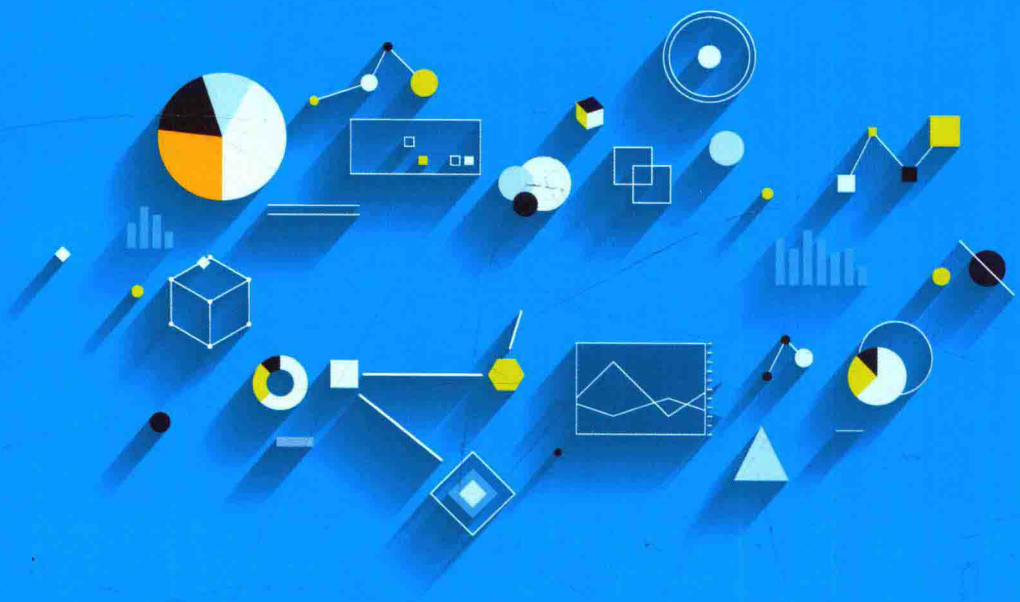
金融大数据统计方法与实证

杨　虎　杨玥含　编著



- 注重理论知识与实际问题的紧密结合
- 应用案例大多采用金融证券市场数据
- 网络共享案例及配套 R 软件程序代码

科学出版社

金融大数据统计方法与实证

杨　虎　杨玥含　编著

科学出版社

北　京

内 容 简 介

全书共九章, 内容包括大数据概述、聚类分析、判别分析、主成分分析、因子分析、线性模型、回归诊断、有偏估计、变量选择. 各章都有丰富的案例分析, 为加深读者对每章内容的理解, 将每章的练习分为理论和实证部分, 书后附有参考答案, 为使书中案例贴近数据的应用实际, 采用了获取方便的证券市场高频数据, 并使用国际通用的 R 软件进行数据收集、处理、加工和分析, 便于读者自己动手和实际应用需要. 全书内容讲解简明扼要, 注重应用, 让读者从收集数据开始, 掌握数据收集、整理和大数据统计分析的全过程.

本书可作为统计学、经济学、管理科学、计算机科学等相关专业本科生的教材和教学参考书, 也可作为相关专业硕士生的教材和案例分析参考书. 书中大部分内容也可供大数据分析应用的大学生、研究生、教师、科研人员和统计工作者参考.

图书在版编目(CIP)数据

金融大数据统计方法与实证/杨虎, 杨玥含编著. —北京: 科学出版社, 2016.6

ISBN 978-7-03-048488-8

Ⅰ. ①金… Ⅱ. ①杨… ②杨… Ⅲ. ①数据处理-应用-金融统计-统计方法-研究 Ⅳ. ①F830.2-39

中国版本图书馆 CIP 数据核字 (2016) 第 121605 号

责任编辑: 王胡权 / 责任校对: 邹慧卿
责任印制: 白 洋 / 封面设计: 迷底书装

科学出版社 出版
北京东黄城根北街 16 号
邮政编码: 100717
http://www.sciencep.com

大厂书文印刷有限公司 印刷
科学出版社发行 各地新华书店经销
*
2016 年 6 月第 一 版 开本: 720×1000 1/16
2016 年 6 月第一次印刷 印张: 10
字数: 201 600

定价: 39.00 元

(如有印装质量问题, 我社负责调换)

