易规则对市场的影响，进行科学的市场机制设计；了解市场运行的基本规律，制定科学的市场监管制度来指导金融市场有效、稳定地运行。

(4)掌握价格变动的规律并用适当的模型进行拟合，为进行价格预测及相关决策活动提供可靠信息，同时也可以用于遏制市场的不良投机行为，增强我国金融市场防范风险的能力。

(5)以金融高频数据为基础改进现有的资产定价模型，实现对资产的合理定价并促进我国衍生金融产品市场的形成，从而完善我国的金融市场体系，为投资者提供更多规避风险的金融工具。

金融高频数据分析除了可在我国证券市场应用外，还可用于我国外汇市场和期货市场等。作为一种先进的数据分析工具，高频数据分析迟早都将被我国的理论工作者和金融市场的管理者和投资者所接受。

注释：

①在 NYSEZ 中，1997 年 6 月 24 日以前，tick sice 为 1/8 美元，2001 年 1 月 29 日之前为 1/16美元。从 2001 年 1 月 29 日开始，所有 NYSE 和 AMEX 的股票都以十进制来交易。

②它们通常认为交易之间的时间间隔不会影响随后的交易价格和市场均衡。

参考文献：

[1]Charles A. E. Goodhart, Maureen 0'Hara. High frequency data in financial markets: Issues and applications[J]. Journal of Empirical Finance, 1997, 4, 73～114.

[2]Ruey S. Tsay. Editor's introduction to panel discussion on analysis of high-frequency data[J]. Journal of Business & Economic Statistics 2000, Apr, 139.

[3]Torben G. Andersen. Some reflections on analysis of high-frequency data[J]. Journal of Business & Economic Statistics, 2000, Apr, 146～153.

[4]Nicholas G. Polson, Bernard V. Tew. Bayseian portfolio selection: an empirical analysis of the S&P 500 index 1970～1996. [J]. Journal of Business & Economic Statistics, 2000, Apr, 164～176.

[5]Eric Ghysels. Some econometric recipes for high-frequency data cooking[J]. Journal of Business & Economic Statistics, 2000, Apr. 154～162.

[6]Robert. Wood. Markét microstructure research databases: history and projections[J]. Journal of Business & Economic Statistics, 2000, Apr, 140～145.

[7]Owain ap Gwilym, Charles Sutcliffe. Problems encountered when using high frequency financial market data: suggested solutions[J]. Journal of Financial Management and Analysis, 2001, 14(1), 38～51.

[8]Saji Gopinath. Number of transactions and volatility: an empirical study using high-frequency data from NASDAQ stocks[J]. The Journal of Finance Research, 2001, summer, 205～218.

[9]Richard Olsen. High frequency data——an essential resource[J]. IFC Bulletin, 2002, Dec.

A Study on the Present Status and Problems of Financial High-Frequency Data Analysis

CHANG Ning[1], XU Guo-xiang[2]

(1. *Department of Statistics, Shanghai University of Finance and Economics, Shanghai* 200433, *China*;
2. *Research Center for Applied Statistics, Shanghai University of Finance and Economics, Shanghai* 200433, *China*)

Abstract: In recent years, the analysis of financial high-frequency (HF) data has been one of the key and difficult issues in the western financial community and research circle. The paper first discusses the definition and characteristics of financial high-frequency data, analyzes the fundamental causes for its analysis, expounds the fields involved and probes into the problems encountered in its analysis. Finally, the paper makes a prospect on the development of the analysis of financial high-frequency data and the key issue to research in this field in China.

Key words: financial market; securities market; the analysis of financial high-frequency (HF) data; microstructure of market