王亚楠 著

金融实时数据分析方法

F

Financial Real Time Data Analysis Method

经济管理出版社

本书出版受河北科技大学应急管理研究中心项目、河

Financial Real Time Data Analysis Method

金融实时数据分析方法

王亚楠 著

经济管理出版社
ECONOMY & MANAGEMENT PUBLISHING HOUSE

图书在版编目（CIP）数据

金融实时数据分析方法/王亚楠著. —北京：经济管理出版社，2015.2
ISBN 978-7-5096-3639-8

Ⅰ. ①金… Ⅱ. ①王… Ⅲ. ①金融—实时数据采集—分析 Ⅳ. ①F830.41

中国版本图书馆 CIP 数据核字（2015）第 039447 号

组稿编辑：杨　雪
责任编辑：杨　雪　许　艳
责任印制：黄章平
责任校对：张　青

出版发行：经济管理出版社
（北京市海淀区北蜂窝 8 号中雅大厦 A 座 11 层　100038）
网　　址：www. E-mp. com. cn
电　　话：（010）51915602
印　　刷：北京易丰印捷科技股份有限公司
经　　销：新华书店
开　　本：720mm × 1000mm/16
印　　张：12.75
字　　数：200 千字
版　　次：2015 年 2 月第 1 版　　2015 年 2 月第 1 次印刷
书　　号：ISBN 978-7-5096-3639-8
定　　价：49.00 元

前　言

金融实时数据是与金融低频数据、高频数据相比较而言的。实时数据和后两者最大的区别是：前者的时间间隔是时变的，后两者是等时间间隔的。这种采集信息的差异使实时数据保存的市场信息更多，对市场状况的反映更真实。随着计算机技术的迅猛发展，人们可以利用的存储空间越来越大，数据处理能力越来越强，这些能力足以保证人们实时采集金融实时数据。

在计量经济学近30年的发展过程中，以2003年诺贝尔经济学奖得主Engle和Granger为代表的计量经济学家，创建了系统化的时间序列分析方法，这些方法可以保证人们有效分析处理低频数据和高频数据，但在处理实时数据时，这些方法存在明显不足。因此，寻求新的计量方法，有效处理实时金融数据，是近年来计量经济学研究的一个新的发展方向。

在实践方面，由于各个国家股票市场交易制度的差异，所存储的实时数据也具有不同的特征。西方国家股票市场以报价驱动为主，其主要研究也是围绕这类市场展开。而我国股票市场属于与之对应的另一类交易市场——指令驱动市场，它包括集合竞价和连续竞价。不同的交易机制，导致不同的市场微观结构，也造成了实时数据的不同特征。这种特殊性要求在处理实时数据过程中，既要寻求分析方法突破，也要结合国内股票市场的特殊性。

本书通过对现有的金融数据分析方法比较研究，在分析方法上寻求创新，并利用构建的金融实时数据模型来分析国内股票市场实时数据的信息含量，以期对交易者行为作出合理解释。对金融实时数据模型研究，有助于优化市场信息，为市场监管者和投资者提供有益的决策参考和理论依据。

全书共分9章：第1章简单介绍了金融数据的基本特征及常用分析方法；第

2章介绍了金融数据分析的基本理论与方法；第3~4章介绍了低频数据分析常用模型；第5章分析了实时金融数据统计特征；第6~8章介绍了常用的金融实时数据分析模型；第9章介绍了金融实时数据分析模型在国内股票市场中的应用。

本书的撰写工作得到我的博士指导老师吴祈宗教授的帮助。是吴祈宗教授把我引入到这个十分宽广的学术领域，开启了我的学术生涯，在此对吴祈宗教授深表谢意。同时还要感谢我的同门张肖肖博士对本书提供的帮助。此外，本书还引用了其他作者公开出版的一些重要成果，在此一并表示感谢。