前　言

1952 年, 芝加哥大学的马科维兹 (Markowitz) 首次采用股票收益率历史数据的方差, 作为风险衡量指标, 并指出与证券市场的整体运行相关联的宏观系统风险不能通过投资分散化加以消除, 称为不可分散风险. 马科维兹在投资者效用最大化的基础上, 将复杂的投资决策问题简化为一个风险 (方差)–收益 (均值) 的二维问题, 即在相同的期望收益条件下, 投资者选择投资风险最小的证券 (组合), 或在相同的投资风险下, 选择预期收益率最大的证券 (组合). 开统计方法应用于金融市场之先河. 1978 年, 西蒙斯 (Simons) 开发了许多数学模型用来进行分析和交易, 这些基本上是自动完成. 他用计算机编程建立模型分析股票价格, 从而能进行很轻松的交易并获利. 这些模型是建立在海量的数据基础上的, 所以具有可靠性并可进行实际预测, 1989~2009 年, 他操盘的大奖章基金平均年回报率高达 35%, 较同期标普 500 指数年均回报率高 20 多个百分点, 比金融大鳄索罗斯和股神巴菲特的操盘表现都高出 10 余个百分点. 即便是在次贷危机爆发的 2007 年, 该基金的回报率仍高达 85%. 西蒙斯成就了世界上最伟大的对冲基金之一: 大奖章基金. 大数据的历史相对较晚一些, 2008 年年末, 大数据才得到部分美国知名计算机科学研究人员的认可, 但在 2013 年, 大数据就已经风靡全球, 成为一个时代的符号. 我们早在 2002 年开始从事金融数据挖掘研究和教学, 2011 年正式给本科生开设证券数据统计建模与实证分析课程, 2013 年结合大数据发展, 给硕士生和博士生开设了金融大数据统计方法与实证的课程.

本书是作者历年在金融大数据统计应用研究与教学讲稿内容的系统总结, 相关的讲义在重庆大学和中央财经大学的研究生与本科生教学和专题讨论中已反复使用过, 并且积累了大量的金融大数据资料、R 程序、PPT 讨论专题习作等 (见附录中的 QQ 群). 在内容讲授与学习上做了大胆的尝试, 加强了学生的参与程度, 通过公开讨论课、课程论文等方式激发学生的学习热情, 提升动手能力和自学能力.

本书在取材和写作上, 有如下特色:

1. 注重理论知识与实际问题的结合, 注重运用统计知识解决大数据问题的能力, 注意吸收国内外优秀教材的长处, 图文并茂, 使该书通俗易懂, 可读性强.

2. 和传统教材相比, 本书引入了最新的研究成果, 应用案例大多来自金融证券市场, 数据量大, 变量众多, 传统的分析方法和计算手段受到很大限制, 为了培养学生的动手能力, 本书借助于开源的免费统计软件 R, 对来自市场的第一手数据进行统计分析, 在实践中体会大数据统计方法的思想和应用.

3. 在例题与案例编排上尽可能做到前后衔接, 从而让一些例子能够前后对照, 突出各种统计方法的优良性质, 所有案例都用 R 软件进行了计算. 为了全书的可读性, 较难的证明较少涉及, 必要的内容放入练习并通过练习解答给出. 程序通过互联网免费下载, 便于正文的流畅阅读, 也方便程序的编辑和使用. 程序的使用可增加运用统计知识解决实际大数据问题的感性认识, 对本书所授知识产生浓厚兴趣和动手愿望, 从而变被动学习为主动学习.

4. 对统计学等相关专业的学生而言, 必要的理论证明和逻辑推导训练还是必不可少的. 我们在本书练习的编排上兼顾了理论题目和实践性题目, 照顾各类读者的实际需求.

由于本书很多内容至今仍然是国际学术研究领域的前沿研究方向, 成为众多的优秀学者和青年学生的热门选题, 如果作为教材可深可浅, 尤其是实证研究, 需要更多的利用课余时间进行金融证券理论知识和 R 程序知识的扩充. 本书无论理论研究和应用需要, 都要求读者结合 R 软件进行大量的应用案例分析. 在应用问题的分析和编程实践中体会金融大数据统计分析方法的广博内容.

本书部分案例来自授课班级学生的课余习作, 不一一说明, 特此一并致谢.

本书是国家自然科学基金项目 (编号：11171361) 的应用研究和金融实证研究组成部分.

由于编者水平所限, 疏漏乃至不当之处在所难免, 恳请国内同行及广大读者不吝赐教.

编　者

2016 年 1 月 19 日

目　录

第 1 章　大数据概述

数据是我们通过观察、实验或计算得出的结果. 数据有很多种, 数据也可以是文字、图像、声音等, 但最简单的就是数字.

