管理学精选教材译丛

APPLIED MULTIVARIATE STATISTICAL ANALYSIS

应用多元统计分析

[第2版]

〔德〕沃尔夫冈·哈德勒（Wolfgang Härdle）
〔比〕利奥波德·西马（Léopold Simar） 著

陈诗一 译

北京大学出版社
PEKING UNIVERSITY PRESS

管理学精选教材译丛

APPLIED MULTIVARIATE STATISTICAL ANALYSIS

应用多元统计分析

[第2版]

北京大学出版社
PEKING UNIVERSITY PRESS

北京市版权局著作权合同登记图字:01 - 2008 - 5020

图书在版编目(CIP)数据

应用多元统计分析:第2版:翻译版/(德)哈德勒,(比)西马著;陈诗一译. —北京:北京大学出版社,2011.1

(管理学精选教材译丛)

ISBN 978 - 7 - 301 - 16772 - 4

Ⅰ. ①应… Ⅱ. ①哈… ②西… ③陈… Ⅲ. ①多元分析:统计分析 - 高等学校 - 教材 Ⅳ. ①O212.4

中国版本图书馆CIP数据核字(2010)第238512号

书　　名:应用多元统计分析(第2版)

著作责任者:〔德〕沃尔夫冈·哈德勒(Wolfgang Härdle)

〔比〕利奥波德·西马(Léopold Simar)　著

陈诗一　译

策划编辑:朱启兵

责任编辑:谢　超

标准书号:ISBN 978 - 7 - 301 - 16772 - 4/F·2670

出版发行:北京大学出版社

地　　址:北京市海淀区成府路205号　100871

网　　址:http://www.pup.cn　电子邮箱:em@pup.cn

电　　话:邮购部 62752015　发行部 62750672　编辑部 62752926　出版部 62754962

印　刷　者:北京飞达印刷有限责任公司

经　销　者:新华书店

850毫米×1168毫米　16开本　28.75印张　670千字

2011年1月第1版　2011年1月第1次印刷

印　　数:0001—4000册

定　　价:65.00元

译者的话

我是在韩国攻读计量经济学博士学位期间认识沃尔夫冈·哈德勒(Wolfgang Härdle)教授的,他是我韩国主要导师的长期合作者,两人经常互访对方所在的研究机构。当时我对哈德勒教授的认识是他是国际上非参数计量经济学和统计学领域的大家,在国际顶尖的统计学、计量经济学和金融学学术期刊上发表了超过200篇高质量的学术论文,根据多达十几项指标的综合科学信息标准统计,哈德勒教授是国际上前5%高引用率的科学家之一。

我在2006年初来到复旦大学经济学院工作,经韩国导师的推荐,该年暑期我第一次有机会访问了哈德勒教授所在的德国洪堡大学统计与计量研究所,利用其提供的经济风险数据进行合作研究。我来到了哈德勒教授无比宽敞且古色古香的办公室,在这里,我第一次看到了书橱里整齐陈列的哈德勒教授所著的多达三十几本专著和教材。例如,1990年由剑桥大学出版社出版的*Applied Nonparametric Regression*,这是世界计量经济学协会出版的系列专著之一,是第一本讲述非参数回归分析的教材,也是欧美攻读经济学博士学位学生的必修教程,广泛影响了一代研究者,也奠定了哈德勒教授在该领域的学术地位。又如,*Statistics of Financial Markets: An Introduction*等书在欧美销量特别大,这是因为哈德勒教授的教材非常标准化和专业化,其又一大特色是应用性,而且电子讲义、软件程序、案例数据、习题答案配备齐全,极大地方便了读者。我于是萌生了把哈德勒教授的著作翻译成中文出版的念头。

哈德勒教授对我的提议非常感兴趣,他的著作大部分以英文或德文出版,也有以俄文和日文发表的著作。用中文出版的著作可是没有的啊!而且,我看得出,哈德勒教授对神奇的中文和神秘的中国文化非常喜爱。他推荐我翻译的第一本教材是*Applied Multivariate Statistical Analysis*,就是读者现在看到的这本《应用多元统计分析》。哈德勒教授的言下之意似乎是,如果这本书翻译效果不错的话,他就会让我继续翻译他的其他著作。一个可爱的绅士!这让我有点儿诚惶诚恐了,下面的活儿是否接到还真不好说呢!

