学校编码：**10384** 密级__________

学号：**15420100153752**

厦门大学

博士学位论文

统计学视角下的金融高频数据挖掘理论与方法研究

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

魏瑾瑞

指导教师姓名：谢 邦 昌 教授
专 业 名 称：统 计 学
论文提交日期：2013 年 4 月
论文答辩时间：2013 年 5 月
学位授予日期：2013 年 月

答辩委员会主席：________
评 阅 人：________

2013 年 6 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行)》。

另外，该学位论文为(　金融高频数据挖掘方法及应用研究　)课题（组）的研究成果，获得（　　　　　　　　　　　）课题（组）经费或实验室的资助，在（　　　　　　　　　　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）:

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　　年　　月　　日解密，解密后适用上述授权。

（　　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘要

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，而高频数据则行不通。频率从日到分钟，与频率从月到日，是有本质区别的[1]。

一般而言，金融高频数据分析主要涉及到：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）[2]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。

本文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、并提出全文的总体分析框架；在数据准备的同时对文题做了必要的注释。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。最后给出全文结论。

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了（如，可得每天的平均价格、最高价、最低价等，而不只是收盘价；再如，高频时间序列建模所使用的更为细致的时间序列数据）；“$t(i)$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。另一方面，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。接下去我们对金融高频数据的经验和理论特征分别予以考察，其中，（1）经验特征主要有：平均日内价格呈 U 型，但平均已实现波动率并不呈 U 型；高频收益率仍呈尖峰厚尾，但经已实现波动率调整之后则近似正态分布；价格变动的基本特征是离散的、跳跃的、惰性的（矩形水平方向），同时交易的价格也可能变化（矩形的垂直方向）……显然这些因素都增加了日内价格波动建模的困难。（2）理论特征方面主要是证明了，相对于波动率而言，为什么在高频数据中均值或均值矫正并不是很重要；市场微结构噪声如何导致日内平均收益率一阶负相关和方差高估，并且随着取样频率增加（微结构噪声突出）尤为显著；市场微结构噪声如何引出最优取样间隔的权衡；为什么采用高频数据建立连续时间模型需要包含跳跃成分等类似问题。

第三章首先对论题做了解释，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（非严格正交，统计正交），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成分、周期成分与随机成分。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；同时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。

第七章从统计学习理论（支持向量机的逻辑背景）和经验分析（金融高频数据交易方向推断）两方面质疑了混合核函数的做法。（1）尽管各种核函数秉性各异，但不同的核函数得到的共同支持向量的比例很高，因而对结果的影响并不显著，此时采用混合核函数是得不偿失的（大量计算换来了微小的精度提升）。（2）从支持向量机提出的逻辑背景（统计学习理论）来看，SVM 是建立在 VC 维和结构风险最小化基础之上以期获得最好的推广能力，特别是当样本量有限或比较

小时，欲提高学习机的推广能力，自然就需要控制算法的复杂程度（使 VC 维成为一个可以控制的量），所以从这个角度来讲，混合核函数也不是被鼓励的做法。（3）事实上，退一步来讲，模型的优劣也并不体现在其复杂程度上，因为模型并不是现实的复制，而是现实的抽象和简化。我们应该学会在既真实又易于处理之间权衡得失。

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，这涉及到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。

关键词：金融高频数据；数据挖掘；函数数据分析；协同波动率；混合核函数；市场微观结构；随机交易间隔

Abstract

As technology continues to mature, the frequency of financial data observed became finer and even real-time tracking of transaction data on the accuracy of milliseconds and microseconds. These data can help to understand the investor's behavior and the details of the transaction process, but also challenges the classic analysis tools. For instance, how to deal with the complex large-scale data sets, jump and irregularly spaced observations with intraday patterns and complex dependence structure.

In the physical and biological sciences, when the analysis scale is reduced to the level of molecules or atoms, something ignored becomes important. So as financial markets: market microstructure can be ignored in the low-frequency case, but it is important to the high-frequency data. The reason is that different frequency scales with its unique characteristics is not self-similarity. Such as low-frequency data can be modeled by Geometry Brownian motion, but high-frequency data does not work. Note that the frequency from days to minutes is totally different from months to days.

In general, financial high-frequency data analysis mainly concerned: the summarization of empirical facts, the market micro-structure analysis, and econometric modeling. According to the views of Herwartz (2006), there are at least three categories of econometric model: (1) price discrete changes (without regard to the time), (2) the fixed time interval (interval as exogenous variables), (3) random interval (interval is a function of the transaction). However, taking into account the non-synchronous trading and microstructural dynamic, many recent models use random intervals.

Firstly, based on literature review, we discriminate some related concepts and study the characteristic of their own. The overall framework also is presented. Secondly, there is data preparation and necessary note of the title. Thirdly, Passages remained are divided into two parts: methodology and theoretical exploration. Finally, we give conclusions of the full text.

The Methodology is made of four chapters: Chapters IV and V is the typical exploratory data analysis; Chapter VI discusses the volatility and proposes a model-free estimator. Chapter VII gives a dissent from mixed kernel function for support vector machine, based on a case of inference of transactions direction and the theoretical background of support vector machine (Statistical Learning Theory). The theoretical exploration includes two chapters: Chapter VIII is the market microstructure analysis, which studies the practice, operating environment as well as the theory and method; Chapter IX considers the random interval, especially the boundary of information and noise.