今天, 整个人类社会已经进入到大数据时代, 这不单单是称谓的不同, 这里所谓 "大" 通常是指数据规模, 一般指在 10TB(1TB=1024GB) 规模以上的数据量. 大数据同过去的海量数据的区别就在于其基本特征可以用 5 个 V(Volumes、Variety、Velocity、Veracity 和 Value) 来总结, 即样本大、维数高、速度快、实时强、价值足.

样本大

用计算机业界的话是体量大, 数据体量巨大, 从 TB 级别, 跃升到 PB 级别, 通常在 10TB 以上, 比如证券市场的超高频分笔数据体量就非常大, 路网视频监测系统的数据也异常庞大. 但这个大仍然是相对而言, 和统计学里的大样本还不能相提并论, 这里的规模还是属于有限样本范围.

维数高

很多说法是数据类型繁多, 如网络日志、视频、图片、地理位置信息等. 其实, 在统计学上, 前者是 n, 是样本容量, 这个就是 p 了, 是变量数. 大数据框架下, 高维是普遍特征, 比如基因测序, 风险对冲, 都存在 $p > n$ 的情况, 有些甚至 p 关于 n 呈现指数增长趋势, 这就属于超高维概念了.

大数据认识上的误区就在于此, 样本仍然受到很多因素的制约, 出现局部的样本不足仍然是普遍的.

速度快

处理速度快. 1 秒定律. 或秒级定律, 就是对处理速度有要求, 一般要在 1 秒或几秒时间范围内给出分析结果, 时间太长就失去价值了. 能否抓住金融市场上的机会, 处理速度是关键. 这个速度要求是大数据处理技术和传统的数据挖掘技术最大的区别.

实时强

在数据量非常庞大的情况下, 也能够做到数据的实时处理. 这在量化投资领域非常重要, 信息时效相当有限, 很多时候需要快速决策, 抓住转瞬即逝的市场机会.

又如登月火箭发射, 选择的时间窗分秒都不能差, 后期的变轨也一样; 又比如云南鲁甸地震, 据网上消息, 地震台成功在 60 秒前作出了预报, 但由于通信的限制, 未能及时传递给每一个人, 最后仍然造成很大的伤亡.

价值足

通常的解释是价值密度低. 以视频为例, 连续不间断监控过程中, 可能有用的数据仅仅有一两秒. 这在大数据时代, 虽然数据进入 PB 规模, 但有用的信息仍然是稀疏的.

大数据技术是指从各种各样类型的巨量数据中, 快速获得有价值信息的技术. 解决大数据问题的核心是大数据技术. 目前所说的 "大数据" 不仅指数据本身的规模, 也包括采集数据的工具、平台和数据分析系统. 大数据研发目的是发展大数据技术并将其应用到相关领域, 通过解决巨量数据处理问题促进其突破性发展. 因此, 大数据时代带来的挑战不仅体现在如何处理巨量数据从中获取有价值的信息, 也体现在如何加强大数据技术研发, 抢占时代发展的前沿. 统计方法是最重要的大数据技术之一, 尤其针对金融大数据, 统计方法的应用往往起到事半功倍的效果.

古时候要理解万物皆数估计很困难, 但现在如果稍稍探究一下计算机的工作原理, 高清电视的制作与播放, 美轮美奂的图片传输方式, 就知道我们看见的形形色色的图片和影视画面, 与我们看见的数字没有两样, 都是将它们数字化后进行编码、传输、解码和重现的.

接收的图像如果不重现, 看见的就是一堆 "杂乱" 无章的数字. 如果不知道原始编码, 要通过统计分析将大数据变成图像是相当费力的, 很多时候甚至是根本不可行的.

本书就是要尝试通过统计分析还原大数据海量信息中的有用信息, 其难度和现实需求都是不可估量的.

统计学家的根本任务就在于此, 由于数据的局限, 很多时候, 我们的统计研究背离了这条主线, 陷入数学游戏的幻境, 随着人类社会快速的步入大数据时代, 统计学家必须正视现实, 不拘泥于传统的统计方法, 探索更多更新的方法与工具, 研究大数据海量信息中隐藏的统计规律性.

早期的统计学者要想获得数据, 往往会很艰难, 牛顿的万有引力定律是从哪里来的? 来自苹果掉地? 这仅仅是唯美的传说, 是文学家的臆想罢了, 其实是来源于天文观测, 那可是经年累月的观察和积累, 同样的观测今天还存在; 又比如水文地质观测, 农业选种实验, 学者的一生如果必须面对这样漫长数据收集的学科太残酷了, 需要好几年到好几十年的观测才能获得第一手完整的观测资料, 人生能有多少个几十年呢?