本书的相关研究受到河北科技大学应急管理研究中心、河北科技大学博士基金（QD201306）资助。

由于作者水平有限，书中难免存在不足之处，敬请读者批评指正。

目　录

1 金融数据简介

当今社会，金融与人们生活息息相关，2008 年华尔街金融海啸波及全球，亿万美元资产瞬间化为乌有，损失殃及千万家庭。数日间，许多家庭生活水平倒退数十年，对整个人类社会发展造成重大影响。但这样的事件并非偶然，在 10 多年前，还发生过类似的东南亚金融危机。面对如此重大的危机事件，人们应该尽可能提前获取征兆信息，及时做出防范，以化解危机，降低损失。这使人们对金融市场的数据分析和信息处理日益重视，以期能获取有效信息，改进市场功能，防患于未然。

近年来，金融数据分析一直是人们研究的一个热点问题。一方面是因为金融数据具有自身的特点，如随机性、受外界因素影响较大等特点，且大部分金融数据为高维、复杂、动态数据，因而如何准确地从这些数据中挖掘有价值的规则是一个难点；另一方面是因为对金融数据的分析可以为金融企业或金融投资者带来很多有价值的信息。对金融数据进行分析不仅有利于金融企业或机构了解自己目前的运营情况，获得有价值的商业信息，防范金融风险；而且有利于金融投资者了解金融市场的本质，更好地进行金融投资。

金融数据一般分为三类：金融低频数据、金融高频数据、金融实时数据（金融超高频数据）。金融低频数据（Low Frequency Data）是指以等于或大于一天的等时间间隔采集的数据。金融高频数据（High Frequency Data）是指以小时、分钟或秒为采集频率而采集的数据。金融实时数据（Ultra High Frequency Data）是指交易过程中实时采集的数据，这些数据不仅包括实时价格与交易量信息，还包括交易发生时间间隔，交易发生之际市场交易指令簿实时状态信息。实时数据和前两种数据的最大区别是：实时数据的时间间隔是时变的，而前两者是等时间间

隔的。这种采集信息的差异使实时数据保存的市场信息更多，对市场状况的反映更真实。随着计算机技术的迅猛发展，人们可以利用的存储空间越来越大，数据处理能力越来越强，这些能力足以保证人们实时采集金融实时数据。对这些金融数据进行分析通常有两类方法：一类是时间序列分析方法；另一类是数据挖掘方法。本书主要讨论第一类方法，这类方法以数理统计模型为基础，通过假设、参数估计、检验等方法得到描述时间序列规律的模型。在计量经济学近 30 年的发展过程中，以 2003 年诺贝尔经济学奖得主 Engle 和 Granger 为代表的计量经济学家，创建了系统化的时间序列分析方法，这些方法可以保证人们有效分析处理低频数据及一些高频数据，但是对于实时数据的分析与处理方法则需要不断完善，相关领域也是目前研究的热点。

1.1 金融时间序列分析

时间序列分析是统计学研究的一个重要分支，它的研究对象——时间序列是一类重要的复杂数据对象。时间序列根据其研究依据的不同，可有不同的分类：按所研究的对象的维度，可分为一元时间序列和多元时间序列；按时间的连续性，可分为离散时间序列和连续时间序列两种；按序列的统计特性，可分为平稳时间序列和非平稳时间序列两类；按序列的分布规律，可分为高斯型（Guassian）和非高斯型（Non-Guassian）时间序列。所有时间序列的基本点就是每一个序列包含了产生该序列的系统的历史行为的全部信息。

时间序列的含义根据分析的角度不同而不同：从统计意义上来讲，所谓时间序列就是将某一指标在不同时间上的不同数值，按照时间的先后顺序排列而成的数列，这种数列由于受到各种偶然因素的影响，往往表现出某种随机性，彼此之间存在着在统计上的依赖关系；从数学意义上来讲，时间序列就是对某一过程中的某一变量或一组变量的样本观测数据；从系统意义上来讲，时间序列就是某一系统在不同时间（地点、条件等）的响应。

金融时间序列作为时间序列的一类，同样也具有时间序列的一切共性，对金融时间序列进行分析，从数量的角度揭示其发展变化规律或从动态的角度刻画其变化规律，进而对其进行预测就显得尤为重要。只有正确地预测未来，才能做出正确的决策。在这个经济飞速发展的时代，对于事物随时间变化的理解将是十分宝贵的知识。对金融时间序列进行分析，深入洞悉其变换的内在原因，是获得这一知识的有效途径。

