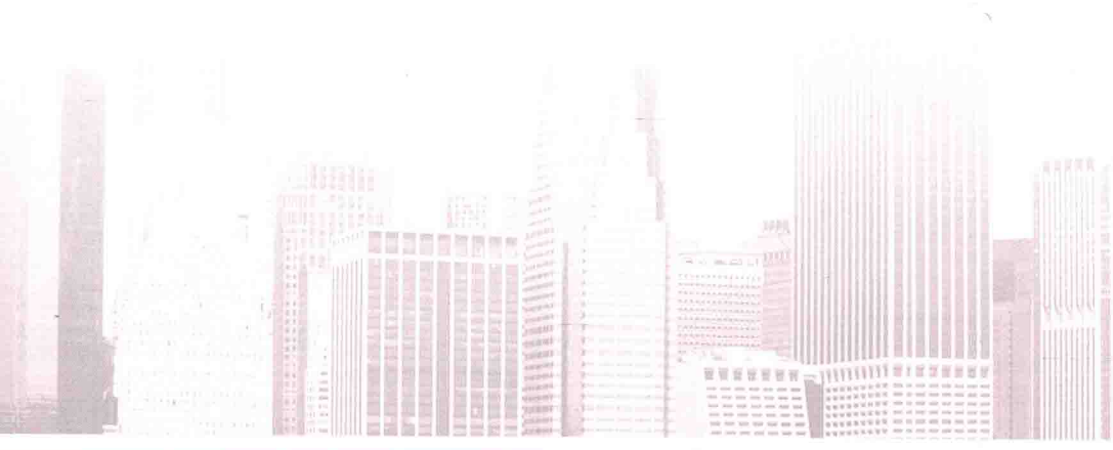
Theory and Methodology in Financial High-Frequency Data Mining Based on the Perspective of Statistics

统计学视角下的金融高频数据挖掘理论与方法研究

魏瑾瑞 著



中国社会科学出版社

东北财经大学统计学院资助出版

Theory and Methodology in Financial High-Frequency Data Mining Based on the Perspective of Statistics

统计学视角下的金融高频数据挖掘理论与方法研究

魏瑾瑞 著

中国社会科学出版社

图书在版编目（CIP）数据

统计学视角下的金融高频数据挖掘理论与方法研究/魏瑾瑞著. —北京：中国社会科学出版社，2015.6

ISBN 978-7-5161-5841-8

Ⅰ.①统… Ⅱ.①魏… Ⅲ.①金融—数据收集—研究 Ⅳ.①F830.41

中国版本图书馆 CIP 数据核字(2015)第 063898 号

出 版 人	赵剑英
选题策划	侯苗苗
责任编辑	侯苗苗
责任校对	周晓东
责任印制	戴宽
出 版	中国社会科学出版社
社 址	北京鼓楼西大街甲 158 号
邮 编	100720
网 址	http://www.csspw.cn
发 行 部	010-84083635
门 市 部	010-84029450
经 销	新华书店及其他书店
印 刷	北京君升印刷有限公司
装 订	廊坊市广阳区广增装订厂
版 次	2015 年 6 月第 1 版
印 次	2015 年 6 月第 1 次印刷
开 本	710×1000 1/16
印 张	17.5
插 页	2
字 数	306 千字
定 价	56.00 元

凡购买中国社会科学出版社图书，如有质量问题请与本社发行部联系调换

电话：010-84083683

前　言

随着技术的不断成熟，对金融数据观测的频率越来越细致，甚至可以实时跟踪交易数据并在精度上达到毫秒、微秒。这类数据有助于理解投资行为和交易过程的细节，同时也对经典的分析工具提出了挑战，比如，如何处理复杂的大规模数据集、跳跃成分以及伴随日内模式和复杂关联结构的随机交易间隔。

在物理和生物科学中，当分析的尺度降为分子或原子时，有些被略去的成分逐渐变得重要起来。金融市场亦如此，市场微结构在低频情况下可以忽略，但在高频数据中却是重要的；低频数据可以用几何布朗运动来近似，而高频数据却行不通。频率从日到分钟，与频率从月到日，是有本质区别的。