2007年8月我在里斯本参加国际统计学协会(ISI)第56届年会的时候遇到了该书的另一位作者,即曾任比利时统计学会主席的利奥波德·西马(Léopold Simar)教授,他也对我翻译这本书给予了热情的鼓励。任务很早就接下来了,但是遗憾的是,翻译的时间拖得太久了点儿,本来想作为西马教授2009年5月退休典礼的礼物,也没有来得及。我想理由无非是作为国内高校的教师太忙了,教学科研任务一大堆,很难抽出整块的时间来进行这种学术翻译。最初,我找了几位学生进行预翻译,但还是感觉离出版要求有不少距离,于是,我不得不

挤出宝贵的时间埋首这早就答应了的任务，逐字逐句进行翻译，算算从2008年6月开始到现在翻译任务初步完成，也花了大概一年多的时间吧，希望能让读者满意。值得一提的是，几年来我和哈德勒教授合作的学术论文中已经有两篇今年以来分别被国际一流的SSCI学术期刊 *Quantitative Finance* 和 *Journal of Forecasting* 正式接受了，真的盼望这本译著今年也能够借此东风顺利出版！

这本教材不仅适合统计学和经济学具有数理计量背景的高年级本科生、研究生和高校教师使用，也适合那些对数理要求不高的一般社会科学研究者使用和参考。最后，译者借此机会感谢美国普林斯顿大学讲座教授和中国科学院特聘教授范剑青、美国康奈尔大学教授和厦门大学王亚南经济研究院院长洪永淼以及以色列统计学会前主席和耶路撒冷希伯来大学教授 Ya'acov Ritov 在百忙中为该书所作的精辟点评！感谢北京大学出版社朱启兵编辑、谢超编辑的辛苦工作和哈德勒教授指导的博士生宋颂（Richard Song）所提供的方便！感谢冯倩、高远和范子英同学对该书的初译工作！感谢上海市重点学科建设项目（编号B101）和复旦大学985国家哲学社会科学创新基地“中国经济国际竞争力研究”课题的资助！最后，我还要感谢我的妻子陈梅女士和儿子陈骁禹对我一直以来的大力支持！

当然，由于时间紧迫和精力有限，书中译误在所难免，欢迎读者积极提出，译者将进一步修改更正（译者的E-mail：shiyichen@fudan.edu.cn）！

陈诗一

2010年8月于复旦大学

第2版序言

本书第2版拓展了应用多元统计分析的方法和应用范围。为了使本书更适时更具有应用特色,第2版例子引入了更多最新的数据集。既然多元统计方法被大量应用于数量金融和风险管理领域,新版本使用更大的篇幅来讨论相关变量的分布及其密度问题。

这一版更详细地讨论了不同的厚尾分布家族,比如 Laplace 和广义双曲线分布。同时新增一节来专门讨论金融风险管理和信用等级排列领域有关依存关系的最新概念——copulae。在计算功能强大的方法和理论章节增加了支持向量机方法,这是统计学习理论中最新的分类和回归技术,所给出的相应例子是把该方法应用于公司破产和信用等级排序分析。该章节中同时增加了分类和回归树技术并应用于公司评级案例。

第2版为增加本书可读性和友好性的最重要的改进是把原来使用的 Quantlet 语言同步翻译成了 R 和 Matlab 语言,其算法可以从 www. quantlet. com 上下载。我们感谢 Anton Andriyashin、Ying Chen、Song Song 和 Uwe Ziegenhagen 为本书第2版的出版所做的工作!

沃尔夫冈·卡尔·哈德勒(Wolfgang Karl Härdle)

利奥波德·西马(Léopold Simar)

2007年6月分别于柏林和新鲁汶

第1版序言

在实证科学中，大部分观测到的现象具有多元的性质。例如，金融研究中，同时观察股票市场的各种资产以及分析它们的联动发展有助于更好地理解其一般变化趋势并跟踪指标。在不同场所对不同医疗科目的医学记录意见是进行可靠诊断和药物治疗的基础。在营销定量研究中，消费者的各种偏好被收集来构造消费者行为模型。可见，这些例子以及许多其他应用科学的定量研究的潜在理论结构是多元的。《应用多元统计分析》这本书将为多变量数据分析特别是其应用提供各种必要的概念和分析手段。

这本书的目的在于给那些每天不得不面对大量统计数据的非数学专业的研究者们提供一种更容易理解的多元数据分析方法。一方面，通过提供较多的实证案例来实现这个目标；另一方面，购买本书同时所提供的电子图书可以帮助读者重新运算甚至修改书中的所有例子，而这通过标准网络浏览器就可以实现，并不需要依靠任何具体的统计软件。