今天的统计学家无疑是幸运的, 数据比比皆是, 甚至已经泛滥成灾, 这是机遇也是挑战, 但我们需要的数据从哪里来, 统计年鉴吗? 那些不是自己亲手调查得来的数据分析起来可得小心, 统计中的模型很依赖条件, 如果条件是错误的, 结果可想而知, 对统计学家而言, 数据如果有问题, 结果也是荒谬的.

真实又方便获取的数据有没有? 有! 而且很多, 比如金融市场的行情数据, 眼

观为实, 这是你可以方便获取的海量数据源.

为什么一些人舍近求远, 不肯关注信手拈来的金融证券市场数据呢? 这个市场太真实, 太现实也太残酷了, 市场是检验模型和公式的唯一客观标准, 理论再优美, 也要经得起市场的检验, 方法再先进, 不能带来利润也是水中月镜中花.

研究的人少, 研究的水平低直接导致很多明显的错误, 比如股市分析家们经常会说资金流入了多少, 流出了多少? 这其实是市值的增减问题, 资金进出始终是平衡的, 有流出, 一定有同等的流入, 这是常识.

仅就中国沪深股市而言, 截至 2012 年 9 月 3 日, 仅 A 股就有 2437 支股票, 市值 21.218 万亿, 而 2 月 26 日是 27.718 万亿, 7 个月蒸发了 6.5 万亿, 看看 2012 年 9 月 3 日中午 11:30 瞬间的证券数据表, 有 2437 行, 116 列 (加上扩展行情), 共 282692 个数据, 这可是每秒都更新的数据表, 这还仅仅是沪深 A 股. 可见金融市场每天的信息量确实惊人!

本书针对金融大数据尤其是证券和汇率数据, 介绍各类相关的统计分析方法, 并结合 R 程序进行实证研究. 为了方便后续章节对这些数据的分析, 我们先介绍一些符号和概念.

一、大数据的数字特征

如果有用的信息存在于大数据之中, 分析大数据中包含的有用信息及主要特征是很有意义的. 我们可以通过这些特征的描述, 去了解整个大数据. 这就是大数据的数字特征.

对于 p 个随机变量组成的向量 $(X_1, X_2, \cdots, X_p)$, 如果获得 n 次观测, 得到 n 个观测样本, 设 $1 \leqslant i \leqslant n$, 第 i 个样本记为 $(x_{i1}, x_{i2}, \cdots, x_{ip})$, 从而样本的第 j 个分量的**均值**(mean) 定义为

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, j = 1, 2, \cdots, p, \tag{1.1}$$

样本的第 j 个分量的**方差**(variance) 为

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2, j = 1, 2, \cdots, p, \tag{1.2}$$

样本的第 j 个分量与第 k 个分量的**协方差**(covariance) 为

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), j, k = 1, 2, \cdots, p, \tag{1.3}$$

$X = (X_1, X_2, \cdots, X_p)$ 的均值记为 $\bar{x} = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p)$, 称矩阵 $S = (s_{jk})_{p\times p}$ 为样本的协方差矩阵. 样本的第 j 个分量与第 k 个分量的样本**相关系数**(correlation

coefficient) 为

$$r_{jk}=\frac{s_{jk}}{s_j s_k}, j,k=1,2,\cdots,p. \tag{1.4}$$

称 $R=(r_{jk})_{p\times p}$ 为样本**相关矩阵** (correlation matrix, 又称皮尔逊 (Pearson) 相关矩阵).

在大数据的今天, 样本矩阵 $C=(x_{ij})_{n\times p}$ 无论行列往往会非常巨大, 虽然给数据收集和整理增添了很多琐碎的工作, 但给统计分析带来了便利, 由于数据充足, 对获取统计规律更加便利和可靠, 大样本性质也确保了进行大数据统计分析的准确性.

在 R 软件 (软件介绍见附录) 中, 我们可以借助于 Excel 软件①的. csv 文件直接读取数据, 这给数据处理带来了极大的方便. 可以通过免费的证券行情软件中的数据管理菜单, 使用数据导出功能, 直接将数据导出到 Excel 文件, 打开后选择 "另存为" 将文件转化为. csv 文件格式存储, 可能会有不兼容的提示, 选择 "是" 即可.

命令: C=read.csv(file=file.choose( ), head=T) 将会打开一个选择窗口, 找到你存储的. csv 数据文件的位置, 执行后, 数据就放入变量 (样本矩阵) C 中了, 后续分析, 只要调用这个变量就能提取相关数据进行分析了. 如果打开 R 软件后, 通过菜单改变工作目录到当前目录 (或通过 R 命令 setwd("D:/R/Data") 改变工作目录), 可以用命令: C=read.csv("2011Dhs300.csv", header=T), 把当前目录中的文件读到矩阵 C 中.

