成分数据多元分析方法研究

孟洁 著

中央财经大学统计学院学术文库

中国统计出版社
China Statistics Press

多元数据成分研究方法分析

孟洁 著

中央财经大学统计学院学术文库

中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

成分数据多元分析方法研究/孟洁著.
—北京:中国统计出版社,2008.4
(中央财经大学统计学院学术文库)
ISBN 978-7-5037-5371-8

Ⅰ.成…
Ⅱ.孟…
Ⅲ.统计数据—统计分析(数学):多元分析—分析方法
Ⅳ.0212

中国版本图书馆 CIP 数据核字(2008)第 026645 号

成分数据多元分析方法研究

作　　者/孟　洁
责任编辑/吕　军　谢蕾蕾
装帧设计/艺编广告
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 57 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
网　　址/www.stats.gov.cn/tjshujia
电　　话/邮购(010)63376907　书店(010)68783172
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/880×1230mm　1/32
字　　数/120 千字
印　　张/4.75
版　　别/2008 年 10 月第 1 版
版　　次/2008 年 10 月第 1 次印刷
书　　号/ISBN 978-7-5037-5371-8/O·63
定　　价/14.00 元

序

长期以来，统计学在自然科学和社会科学中都有着极其重要而广泛的应用。统计方法作为数量分析的一种有效方法已经成为理、工、农、医、人文、社会、经济、管理、军事、法律等诸多学科领域的基本方法。特别是近几年来，随着社会经济、科学技术的飞速发展，统计方法与具体的实际问题紧密结合，并辅之以高效的计算机技术，更从广度和深度上大大拓展了统计学在理论、方法和应用等方面的内容体系。可以说，在当今信息爆炸的时代，离开数据、离开数据分析、离开统计应用，就意味着远离科学、远离进步、远离财富。

为繁荣统计与数据分析领域的科学研究事业，鼓励广大教师积极开展学术研究，并使研究成果能够尽快地服务于经济和社会的发展，在中央财经大学学术著作出版资助项目的支持下，中央财经大学统计学院组织出版了“中央财经大学统计学院学术文库”（以下简称文库）。希望它能够为从事统计理论、方法和应用的研究人员以及实际工作者们提供及时而有效的帮助，也能够对广大学生读者深入理解统计、数据分析的内涵有所帮助和启迪。

文库的选题主要源于统计理论、方法和应用领域的相关博士论文，一旦有好的选题，即可列入出版计划。此次出版的书目包括下列四部：

《经济周期波动：测度方法与中国经验分析》，吕光明　著

《中国各地区金融发展与固定资产投资实证研究》，赵楠　著

《成分数据多元分析方法研究》，孟洁　著

《含组结构和层次结构模型的规则化路径估计》，马景义　著

文库的选题注重引进当前国内外在统计理论、方法和应用方面的前沿研究内容，注重与计算机技术相结合，着力突出以下特点：

1. 创新性：文库所收录的专著，在经典统计理论的基础上，进行方法或应用方面的创新，并付之于实践应用，用以解决实际问题。

2. 广泛性：文库包括统计方法在经济、金融领域的实证研究、特殊数据类型的统计建模研究、数据挖掘算法研究等多方面的内容，对自然科学和社会科学领域的广大读者具有一定的参考和借鉴价值。

3. 可操作性：文库中涉及到的统计方法、应用案例等，读者都可以按照其中具体介绍的实施步骤进行演练或用于解决自己工作生活中的数据处理问题；同时，充分考虑读者需求，文库中部分涉及并介绍了实现相关模型方法的应用软件或可编程软件。

4. 通俗性：行文按照“统计原理→模型算法→案例应用”的组织形式，努力体现深入浅出的结构安排和文字风格，便于读者的理解，不同读者群可以有所取舍地阅读和学习。

总之，统计学作为数据处理的方法论，具有广泛的应用领域。本文库的初衷正是希望为相关读者奉献一系列具有一定理论高度，且具备一定指导性和实战性的统计方法和应用书籍，可以让众多领域的科研人员、管理人员、高校学生从中获益。本文库的出版感谢中国统计出版社的大力协助和支持，希望我们的共同努力能够筑就统计学未来的辉煌。同时，欢迎广大读者提出批评意见，以利于我们不断提高。

邱　东

2008 年 4 月

前　言

成分数据(Compositional Data)是指由若干在 范围取值、加和为 1 的分量构成的向量,在社会、经济、技术等许多领域,都有着非常重要而广泛的应用。与普通数据相比,采用成分数据建模分析的优势在于:①成分数据能够揭示绝对数据背后的相对信息;②原始无关联数据通过转化为成分,建立起相互约束关系,可以作为一个结构性整体建立模型;③可以对多变量组的整体性特征进行综合分析;④可以对多变量组的结构性特征、差异性特征、相互关联关系进行内涵分析。