一般而言，金融高频数据分析主要涉及基本经验事实的归纳、市场微结构分析以及计量经济建模等几个方面。其中，根据 Herwartz（2006）的观点，高频数据建模至少可以分为三类：（1）价格离散变动建模（不考虑取样的时间维）；（2）固定时间间隔建模（间隔作为外生变量）；（3）随机交易间隔建模（间隔是交易的函数）。但考虑到等间隔的非同时性、微结构动态等因素，事实上最近的很多模型都兼顾到随机间隔的情形。

本书首先在回顾历史文献的基础上，界定高频数据的相关概念、研究其性质，并提出全书的总体分析框架；在数据准备的同时对“统计视角”做了必要的注释。接下来的内容分为两个部分：方法和理论探索，其中，前者在更一般的意义上讨论方法；后者则将重点放在高频数据及其理论基础上。最后给出全书结论。

方法探讨由四个章节组成，其中，第四章和第五章属于典型的探索性数据分析（Explotary Data Analysis，EDA）；第六章讨论波动率问题，提出协同波动率，它是一类模型自由的波动率估计方法；第七章以金融高频

数据交易方向推断为例，结合支持向量机提出的理论背景（统计学习理论），对支持向量机混合核函数的做法提出了异议。理论探讨部分包括两个章节，其中，第八章市场微观结构分析对金融高频数据的现实背景、运行环境以及相关理论和方法进行了深入研究；第九章是随机交易间隔分析，着重分析了信息与噪声的边界问题。具体地：

在第一章对金融高频数据相关研究领域文献梳理的基础上，第二章首先提出，金融高频数据不仅仅是作为一个优质的时间序列用来验证在以往粗糙信息下建立的经典理论与模型，因为金融高频数据不能单纯理解为时间序列，这样至少忽略了日内与日间两个不同维度各自所具有的分布特征。为此提出了序贯面板数据变换，得到一个看待金融高频数据的双重视角。其中，“$i(t)$视角”本质上是样本的细化，它分析的对象仍然以天为单位，只是每天的数据更细致而已；“$t(i)$视角”相当于对“交易日”的重复观测，它分析的对象就是这一天，关心的是短期行为（微观结构）。在第二章，我们还区分了“交易高频数据”与“高频交易数据”，其中，后者是对“高频交易”的记录，而前者很大程度上是对“一般交易”的实时记录。二者的共同点是对短期的关注。另外，采用高频数据验证市场有效性可以为高频交易是否存在获利机会提供佐证。接下来我们对金融高频数据的经验和理论特征分别予以考察。

第三章首先对“统计视角”做了解释，讨论了数据挖掘的统计学内涵以及区别于统计学的显著特征，指出了统计分析的本质属性是对数据的阅读（提取其中的信息和知识，这在一定程度上决定了理解数据背景或环境的重要性，统计分析离不开它所应用的土壤），最后着重从云计算的角度探讨了大规模数据处理的基本逻辑。

第四章从一个统一的框架考察了连续信号与离散信号之间的关系。在不含有微结构噪声条件下，基于数字信号处理探讨了连续信号离散化的理论基础，论证了采样的本质（对采样函数偏移后做基展开）。函数数据分析的一个重要步骤是将离散数据连续化（含有微结构噪声），研究了函数数据与面板数据、符号数据之间的异同，以及函数数据分析的基本原理，特别是对基展开的本质做了广泛的讨论。基展开（可以是正交基或非正交基）就是在基构成的子空间下求得相应的坐标（将波动分解为在各基方向上的波动），这相当于变换到时域以外的（频）域进行分析。插值与平滑都是函数逼近（拟合）问题，从一般意义上（度量空间）对此做了

规范分析。最后，我们用一个例子说明了函数数据分析如何有助于对金融市场行为细节的刻画。

第五章研究了希尔伯特—黄变换提出的理论背景和基本逻辑，并与傅里叶变换和小波变换做了对比；讨论了 IMF 的正交性，并从成分数据分析的角度研究了约束条件带来的影响。以金融高频数据为例进行实证分析，讨论了序列的分解与重构问题，并仿效时间序列加法因素分解，将非线性非平稳序列也分解为趋势、周期与随机波动。不同之处在于，这里的周期是可变的，即这里的分解是动态的，且分解的对象可以是非平稳非线性序列。