下面我们采集从 2011 年 8 月 8 日到 2011 年 12 月 30 日的沪深 300 指数及其 300 支成分股的收盘价 (可以通过 R 程序对采集的 300 个成分股日线数据进行自动整理, 形成需要的数据文件, 详见案例 6.3 的数据收集整理说明). 在这段时间, 有的股票并不是每一天都开盘, 对于缺失的数据分别用前一天的收盘价补齐. 部分数据见表 1.1.

表 1.1　股票数据

沪深 300 指数	沪深 300 成分股日收盘价						
Y	X_1	X_2	X_3	X_4	X_5	$\cdots$	X_{300}
2793.9	16	8.06	18.65	15.55	7.42	$\cdots$	4.24
2798.19	16.03	8.3	18.59	15.67	7.26	$\cdots$	4.29
2824.12	16.09	8.44	18.9	15.93	7.29	$\cdots$	4.32
2866.92	16.43	8.55	19.6	16.2	7.36	$\cdots$	4.37
2875.36	16.64	8.43	19.59	16.2	7.46	$\cdots$	4.37
2917.88	17.13	8.5	19.87	16.67	7.6	$\cdots$	4.54

① Excel2007 以后的版本没有列数 256 的限制, 较低的版本或免费的 WPS 表格列数都不能超过 256 列.

续表

沪深 300 指数	沪深 300 成分股日收盘价						
Y	X_1	X_2	X_3	X_4	X_5	$\cdots$	X_{300}
2897.58	17.05	8.36	19.57	16.47	7.54	$\cdots$	4.51
2886.01	17.05	8.22	19.27	16.14	7.41	$\cdots$	4.5
2834.25	16.73	8.14	18.95	16.02	7.32	$\cdots$	4.42
2807.66	16.6	8.14	19.01	15.72	7.28	$\cdots$	4.39
2777.79	16.42	8.08	18.69	15.45	7.21	$\cdots$	4.37
2821.00	16.64	8.22	19.16	15.7	7.34	$\cdots$	4.52
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
2307.93	15.3	7.34	11.2	9.07	5.06	$\cdots$	3.97
2311.36	15.31	7.33	10.67	9.16	5.03	$\cdots$	3.98
2345.74	15.59	7.47	10.9	9.06	5.13	$\cdots$	4.04

该数据表一共有 301 列 99 行. 另存到文件 2011Dhs300.csv 中. 通过前述的数据读入命令读入 R 软件数据矩阵 C 中, 然后用 colMeans(C) 计算每列均值, sapply(C, sd), 计算每列标准差, cov(C), cor(C) 函数可以方便的计算出样本协方差矩阵和相关矩阵.

案例 1.1　为了方便演示, 我们选择 X_1, X_2, X_3 三个变量的前 12 个样本, 数据见表 1.1, 试计算样本均值、样本协方差和样本相关系数, 并用皮尔逊相关性检验确认这三个变量是否相关?

建立数据文件有两种方法, 一种是直接把这些数据拷贝存入单独的文件再通过 R 读入, 二是通过数据矩阵 C 裁剪, 使用 R 命令 C1=C[1:12, 1:3] 即可. 通过计算, 容易得到

$$\bar{x}_1 = \frac{1}{12}\sum_{i=1}^{12} x_{i1} = 16.5675, \quad \bar{x}_2 = \frac{1}{12}\sum_{i=1}^{12} x_{i2} = 8.2867, \quad \bar{x}_3 = \frac{1}{12}\sum_{i=1}^{12} x_{i3} = 9.1542.$$

样本协方差矩阵为

$$S = \begin{pmatrix} 0.1545 & 0.013 & 0.1247 \\ 0.013 & 0.0282 & 0.0518 \\ 0.1247 & 0.0518 & 0.182 \end{pmatrix},$$

相关系数矩阵

$$R=\begin{pmatrix} 1 & 0.1964 & 0.7433 \\ 0.1964 & 1 & 0.7226 \\ 0.7433 & 0.7226 & 1 \end{pmatrix}.$$

然后根据相关系数进行皮尔逊相关性检验, 当 (X_1, X_2) 服从二元正态分布且 $\rho = 0$, 有

$$t = \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}} t(n-2), \tag{1.5}$$

因此, 对于假设检验 H_0: $r_{12} = 0$, H_1: $r_{12} \neq 0$, 当显著水平取 0.05 时, 利用 R 函数 cor.test( ), 对于例中 (X_1, X_2) 的相关系数 $r_{12} = 0.1964$ 算得 (1.5) 的值 $t = 0.6336$, 在原假设成立的条件下, 出现这么大的 t 值的概率为 0.5406, 这个概率只有低于 0.05 才认为显著, 从而判定相关系数显著异于零, 可见, 这里的结果是不显著, 接受原假设, 认为相关系数为零, 即 X_1 与 X_2 不相关.