这本书主要分为三个部分。第一部分讨论图形分析技术，并描述所涉变量的分布。第二部分主要处理多元随机变量，从理论的角度来讨论其分布，分析其在各种实际情况下的估计和假设检验。最后一部分讨论多变量处理技术，并且引导读者来选择合适的工具用于多变量数据分析。所有数据集均在附录中列出，并可以从 www. md-stat. com 网站下载。本书包含了大量的练习题，其答案在另外的习题集一书中给出。另外，与本书教学相关的讲义以一套完整幻灯片的形式提供给教师，以方便授课，这也可以从 www. md-stat. com 网站上下载。所有幻灯片包含的超链接均可链接到统计网页服务器，以便教师和学生通过标准网络浏览器来再现所有的例子。

具体而言，本书第一部分，即图形描述性分析技术部分，首先讨论箱形图（boxplot）的构造。这里使用的案例主要是关于银行真钞和假钞以及波士顿住房的标准数据集。1.5 节介绍夫洛瑞（Flury）脸谱图，1.6 节介绍安德鲁（Andrew）曲线和平行坐标图形。最后还介绍了直方图、核密度图和散点图等。从这些图形的绘制中，读者可以了解到偏度和相关的概念。

本书第二部分一开始首先帮助读者简单回忆一下矩阵代数的知识以及诸如协方差、相关、线性回归模型等基本概念。然后介绍方差分析（ANOVA）技术及其在多元线性回归模型中的应用。第 4 章是关于多元分布特别是多元正态分布的知识。在多元随机变量内容的最后将介绍估计和假设检验的理论。

本书第三部分及其后内容从数据矩阵的几何分解开始，这受到了数据分析法国学派的

影响。这种几何的观点与第10章的主成分分析相关。因子分析章提供了大量的心理学和经济学方面的例子。聚类分析部分讨论了许多聚类技术,并且自然延伸到判别分析章。接着讨论了因素间的对应分析问题。典型相关分析章描述了数据集的联合结构,并给出了一个有关小汽车安全特性和价格间实证研究的案例。接下来讨论的是重要的多维标尺问题和联合测度分析,后者常用于心理学和市场营销领域,比如度量特定商品的偏好排序。在金融领域的应用也很多,见第18章,比如CAPM模型和有效资产组合构成等。本书讨论的这些技术具有高度互动和计算能力强大的特征。

这本书适合于高年级的本科生、低年级的研究生以及那些期望使用各种多元数据分析工具的经验还不丰富的数据分析师。而具有足够代数知识的有经验的读者可以跳过那些关于多元随机变量知识的基础章节而专注于各种多元分析技术的数学根源内容。对于一名研究生来说,第一部分的数据描述性分析技术应该在基础性的统计学课程中就学习过了,第二和第三部分的数理根源和应用技术才是进一步的关于多元统计分析技术的必须掌握内容。

本书及其电子版本将通过不同的应用案例帮助那些没有个人电脑使用经验的读者逐步熟悉跨学科间的统计思维方法。特别是本书的电子图书是一个有效的互动学习的资料,它的完整版本可以根据本书最后一页所附的出版社提供的许可密码从 www. xplore-stat. de 网站下载。同时提供的还有本书完整的PDF和HTML版本。本书读者还可以通过XploRe Quantlet Server(XQS)服务器根据所提供数据来再现书中所使用的各种方法,而不需要下载或者购买另外的统计软件。该服务器可以按照 www. ixplore. de 网页说明免费安装。

这本书的出版得自许多朋友、同仁和同学的帮助。我们感谢 Jörg Feuerhake, Zdeněk Hlávka, Torsten Kleinow, Sigbert Klinke, Heiko Lehmann, Marlene Müller 对本书电子版本的技术支持。Christian Hafner, Mia Huber, Stefan Sperlich, Axel Werwatz 仔细校阅了这本书。我们也感谢 Pavel Čížek, Isabelle De Macq, Holger Gerhardt, Alena Myši čková and Manh Cuong Vu 对本书的练习和统计问题所给出的答案。最后我们要感谢 Springer Verlag 出版社的 Clemens Heine 对本书的写作风格和应该包含的内容所提供的持续支持和有价值的建议!