第六章在回顾时变（条件）波动测度的基本方法的基础上提出了一类模型自由的波动率估计方法。协同波动率强调波动所处的空间并非“真空”，而是考虑受扰于其他相关随机变量波动条件下的波动程度。而通常计算波动率是将变量抽离出来单独计算，或以自身历史为条件从动态的视角切入。协同波动率的构建基于相关分析和随机变量取值的频数（非实际取值），所以它具有对称性，同时不受取样频率所限，也有益于从概率分布的角度来探讨波动。与已实现波动率类似，协同波动率也会随平均组距减少（组数增加）而增加，这可能主要受微结构噪声的影响。

第七章评估混合核函数的有效性。注意到混合核函数方法并没有解决核函数的选择问题，只是将问题等价转换为权重参数的选择。同时该方法还需要分别为两个核函数确定参数，大大增加了算法的复杂程度，限制了支持向量机的泛化能力。事实上，调节核函数的参数对分类结果的影响要远大于选择什么类型的核函数，因此混合核函数方法实属“避轻就重”。实证分析表明，不同核函数对应的共同支持向量比例很高，存在很大程度的一致性，线性组合的意义并不大，这也是混合核函数方法无法有效提升分类性能的一个重要原因。因此有必要对支持向量机的混合核函数方法做进一步的深入研究，讨论混合核函数在支持向量机中的有效性。明确了核函数在支持向量机中的具体作用，继而从算法复杂度对泛化能力的影响以及信息重叠两个方面研究了混合核函数无法有效改进分类结果的原因。

第八章研究市场微观结构理论（market microstructure theory）。不同于传统理论着眼于长期均衡（忽视调整过程中的摩擦），市场微结构理论研究的主要是，在考虑微结构因素影响的条件下，有效均衡价格发现的机理，或向均衡或新均衡的转移动态过程；反过来，价格形成过程中渗漏出来的

信息对交易行为和策略有何影响；市场是通过价格发挥作用的，那么，进一步还可以讨论市场微观结构对市场效率和质量的影响，这涉及市场机制的设计与选择。在这一章，我们还对几种强调微观过程的方法（奥地利学派、芝加哥学派、行为经济学等）做了比较，并从一个综合的视角解释了日历效应和日内收益率一阶负相关等现象，特别是将日历效应推广到一般的间歇性时限情景中加以解释，但这种解释的视角是把交易者当做一个整体来研究的，为寻找其中的微观基础，我们还构造了一个博弈模型。

第九章通过经验分析验证了随机交易间隔存在很强的聚集性，其概率分布与指数分布相近，从而倒推出单位时间内的交易次数服从 Poisson 分布，这些都与经典的假定（如 ACD 模型扰动项服从指数分布，跳跃成分假定由 Poisson 过程驱动等）相吻合。同时推导了随机交易间隔下的收益率计算方法。事实上，尽管随机交易间隔含有重要的交易信息，但并非“字字玑珠”（受微结构噪声干扰），所以这里面有一个信息提取的问题。尽管在研究间隔分布时，噪声并不是一个重要的因素（被解释变量与噪声的概率分布是相同的），但是，变量之间的关系很可能被噪声掩盖。剔除噪声之后我们发现：（1）收益率对随机间隔的变化并不敏感；（2）价格与随机间隔之间可能存在非线性关系，但价格变动与随机间隔之间不存在显著关系；（3）交易量与随机间隔之间可能存在负相关关系。

目　录

第一章　绪论 …… 1

第一节　研究背景与意义 …… 1
第二节　国内外文献综述 …… 5
一　日内模式、随机交易间隔建模与市场微结构理论 …… 5
二　波动率、微结构噪声与最优取样间隔 …… 8
三　连续时间模型 …… 13
四　国内研究现状 …… 14
第三节　研究内容及创新 …… 15

第二章　金融高频数据挖掘的概念与统计特征 …… 18

第一节　基本分析框架 …… 18
一　时间序列：理解高频数据的起点 …… 18
二　序贯面板数据变换 …… 24
第二节　相关概念辨析 …… 29
一　高频交易数据 …… 29
二　交易高频数据 …… 35
第三节　典型统计特征 …… 41
一　基本描述 …… 41
二　经验特征 …… 41
三　理论特征 …… 46
第四节　本章小结 …… 52