同理, 对变量 (X_1, X_3) 进行皮尔逊相关性检验, 算得 $t = 3.51$, 概率为 0.0056, 远低于 0.05 的显著水平, 因此认为 X_1 与 X_3 相关. 对变量 (X_2, X_3) 进行皮尔逊相关性检验, 算得 $t = 3.31$, 概率为 0.0079, 远低于 0.05 的显著水平, 因此认为 X_2 与 X_3 相关.

此外, 可以算出相关系数的置信区间, 比如 r_{23} 的 95% 的置信区间为 (0.25, 0.92).

二、大数据的图表示

直观的图形一直是有效的分析手段, 从定积分简单绘制的积分区域图形, 到几何概率的图形描述都给实际的计算带来很大的帮助, 但三维以上的图形就没有形象逼真的解决办法了. 这里介绍一些方法来刻画大数据的直观特征, 这就是大数据的图表示.

1. 轮廓图

轮廓图由如下作图步骤完成:

(1) 作直角坐标系, 横坐标取 p 个点, 以表示 p 个变量;

(2) 对给定的一次观测值, 在 p 个点上的纵坐标与之对应的变量取值成正比;

(3) 连接 p 个点得一折线, 即为该次观测值的一条轮廓线;

(4) 对于 n 次观测值, 每次都重复上述步骤, 可画出 n 条折线, 构成 n 次观测值的轮廓图.

可以使用画轮廓图的 R 函数 outline( ) 进行此类计算.

案例 1.2　考虑海南省 19 只 A 股 2012 年 6 月 29 日的数据, 选择 11 个指标 (变量), 如表 1.2, 为了减少量纲的影响, 对各个指标通过量纲变化进行了放大和缩小. 试画出这 11 个变量的轮廓图.

表 1.2　海南股票数据*

名称	最新	涨跌	涨幅	换手率	内盘	外盘	振幅	量比	委比	市净率	市值
海虹控股	5.81	2	3.570	1.800	6.0679	6.2368	6.950	1.13	−5.342	4.2	3.96694
海南海药	19.3	6.1	3.260	1.030	1.0277	1.1565	4.980	0.83	−1.976	3.56	4.07751
海德股份	6.42	1.7	2.720	1.830	1.1859	1.2193	6.240	1.01	3.502	4.87	0.84543
新大洲A	4.97	0.7	1.430	0.760	2.5337	3.074	4.490	0.37	−1.242	2.79	3.65596
海马汽车	3.26	0.6	1.870	0.320	2.1237	3.2027	2.190	1.7	−0.909	0.82	5.35119
亚太实业	6.33	5.8	10.090	10.660	19.3804	11.5906	11.830	14.59	10.000	14.42	1.83878
罗牛山	4.83	0.1	0.210	1.600	7.4561	6.5865	4.360	1.34	−2.349	2.48	4.25029
华闻传媒	6.18	0.4	0.650	0.260	1.8604	1.6762	2.770	0.83	−2.620	3.12	8.40091
海南高速	3.46	0.2	0.580	1.110	5.7488	4.7723	2.030	3.52	0.721	1.34	3.29091
海峡股份	12.18	−2.7	−2.170	9.830	10.4588	7.7077	4.660	23.3	4.384	2.92	2.25161
海南瑞泽	15.8	−7.4	−4.470	39.700	7.2633	6.235	6.890	14.9	1.550	2.6	0.5372
康芝药业	12.48	2.3	1.880	1.510	0.4926	0.5719	4.000	0.43	−1.632	1.41	0.88221
神农大丰	15.61	2.2	1.430	1.450	0.4885	0.4753	2.530	0.29	3.474	1.91	1.0365
海南航空	4.88	1	2.090	0.340	5.0023	6.2737	3.560	0.38	−5.296	1.47	16.32455
海南椰岛	10.28	2.7	2.700	1.300	2.7187	3.0166	4.500	1.79	−2.587	6.95	4.54408
广晟有色	65.93	35.8	5.740	1.610	1.6271	2.3965	8.920	1.26	−4.977	34.71	16.44294
正和股份	5.54	0.5	0.910	0.760	4.3726	4.8295	2.730	1.03	−4.955	3.47	6.73677
中海海盛	4.38	0.7	1.620	5.280	12.5694	18.1067	8.820	24.77	−4.876	1.25	2.54616
海南橡胶	6.87	1.1	1.630	1.590	6.1018	6.6039	4.590	1.76	−4.871	2.94	5.50287

* 本书涉及很多股市术语, 为节约篇幅, 希望读者通过百度了解, 比如本表中出现的换手率, 内盘, 外盘, 振幅, 量比, 委比, 市净率, 市值等术语在股市很常见. 不方便查阅的读者可以把这些术语都看成变量名, 不影响数据分析.