广义地讲，将某种金融随机变量按出现时间的先后顺序排列起来称为金融时间序列。从现实世界的角度看，金融时间序列就是指在一定时期内按时间先后顺序排列的金融随机变量。金融时间序列最显著的特征就是其与“时间”紧密相连。一般来说，金融时间序列变量，有时也简称为金融时序变量，由两个明显的要素组成，即时间跨度和序列的频率。

1.2 收益率

金融市场中资产的标价方式一般是用价格来表示，但是对金融市场进行定量分析时经常需要把价格变换成收益率。把支付利息的时间长度作为一个周期，计算这段时间内的收益率叫作单周期收益率。单周期收益率的计算有两种方式：一种是简单收益率；另一种是连续复利收益率。

收益率的大小与单周期的时间长度有关，为了比较不同时间长度的投资的收益率大小，需要把收益率标准化为年收益率。把投资周期短于 1 年（或支付利息的时间长度短于 1 年）的收益率标准化为年收益率有三种方法：简单年收益率、复利年收益率（也称作有效年收益率）和连续复利年收益率（也称作对数收益率）。

1.2.1 单周期收益率

假设在某个时点 T_0 进行初始投资 P_0，到时刻 T_1 获得收入 P_1。在 T_0 到 T_1 之

间没有任何的现金流，T_0 到 T_1 的收益率称为单周期收益率。因为以该时间长度为一个周期获得收益。单周期收益率分两种：简单收益率和连续复利收益率，定义如下：

简单收益率：$R = \frac{P_1 - P_0}{P_0}$

连续复利收益率：$r = \ln(P_1) - \ln(P_0)$

简单收益率与连续复利收益率之间的关系是：$r = \ln(1 + R)$

1.2.2 多周期收益率

如果投资长度不止一个周期，则需要计算多周期收益率。如图 1.1 所示，初始投资额 P_0，t 期资产总额 P_t，期间无任何收益。

图 1.1 投资期限与资产价格

则 t 期收益率为：

$$R(t) = \frac{P_t - P_0}{P_0}$$

则 t 期连续复利利率为：$r(t) = \ln(P_t) - \ln(P_0)$。

单期收益率与多期收益率之间的关系为：

$$1 + R(t) = \frac{P_t}{P_0} = \frac{P_t}{P_{t-1}} \frac{P_{t-1}}{P_{t-2}} \cdots \frac{P_1}{P_0} = (1 + R_t)(1 + R_{t-1}) \cdots (1 + R_1)$$

$$\begin{aligned} r(t) &= \ln(P_t) - \ln(P_0) \\ &= [\ln(P_t) - \ln(P_{t-1})] + [\ln(P_{t-1}) - \ln(P_{t-2})] + \cdots + [\ln(P_1) - \ln(P_0)] \\ &= r_t + r_{t-1} + \cdots + r_1 \end{aligned}$$

1.3 收益率分布性质

要研究资产收益率，最好是从它们的分布性质开始。目的是弄清不同资产、不同时间收益率的表现。考虑 N 个资产，持有这 N 个资产 T 个时间周期，如 t = 1，…，T。对每个资产 i，r_{it} 表示它在 t 时刻的对数收益率。所要研究的对数收益率为 $\{r_{it};\ i=1,\ \cdots,\ N;\ t=1,\ \cdots,\ T\}$。也可以考虑简单收益率 $\{R_{it};\ i=1,\ \cdots,\ N;\ t=1,\ \cdots,\ T\}$。

1.3.1 统计分布及其矩的回顾

简短地回顾一下统计分布的一些基本性质和随机变量的矩。R^k 表示 k 维欧几里得空间，$x\in R^k$ 表示 x 是 R^k 中的点，考虑两个随机向量 $X=(X_1,\ \cdots,\ X_k)'$ 和 $Y=(Y_1,\ \cdots,\ Y_q)'$。$P(X\in A,\ Y\in B)$ 表示 X 在子空间 $A\subset R^k$ 中、Y 在子空间 $B\subset R^q$ 中的概率。