第三章　数据准备及大规模数据集的分析逻辑 …… 54

第一节　数据挖掘的统计学内涵 …… 55

一 参数与非参数方法 …… 55
二 验证性与探索性分析 …… 56
三 渐进理论与统计学习理论 …… 56
四 数据规模：实录数据与系统收集数据 …… 58
五 再论数据挖掘与统计学 …… 59
第二节 统计分析的本质属性 …… 61
第三节 样本数据的来源与结构 …… 67
第四节 大规模数据集的分析逻辑 …… 70
一 定义及特征 …… 70
二 分析逻辑 …… 71
第五节 本章小结 …… 77

第四章 函数数据分析的基本逻辑及实证分析 …… 79

第一节 信号与随机信号 …… 79
一 信号的定义及分类 …… 79
二 随机信号的定义及分类 …… 79
第二节 连续信号离散化 …… 81
一 数字信号处理 …… 82
二 Shannon 采样定理 …… 82
三 采样的本质 …… 83
第三节 离散数据连续化 …… 86
一 函数数据、面板数据与符号数据 …… 86
二 函数数据分析的要点 …… 90
三 基本原理与步骤 …… 92
第四节 基展开（频域分析）的逻辑 …… 99
一 基展开的本质 …… 99
二 何为基 …… 99
三 两类重要的变换 …… 102
四 基函数的比较 …… 102
五 再论逼近问题 …… 108
第五节 基于 FDA 的日内结构分析 …… 111
一 序贯面板数据变换 …… 111

二 情形1（$N=48$，$T=218$） …… 113
三 情形2（$N=218$，$T=48$） …… 119
第六节 本章小结 …… 125

第五章 非平稳非线性序列分析的EMD方法 …… 126

第一节 传统方法及其比较 …… 126
第二节 HHT的基本思想 …… 128
第三节 EMD分解与原序列重构 …… 130
第四节 正交性检验与成分分析 …… 133
一 正交性检验 …… 133
二 成分数据分析 …… 135
第五节 本章小结 …… 137

第六章 一类模型自由的波动率估计方法 …… 139

第一节 典型特征对建模的启示 …… 139
第二节 历史波动率与隐含波动率 …… 141
第三节 波动率的基本估计方法 …… 145
一 ARCH族和SV族模型的基本逻辑（MEM模型） …… 145
二 用RV估计IV …… 148
第四节 协同波动率方法 …… 150
一 协同波动率的定义 …… 150
二 相关性与波动性的分解与关联 …… 152
三 数值模拟：取样频率与相关性对协同波动率的影响 …… 154
四 方差—协方差随取样频率增加而下降的事实（不含有微结构噪声） …… 156
第五节 实证分析 …… 160
第六节 本章小结 …… 162

第七章 对支持向量机混合核函数方法的再评估 …… 164

第一节 混合核函数的基本思路 …… 165
第二节 核函数在支持向量机中的作用 …… 166
第三节 算法复杂度对泛化能力的影响 …… 169

一 基于小样本的统计分析理念…… 169
二 影响支持向量机泛化能力的关键因素…… 170
三 模型选择的基本准则…… 173
第四节 信息重叠弱化了混合核函数的有效性…… 174
一 数据清洗…… 175
二 结果分析…… 176
第五节 本章小结…… 177

第八章 市场微观结构分析…… 180

第一节 市场微观结构理论概述…… 180
一 市场微观结构理论研究的主要内容…… 180
二 价格发现建模与市场有效性检验…… 187
第二节 日历效应的经济学解释…… 192
一 经验分析…… 192
二 博弈论视角…… 193
三 对拥挤现象的剖析…… 194
四 对相关性的剖析…… 194
第三节 微观方法论及其比较分析…… 195
一 奥地利学派与芝加哥学派…… 196
二 奥地利学派与行为经济学…… 202
三 个人与群体的行为逻辑…… 203
四 预期理论…… 205
五 市场过程…… 208
第四节 证券及证券市场的意义…… 209
第五节 本章小结…… 210