沃尔夫冈·卡尔·哈德勒(Wolfgang Karl Härdle)

利奥波德·西马(Léopold Simar)

2003年8月分别于德国柏林和比利时新鲁汶

目录

第一部分　统计描述技术

第二部分　多元随机变量

第三部分 多元技术

第四部分 附 录

第一部分　统计描述技术

第1章 批量数据比较

多元统计分析关注的是如何分析和理解高维数据。假定$\mathbb{R}^p$空间的数据矩阵X的某一个变量列向量为$\{x_i\}_{i=1}^n$,其观察值的个数为n,则p维矩阵X的第i个观察值x_i可以表示为:

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

这里,$X \in \mathbb{R}^p$,即,X由p个随机变量组成:

$$X = (X_1, X_2, \cdots, X_p)$$

其中$X_j, j = 1, \cdots, p$,是一个一元随机变量。如何分析这种类型的数据呢?在我们研究从这些数据可以获得什么样的统计推断之前,我们应该先考虑怎样"看待"这些数据。这就会使用到数据描述性分析技术。我们可以通过数据描述技术回答下述问题:

- X中是否存在具有不同变化特征的成分?
- X中的所有变量是否可以分成不同的组别?
- X中的变量是否存在异常值?
- 数据的分布有多么接近于"正态"?
- 是否存在X的"低维"线性组合可以用来表示"非正态"的行为?

高维数据描述性分析的困难之一就是人类认知体系的缺陷。二维点集图往往容易理解和解释。我们有机会利用现代交互计算技术看到实时3D旋转并因此来帮助理解三维数据。哈德勒和斯科特(Härdle and Scott, 1992)所提出的"滑行技术"(sliding technique)可以通过描绘随第四个变量在一定范围内变化而变化的动态3D密度等高线来探究四维结构数据。如果不能将高维数据映射到低维组合,则描述五维或五维以上的数据就成了一条不可逾越的鸿沟(Klinke and Polzehl, 1995)。不过像亚群聚类和异常值这样的数据特征还是可以运用纯粹的图形技术来识别的。

在本章中,我们将研究基本的数据描述及绘图技术以便进行数据的探索性分析。我们首先用箱形图来解释数据。箱形图是一种简单的单变量描述工具,它通过把数据排序分组来发现异常值和比较不同部分的分布特征。接下来介绍多变量的图形分析方法(如彻诺夫—夫洛瑞脸谱图、安德鲁曲线和平行坐标图),这是本书的主要任务。每种方法的优势和劣势将同时被陈述。

我们将介绍两种估计密度的基本技术:直方图和核密度。通过密度估计可以对数据分布的形状快速一瞥。我们将会看到核密度估计将会克服直方图方法的一些缺点。

最后,散点图在绘制两变量或三变量关系时非常有用:它可以帮助我们理解某个数据集中变量间的本质关系,也可以帮助识别数据点的类别和集聚特征。窗格图或矩阵图将多个两变量的散点图表示在同一个视图中,有助于识别变量之间的条件依赖结构或关系。异常值或者需要特殊关注的观察值可以通过安德鲁曲线和平行坐标图来发现。本章最后通过对波士顿住房数据进行说明性分析而结束。

1.1 箱形图(Boxplots)

例 1.1 瑞士银行数据(见附录表 B.2)包括了瑞士银行钞票的 200 个测量值。前一半观察值来自真钞票,后一半观察值来自假钞票。

如图 1.1 所示,权威的度量指标如下:

X_1 = 钞票长度

X_2 = 钞票宽度(左)

X_3 = 钞票宽度(右)

X_4 = 内框到下边缘距离

X_5 = 内框到上边缘距离

X_6 = 中央画面的对角线长

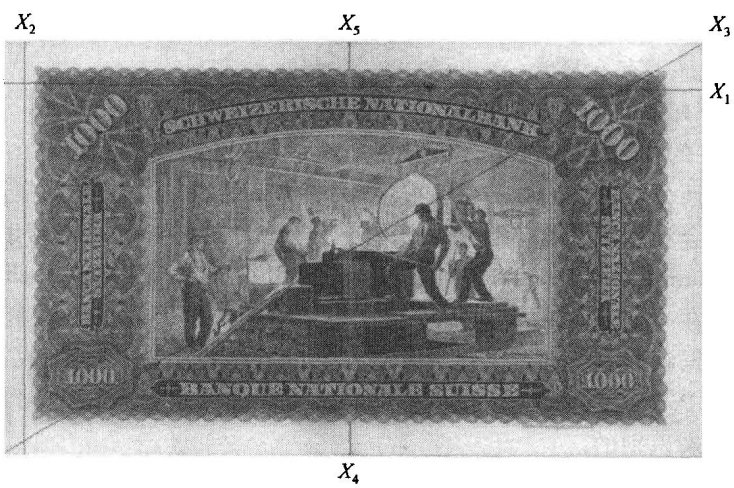

图 1.1 一张旧的瑞士银行 1 000 法郎钞票

这些数据来自夫洛瑞和瑞德维尔(Flury and Riedwyl, 1988)的研究,此项研究的目的是如何运用以上测量值来辨别钞票的真伪。箱形图运用图形技术反映出变量的分布,有助于我们识别其位置、偏度、跨度、尾长和异常点。