Based on the literature review of chapter I, We point out in chapter II that the financial high-frequency data can not merely be understood as a time series of high quality to verify the classical theory and model established under the previous rough information. Because, at least, it ignore the distribution characteristics of days and intradays when the financial high-frequency data being simply treated as a time series. So we propose the sequential panel data transformation and get a dual perspective: "*i(t)*" is essentially the refinement of sampling, with days being analytical unit but more detailed; "*t(i)*" is equivalent to repeated observations of the trading day, with days being analytical object and short-term behavior (microstructure) being concerned.

In chapter II, we also distinguish "(trading) high-frequency data" from "high-frequency trading data", of which the latter is the record of high-frequency trading, while the former is largely the record of general real-time trading. The two have in common is a focus on the short-term. Moreover, the verification of efficient market hypothesis using high-frequency data can provide evidence of the existence of profit opportunities for high-frequency trading. Then we investigate the empirical and theoretical characteristics of financial high-frequency data. (1) the empirical characteristics: average intraday price has a U-shape, but the average realized volatility is not a U-shape; high frequency rates of return show a fat tail, but has been approximate normal distribution with the realized volatility adjustment; basic characteristics of price changes is discrete, jump, inert (rectangular horizontal direction), possible changes in synchronous trading(rectangular vertical direction) ... Obviously, these factors have increased difficulties of modeling the intraday price volatility. (2) the theoretical characteristics: volatility is more important than mean or mean correction in the high-frequency data; how the market micro-structure noise causes negative first-order correlation，overestimated variance and optimal interval; Jump should be considered in continuous-time model.

Chapter III explains what is the statistical perspective, the statistical connotation of data mining and the difference between data mining and statistics. We also explore the logic of large-scale data processing from the perspective of cloud computing.

Chapter IV examines the relationship between the continuous and discrete signal from a unified framework. Without microstructure noise, we explore the theoretical basis of discrete of the continuous signal based on digital signal processing, and demonstrate the nature of sampling. Continue the discrete data is an important step in functional data analysis (with microstructure noise). We study the relationship between the functional data, panel data and symbolic data. Also, the basic principle of FDA has been explored, especially for the nature of the base expansion. The base expansion (orthogonal basis of non-orthogonal basis) is essentially equivalent to obtain coordinates of sub-space consisting of the base (decompose fluctuations into sub-fluctuations in the base direction), as transform into another domain. At last, we illustrate with an example of FDA to help characterize the detailed behavior of financial markets.

Chapter V studies theoretical background and basic logic of Hilbert-Huang Transform，compared with Fourier transform and wavelet transform. IMF is not strictly orthogonal but Statistics orthogonal. We also examine the impact of the constraint condition in the view of composition data analysis. Empirical analysis of financial high-frequency data as an example, we discuss the decomposition and reconstruction of the series. The non-linear non-stationary series can be decomposed into trend, cycle and stochastic components. The difference from additive factors decomposition is that here the cycle component is not a constant and the series can be non-linear or non-stationary.

Chapter VI proposes a model-free estimator of volatility, co-volatility. Before that we give a review of the measurement of (conditional) volatility. Traditional measurements are either separateness or on its own historical conditions. However, co-volatility emphasizes the space is not vacuum, but disturbed by other kinds of volatility. Based on correlation analysis and the frequency of the random variable, co-volatility is symmetric, not limited by the sampling frequency, and also benefit to explore fluctuations from the perspective of the probability distribution function. Similar with realized volatility, co-volatility will increase with frequency, which may be mainly affected by the microstructure noise.

Chapter VII queries if the mixed kernel function is appropriate in view of statistical learning theory (the logical background of support vector machine) and empirical results (inference of financial high-frequency data transactions direction). (1) The empirical analysis shows that, despite of the different kernel functions with their own properties, there has no significant effect on the results. Why？We found that this is mainly because of a high proportion of common support vector, making the results to a large degree of consistency. Under the circumstance, mixed kernel function is not worth the candle in the analysis. (2) According to the philosophy of SVM (Statistical Learning Theory), mixed kernel function is not encouraged. Because we should control the complexity of the model if the generalization ability is concerned, especially with the small sample. (3) In fact, to say the least, the quality of model does not depend on the complexity. Because the model is not the copy of reality, but rather to abstract and simplify it. Therefore, we should learn to tradeoff between complying with the reality and grasping the important aspect of the reality.

Chapter VIII is the market microstructure theory. Unlike traditional theories focus on the long-run equilibrium (ignore the friction in the adjustment process), the market microstructure theory considers (1) the microstructure in the effective equilibrium price discovery mechanism or the dynamic process to the new equilibrium; (2) how the information leaked out impacts strategies of traders; (3) how the microstructure impacts the market efficiency and quality. In this chapter, we also compare several methodologies centered on micro-process (Austrian School, the Chicago School, behavioral economics, etc.). From a comprehensive perspective, we explain the negative first-order correlation and the calendar effect especially extend the calendar effect to the general scenario of intermittent. Because this explanation treats the traders as a whole, we construct a game model to find the microstructure.

Chapter IX verifies that random trading intervals have strong cluster and exponential probability distribution which means that the number of transactions in unit of time have Poisson distribution. All that coincides with the classic assumptions of the ACD models and continuous time model. We also derive the returns of random trading intervals. In fact, the random trading interval contains important information about the transaction, but not every word has its count (influenced by micro-structure noise). So the information extraction should be taken into account. Note that the noise in the distribution of the random intervals is not important (explained variable have the same distribution as the noise), but disturbance in the relationship between variables. After eliminated noise, we found that (1) the rate of return is not sensitive to changes in the random interval; (2) there may be non-linear relationship between price and random intervals; (3) there may be negative relationship between trading volume and random intervals.

Keywords: financial high-frequency data; data mining; functional data analysis; co-volatility; mixed kernel function; market microstructure; random trading interval

目录