第九章 随机交易间隔分析…… 212

第一节 数据以高频记录的成本…… 212
第二节 随机交易间隔的基本特征…… 214
第三节 数据清洗中可能遇到的错误…… 215
第四节 信息与噪声在何处分界…… 218
一 概率分布与反演…… 218

二 更细致的分析…… 219
三 经济含义解读…… 221
第五节 随机交易间隔建模…… 227
第六节 本章小结…… 231

第十章 结论与展望…… 233

第一节 结论…… 233
第二节 展望…… 237

参考文献…… 239
后记·致谢…… 263

第一章　绪论

第一节　研究背景与意义

在 *An Introduction to High – frequency Finance*（2001）这本书的导论里有这样一段对话，大意是，“为什么把一生置于这危险的攀岩？因为山峰等在那里”（Because they are there.）。同样地，浩繁的高频数据也等在那里，等待我们去“攀登”（The feld of Statistics is constantly challenged by the problems that science and industry brings to its door.）。“上帝创造问题，并诱惑我们来解释。”

金融高频数据构成海量数据集，是大规模数据集的一个重要子集，越来越受到学界和业界的广泛关注。其中一个主要的原因是市场结构和交易过程的快速演变，这主要得益于技术进步以及由此推动的交易系统的深远发展、交易所之间的竞争加剧以及日内交易活动日渐频繁。

然而，高频数据本身并不是新事物，地质、气象、工厂生产线、实验观测、高峰时刻的超市、车站和机场、金融市场等各领域的高频数据俯拾皆是。但为何数据泛滥发生在今天？是今天产生的数据果真比过去多了吗？或许是因为我们越来越擅长记录了。但擅长记录是好事吗？一层层的数据，像埋葬自己的墓穴。我们把过多的精力放在了捕捉和存储数据，而后束之高阁，却忽略了对已有数据的“开采挖掘”（mining）和“蒸馏提纯”（distilling）。贪婪？动物大都只存一个冬季的食物。长期眼光，还是不自信的表现？

随着技术的不断发展，不仅各领域记录数据的时间尺度越来越精细，而且也使存储与处理类似的大规模数据集成为可能。过去因记录和存储等方面的限制只能有选择性地存储（如精简的古文、有影响力的文献），而

现在则是泥沙俱下等权记录，论语和一行微博同样载入史册。网络公开课、大规模开放网络课程（massive open online course，MOOC）、开放存取仓储（Open Access Repositories）等也已经逐渐开始对传统的教学和科研产生冲击。我们已身处大数据的洪流，而且是“被卷入”，一如对现代通信工具的被迫回应，特别是目前异常活跃的增速。一方面，数据记录大量产生（数据尾气，商业记录、行政记录等）；另一方面，不仅原始数据而且数据的复制品（报纸、杂志、网页等）也需要存储空间，信息累积的方式也从竹简、纸张、软盘过渡到硬盘、网盘等效率更高的存储媒介。

2008 年 9 月 4 日刊出的《自然》杂志（*Nature*）以“big data”作为专题（封面）探讨了环境科学、生物医药、互联网技术等领域所面临的大数据挑战。2011 年 2 月 11 日，《科学》杂志（*Science*）携其子刊《科学—信号传导》（Science Signaling）、《科学—转译医学》（Science Translational Medicine）、《科学—职业》（Science Careers）专门就日益增长的科学研究数据进行了广泛的讨论。格雷还进一步提出科学研究的“第四范式”（the fourth paradigm）是数据（数据密集型科学，data - intensive science），不同于实验、理论和计算这三种范式，在该范式下，需要“将计算用于数据，而非将数据用于计算”。这种观点实际上是将数据从计算科学中单独区别开来了。在《大数据时代的历史机遇：产业变革与数据科学》（2013）一书中，鄂维南院士也提道：“大数据在科学领域的表现是数据科学的兴起，数据科学将成为科研体系中的重要组成部分，并逐渐达到与物理、化学、生命科学等自然科学分庭抗礼的地位。”然而数据科学目前只是多个相关学科“拼接”起来的一个新兴学科，尚未形成完整的学科框架体系；同时，也鲜有统计学视角下的探讨。