(1) 联合分布

函数：$F_{X,Y}(x,\ y;\ \theta)=P(X\leqslant x,\ Y\leqslant y)$，$x\in R^p$，$y\in R^q$ 是带参数 θ 的 X 与 Y 的联合分布，其中不等号“≤”是分量对分量的运算。x 和 y 的分布由 $F_{X,Y}(x,\ y;\ \theta)$ 刻画。如果 X 和 Y 的联合概率密度函数 $f_{X,Y}(x,\ y;\ \theta)$ 存在，则

$$F_{X,Y}(x,\ y;\ \theta)=\int_{-\infty}^{x}\int_{-\infty}^{y}f_{x,y}(x,\ y;\ \theta)dzdw$$

这时，X 和 Y 是连续随机向量。

(2) 边际分布

X 的边际分布是：$F_X(x;\ \theta)=F_{X,Y}(x,\ \infty,\ \cdots,\ \infty;\ \theta)$

这样，X 的边际分布可通过对 Y 求积分得到。Y 的边际分布也可类似得到。

如果 k = 1，X 是一个一元随机变量，其分布函数为：$F_X(x)=P(X\leqslant x;\ \theta)$，称为 X 的累积分布函数 (Cumulative Distribution Function，CDF)。一个随机变量

的CDF是不减的（即对 $x_1 \leqslant x_2$ 有 $F_X(x_1) \leqslant F_X(x_2)$，且 $F_X(-\infty)=0$，$F_X(\infty)=1$）。对给定的概率 p，使 $p \leqslant F_X(x_p)$ 的最小实数 x_p 叫做随机变量 X 的 p 分位点，即：

$$x_p = \inf_x \{x \mid p \leqslant F_{X(x)}\}$$

（3）条件分布

给定 $Y \leqslant y$ 的条件下 X 的条件分布为：

$$F_{X|Y\leqslant y}(x;\ \theta) = \frac{P(X \leqslant x,\ Y \leqslant y)}{P(Y \leqslant y)}$$

若所对应的概率密度函数存在，则给定 $Y = y$ 的条件下，X 的条件密度为：

$$f_{x|y}(x;\ \theta) = \frac{f_{x,y}(x,\ y;\ \theta)}{f_y(y;\ \theta)}$$

其中边际密度函数 $f_y(y;\ \theta)$ 由下式得到：

$$f_y(y;\ \theta) = \int_{-\infty}^{\infty} f_{x,y}(x,\ y;\ \theta)dx$$

联合分布、边际分布和条件分布之间的关系为：

$$f_{x,y}(x,\ y;\ \theta) = f_{x|y}(x;\ \theta) \times f_y(y;\ \theta)$$

这个相等关系在时间序列分析中经常用到（如在进行最大似然估计时）。X 与 Y 是相互独立的随机变量。当且仅当 $f_{x|y}(x;\ \theta) = f_x(x;\ \theta)$，这时 $f_{x,y}(x,\ y;\ \theta) = f_x(x;\ \theta)f_y(y;\ \theta)$

（4）随机变量的矩

一个连续型随机变量 X 的 l 阶矩定义为：

$$m_l' = E(X^l) = \int_{-\infty}^{\infty} x^l f(x)dx$$

其中“E”表示期望，f（x）是 X 的概率密度函数。一阶矩称为 X 的均值（mean）或期望，它表示的是分布的中心位置，记为 M_x。X 的 l 阶中心矩定义为：

$$m_l = E[(x - M_x)^l] = \int_{-\infty}^{\infty} (x - \mu_x)^l f(x)^{dx}$$

只要式中的积分是存在的，X 的 l 阶中心矩存在。二阶中心矩可度量 X 取值的变化程度，称为 X 的方差（Variance），记为 σ_x^2。方差的正平方根 σ_x 称为 X 的标准差。一个正态分布是由随机变量的头两阶阶矩唯一决定的。对其他分布，可