解　先把上面的数据存入表格文件 hn.csv 中, 然后读入到 R 中的变量 D 中, 然后通过程序 outline(D) 获得轮廓图 1.1.

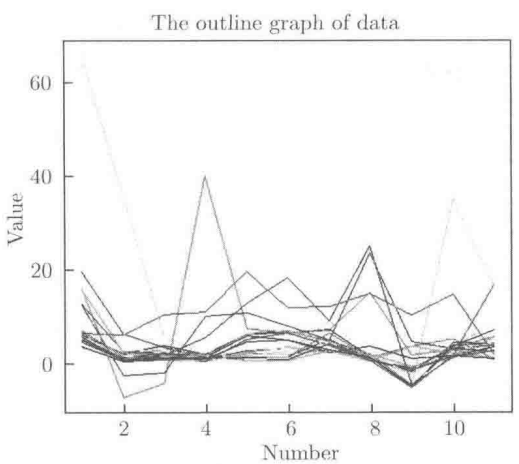


图 1.1　轮廓图

适当的调整变量量纲是必要的, 个别变量的数值过大, 会使轮廓图看不见其它变量的变化. 当然, 为了减少量纲的影响, 还可以对数据施行标准化变换, 即每个变量 $x_j = (x_{1j}, x_{2j}, \cdots, x_{nj})'$, $j = 1, 2, \cdots, p$ 作如下变换

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j},\ i = 1, 2, \cdots, n, \tag{1.6}$$

R 中函数 scale(D) 就可以对数据 D 按列进行标准化处理.

图 1.1 中每条线代表一只股票, 因此表现的是一个 11 维的向量, 从轮廓图可以直观地看出, 哪些股票的相关指标相似, 比如涨跌、市值等, 这种图形在分类中会有帮助.

2. 星图

星图的作图步骤如下:

(1) 作一圆, 并将圆周 p 等分;

(2) 连接圆心和各分点, 把这 p 条半径依次定义为变量的坐标轴, 并标以适当的刻度;

(3) 对给定的一次观测值, 把 p 个变量值分别取在相应的坐标轴上, 然后将它们连结成一个 p 边形;

(4) n 次观测值可画出 n 个 p 边形.

R 软件里有现成的星图命令, 利用 stars(D) 即可画出星图. 对于案例 1.2, 可以画出星图如图 1.2.

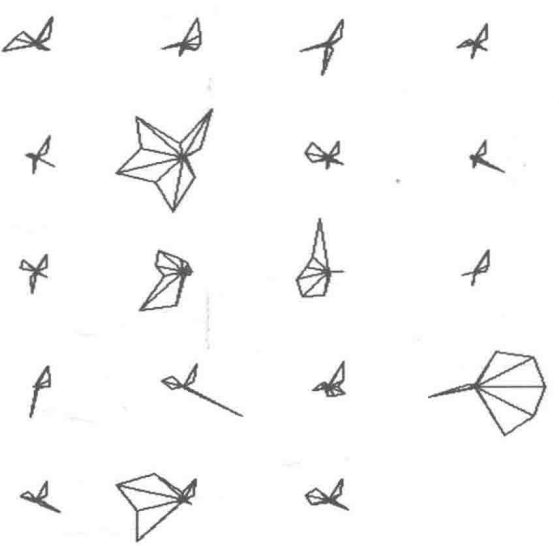


图 1.2　星图

星图中水平轴是变量 X1, 沿逆时针方向, 依次是 X2, X3, ···, 由于星图既像雷达屏幕上看到的图像, 也像一个蜘蛛网, 因此, 星图也称雷达图或蜘蛛图.

图中可以一目了然地观察到各个股票的情况, 是高价股还是低价股, 成交是否活跃等.

函数 stars( ) 可以加各种参数, 从而画出不同的星图, 下面是该程序绘出的半幅星图 (见图 1.3).