其实追溯起来，股票交易至少有两百多年的历史，但直到 20 世纪 50 年代，日收盘价还要等到次日才可知晓。而今天，在流动性良好的市场中，单个交易日的超高频数据量与 30 年按日统计的交易数据量相当。作为金融市场的雏形，外汇市场可以说是最大也是最为复杂的金融市场，如交易约束条件（所在时区、工作时间、交易成本、信息获取方式、交易制度等）迥然相异。[①] 从实务部门的角度来看，股指期货是一个值得研究

① 大量文献以 NASDAQ、NYSE、东京交易所、巴黎证交所等高频数据库为研究对象，但是不能简单地将这里的一些经验分析推广至中国，因为这里面交易机制等各方面都有差异，如是否存在做市商、是否有订单簿、开闭市时间、最小最大价格变动、最小交易量等。

领域，如光大银行（2010）曾出过类似的研究报告。中国量化投资研究院林健武教授在发布“2012 中国量化投资半年报”时提到，“中国可以做高频的就是股指期货和商品期货，这里面发挥了很多量化投资的高等技巧，频率也越来越高”。CSMAR 中国证券市场高频交易数据库（2008Level－1）的使用指南也提出了一些亟待实证和理论分析的研究领域，如中国证券市场的微观结构、交易规则、交易者特征等。事实上十年前 Goodhart 和 O'Hara（1997）对金融市场高频数据所带来的一些问题和实际应用曾予以概述，他们指出，“目前大部分实证文献仍然保持相当的描述性……大部分聚焦于造市商如何从交易中学习（获取信息），然后又怎样影响价格和报价”。Robert Wood（2000）不仅讨论了金融高频数据量的快速增长趋势及其在市场微结构研究中的应用问题，而且还对所用的不同频率数据库的组织形式和特征等进行了分析。

研究金融高频数据的一个直接的意义在于，高频数据是否能提供一些低频数据所不能够提供的信息。如果将其单纯作为一个优质的时间序列来看待，意义可能并不大；就我们目前的研究来看，高频数据（high－frequency data，HFD）的意义可能更多地在于短期分析，比如市场微观结构方面的探索，特别是超高频数据（ultra－high－frequency data，UHFD；tick－by－tick；transaction－by－transaction）所提供的大量交易细节为实证微结构理论提供了丰富的凭据。然而在目前金融高频数据的研究过程中，尚存一些认识上的误区，如混淆了低频数据、高频数据与超高频数据：周橙（2009）认为高频数据是等间隔的，只是加细了抽样间隔，与低频数据没有本质区别，仍然可以采用 ARCH 族模型。这是错误的，因为高频数据有异步交易（non－synchronous trading；asynchronous trading）问题，[①] 而且高频数据也并不是简单加细了取样间隔，而是有其特殊的分

① 以股票交易为例，不同的股票，交易时间（频率）并不相同，也可能在某个时间区间内 A 股票比 B 股票交易更频繁，或 A 股票比 B 股票对某个信息更敏感，然而记录时却以相同的频率取样（如 10 分钟、一天）。这在低频数据中并不是问题，因为分秒差异相对于天而言是可以忽略的，但在高频数据中却不容忽视。比如噪声在低频数据里可以忽略，而在高频数据里却有显著影响。这就好像是在较小的尺度上（比如短期），你可能犯错，导致出现一个凸点，但是在较大的尺度上（长期），这个错误的凸点可能就被“磨圆”了、没有那么明显了。事实上，即便是单只股票，其交易频率在一天内的不同时段也不尽相同（如可能开闭市时段较密集，中间时段较松散）。所以，如果采用固定取样频率，结果可能会有偏差。如 Tsay（2010）中提到的 Lo 和 MacKinlay（1990）的研究，该研究考虑了非同步交易中的交易中止现象（non－trading），结果表明，在有效市场中，非同步交易可以导致观测收益率具有“伪负一阶自相关”。

析目的，如对短期行为的考察（市场微结构理论）。退一步而言，即便是加细了取样间隔，也并不是越细信息就越充分，因为还存在微结构噪声的扰动。为此，需要从更严格的意义上澄清、界定和辨析低频数据、高频数据与超高频数据，进而从统计学和数据挖掘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，也有利于激发统计方法的进一步拓展、深入和创新。

研究金融高频数据挖掘的意义还在于：（1）高频数据可以连续地记录金融市场上发生的变化，因此高频数据具有很多采用低频数据所无法观察到的重要特征，如微结构噪声的干扰、跳跃成分、日内模式、离散变动（discreetness）、[①] 随机交易间隔等，而加总或稀疏取样则可能会掩盖或漏掉这些特征所蕴含的信息。（2）理论上来讲，统计分析通常要求达到一定的样本容量，而低频数据通常难以满足，所以为了保持较高的自由度，模型应尽量简洁（变量或待估参数，要尽可能少），但这样可能会遗漏重要的信息，而高频数据可以为构造理想的模型（可能是复杂的）提供一个很好的起点。（3）从实证分析上来讲，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，[②] 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，也为理解金融市场价格形成机理、市场组成结构与市场交易机制等方面提供丰富的素材。（4）在宏观经济分析中，长期与短期视角下的结论不同，金融高频数据分析在很大程度上是为了探索短期行为特征（如日内波动）与市场微结构（如价格发现）。这种研究尺度的变换改变了分析的单位或尺度，高频数据扩大了我们的视野，就好像用“显微镜”可以看到肉眼看不到的东西，从而更深刻地理解一些现象。比如，慢镜头重播运动员的动作，可以通过这种“放大”找出错误以便于矫正。当然，也不可避免地会遇到一些无特征尺度的现象，即在不同的时间尺度上表现出相似的性质。（5）研究金融高

① 最小价格变动（tick）和最小交易量（lot）。

② 在低频限制下，要想获得一个样本容量足够大的样本，时间跨度就必须相应地足够大，这样导致一个问题是：在这么长的时间跨度里，所研究的对象还是同一个吗？在时间序列分析中，这种时变性也就是所谓的非平稳性。而高频数据可以让我们在一个既定的时间区间（平稳区间）内获得我们想要的样本容量。

频数据一个很重要的目标是减少交易成本或增加交易的灵活性，提高风险管理的能力。（6）目前研究有通过高频数据来验证证券市场的有效性，研究信息传导机制、波动溢出效应、风险测量和异常值检测等问题。对高频数据的研究也有助于回答如下问题：应该披露多少的信息给市场、极端波动对市场流动性的影响有多大、造市商（market maker）是必要的吗、如何利用金融高频数据来进行资产组合的选择，等等。

第二节 国内外文献综述

在文献综述部分，我们有选择性地考察金融高频数据几个相关研究领域的脉络，在梳理各个研究分支的同时也以长期关注这个领域的专家学者为线索。

一 日内模式、随机交易间隔建模与市场微结构理论

关于金融（超）高频数据至少可以追溯到二十五六年前，早期主要是对日内模式的考察，如 Wood 等（1985），McInish 和 Wood（1985a；1985b；1992），Harris（1986），Admati 和 Pfleiderer（1988）等，近期文献，如 Heston 等（2010）。Baillie 和 Bollerslev（1990），Aggarwal 和 Gruca（1993）以及 Andersen 和 Bollerslev（1994；1997），Kim 等（1998）等的研究在不同的金融市场上都发现了交易高频数据的日内模式（如波动率、交易量、交易频率、买卖价差等变量往往都会表现出 U 形特征）。最近的一篇论文提出了这样一个问题：剔除公共信息之后交易量与波动率仍呈 U 形模式吗？Eaves 和 Williams（2010）考察了东京谷物交易所（Tokyo Grain Exchange）的期权数据，发现日内 TGE 交易量呈 U 形而日内波动率呈 L 形，而这些模式在剔除公共信息之后几乎不存在了，所以私人信息可能并不是日内模式的根源。

因为价格变动只能是最小单位（tick）的整数倍，所以还存在价格离散波动问题，如 Harris（1994）。目前已有的建模方法有离散选择模型（discrete choice model；multiple choice model）、排序选择模型（ordered choice model）等。事实上如果考虑对日内价格波动建模，可能事先需要做一些恰当的变换，将价格波动放大，因为通常日内价格存在惰性，再加上熔断机制等限制，所以日内价格本身波动非常小，而建模的目的是为了