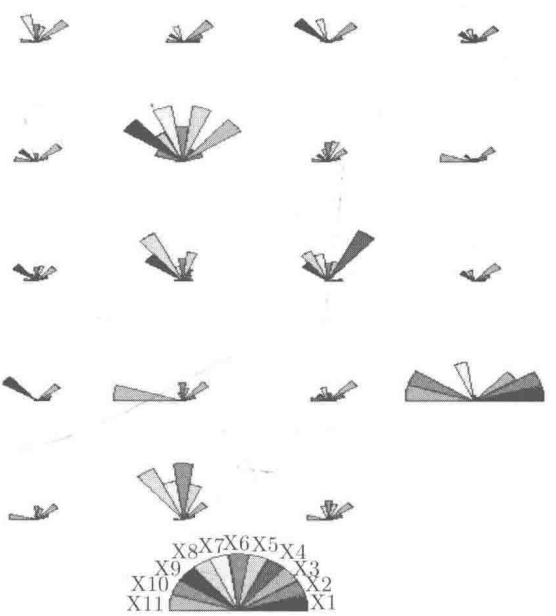


图 1.3　半幅星图

3. 脸谱图

脸谱是非常形象的图形, 戏剧中脸谱的设计独具匠心, 往往能很好的刻画人物的个性. 通过脸的胖瘦、五官特征可以用来描述多元数据, 从而让数据的表现更加形象. 用脸谱表达多变量首先是由美国统计学家切尔诺夫 (Chernoff) 在 1970 年提出, 以后又经过很多学者的改进并收入一些统计软件包, 这里介绍 R 软件中的 faces( ) 函数, 这需要加载程序包 aplpack.

该程序选择了 15 个指标:

1-脸的高度; 2-脸的宽度; 3-脸的形状; 4-嘴的高度; 5-嘴的宽度; 6-笑的形状; 7-眼睛的高度; 8-眼睛的宽度; 9-头发的高度; 10-头发的宽度; 11-头发的类型; 12-鼻子的高度; 13-鼻子的宽度; 14-耳朵的宽度; 15-耳朵的高度.

因此最多可以刻画 15 个变量的数据, 当然可以定义更多的变量进一步细化脸

部特征的描述, 从而适用于更多变量的脸谱描述.

对于案例 1.2 的数据, 可以得到脸谱图 1.4. 脸谱图的不足在于用什么变量来画什么部位没有固定的法则, 因此不同的人画出的脸谱图可能会给人留下不同的印象.

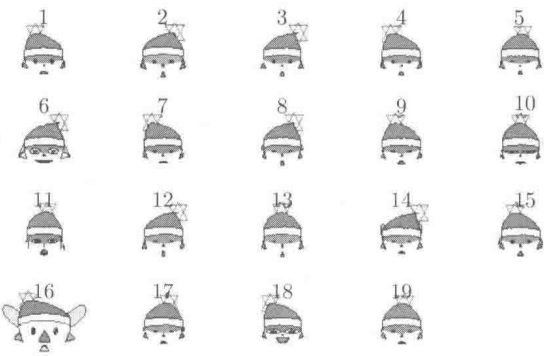


图 1.4　脸谱图

4. 三角多项式图

三角多项式图又称调和曲线图, 是安德鲁斯 (Andrews) 在 1972 年提出的三角表示法, 其思想是将多维空间中的一个点对应于二维平面的一条曲线, 对于第 i 个观测值为 $(x_{i1}, x_{i2}, \cdots, x_{ip})$, 取 $[-\pi, \pi]$ 上的正交函数系 $\{\sin t, \cos t, \sin 2t, \cos 2t, \cdots\}$, 建立如下映射

$$X_i = (x_{i1}, x_{i2}, \cdots, x_{ip}) \rightarrow$$

$$g_i(t) = \frac{x_{i1}}{\sqrt{2}} + x_{i2}\sin(t) + x_{i3}\cos(t) + x_{i4}\sin(2t) + x_{i5}\cos(2t) + \cdots, -\pi \leqslant t \leqslant \pi. \quad (1.7)$$

该映射具有如下优良性质:

(1) 保线性关系;

设 X, Y, Z 均为 p 维向量, a, b 为常数, 若 $Z = aX + bY$, 则

$$g_Z(t) = ag_X(t) + bg_Y(t), -\pi \leqslant t \leqslant \pi. \quad (1.8)$$

(2) 保欧氏距离;

设 X_i, $X_j \in R^m$, 那么在 X_i, X_j 之间的欧氏距离为

$$d_{ij}^2 = (X_i - X_j)'(X_i - X_j), \quad (1.9)$$

由于 $g_i(t)$ 和 $g_j(t)$ 均为 $[-\pi, \pi]$ 上的平方可积函数, 它们之间的欧氏距离可以定义为

$$d_{g_i g_j}^2 = \int_{-\pi}^{\pi} |f_v(t) - f_w(t)|^2 \, \mathrm{d}t, \quad (1.10)$$