成分数据的概念最早来自于 1866 年 Ferrers 的工作。1897 年,Pearson 在一篇讨论伪相关的权威性文章中指出:"在实际的成分数据分析中,定和限制常常被有意或无意地忽略,一些为不带限制条件的数据而设计的传统统计方法经常被不适当地滥用,从而造成灾难性的后果。"直到 1986 年,Aitchison 发表了首部系统研究成分数据的论著《The Statistical Analysis of Compositional Data》,提出采用成分分量比值的对数,即"对数比(logratio)"作为研究成分数据的基本单元,并由此提出对数比协方差结构、对数衬度(logcontrast)线性组合、加法逻辑正态分布等一系列适于成分数据统计分析的基本理论体系,有效解决了成分数据定和约束(sum to unit)条件的限制以及由此带来的闭合效应(closed effect)的影响。并且,Aitchison 在成分数据单形空间中的加法逻辑正态分布假设的基础上,研究了关于成分数据的参数估计、假设检验问题,以及成分数据的线性回归、判别分析、降维技术等多元统计分析模型。

国内有关成分数据的研究,1990 年中国地质大学周蒂教

授等出版了该书的中译本《成分数据的统计分析》。2000 年，张尧庭教授出版专著《成分数据统计分析引论》，由成分数据的统计分布问题出发，着重研究了单形空间的狄氏分布；并以此为基础，探讨了成分数据的估计、回归、判别分析、典型相关分析、贝叶斯方法等统计分析模型。章栋恩(2002)也对成分数据在单形空间的狄氏分布方面进行了较深入的理论研究。

之后，国内有关成分数据的大量研究出自于我的导师王惠文教授等。在成分数据预测方面，2002 年，王惠文等提出球坐标变换方法，解决了对含有零分量的成分数据进行时序预测的问题。在成分数据的回归分析方面，2003、2006 年，王惠文等提出将成分数据对称的对数比变换与偏最小二乘回归分析、偏最小二乘通径分析方法相结合，解决一元成分数据和多元成分数据的线性回归建模问题。本书的研究内容正是在此基础上展开并深入的，主要建模思路沿袭了将成分数据对数比变换与偏最小二乘相关分析技术相结合的做法。同时，本书的理论研究部分主要借鉴了 Aitchison(1986)所提出的成分数据对数衬度主分量分析模型(Logcontrast PCA)，Hinkle 和 Rayens (1995)所提出的对数衬度偏最小二乘回归分析模型(Logcontrast PLS)，在研究过程中特别注意遵循成分数据的基本代数理论性质要求，对已有分析方法进行理论上的深化和完善。

本书出自我的博士论文，在此向我的导师王惠文教授致以最诚挚的谢意！书中关于成分数据的多元分析建模方法根源于她的研究思路，书中的研究更多的凝结着导师的无尽心血。此外，还有曾在北航复杂数据分析研究中心的学长们，他们在成分数据多元分析领域的研究中也作了大量的工作，这些都是本书得以完成并出版所不可或缺的，在这里我表示衷心的感谢！

本书的相关研究工作得到了我的导师所主持的国家自然科学基金项目“成分数据多元分析中的理论方法研究(编号：70371007)”的支持；本书的出版得到了我现在所在单位中央财经大学著作出版基金的资助，以及我所在的统计学院的领导、

老师们的全力支持，在此表示深深的谢意。我还要感谢中国统计出版社的同志，他们为本书的出版和编辑付出了大量的精力。

最后，希望得到读者的批评和帮助，以便改正由于我的水平有限所出现的不足。

孟　洁

2008 年 4 月

目　录

第 1 章　绪 论　1

1.1　研究的背景和意义　1
1.2　成分数据分析方法概述　3
1.3　偏最小二乘分析技术概述　7
1.4　本书的研究内容和结构安排　11

第 2 章　成分数据分析基础　14

2.1　成分数据的概念　14
2.2　成分数据分析的特殊困难　15
2.3　成分数据分析基础　18
2.4　小结　28

第 3 章　成分数据预测建模　31

3.1　基于变换的成分数据预测方法　31
3.2　非线性偏最小二乘回归方法　32
3.3　成分数据预测建模步骤　58
3.4　成分数据预测模型的评价　59
3.5　小结　60

第 4 章　成分数据回归分析　62

4.1　一元普通数据关于一元成分数据的线性回归方法　62
4.2　一元成分数据关于一元成分数据的线性回归方法　67

4.3 一元成分数据关于多元成分数据的线性回归方法 72
4.4 小结 79

第5章 成分数据偏最小二乘通径分析 81

5.1 偏最小二乘通径分析 81
5.2 成分数据的偏最小二乘通径分析 89
5.3 小结 91

第6章 应用研究——北京市三次产业就业结构需求分析 93

6.1 北京市就业需求研究概述 93
6.2 三次产业就业结构的需求预测 100
6.3 三次产业就业结构关于GDP结构的回归分析 105
6.4 三次产业就业结构关于投资结构和GDP结构的回归分析 113
6.5 三次产业的投资结构、GDP结构、就业结构的路径关联分析 124
6.6 相关建议措施 129
6.7 小结 132

主要参考文献 134

第1章 绪论

1.1 研究的背景和意义

在社会、经济、技术等许多领域的数据分析中，成分数据(Compositional Data)都有着非常重要而广泛的应用，可以被用来反映诸如投资结构、产业结构、居民消费结构等问题。国民经济各部门的财政开支比例、竞争企业的市场份额、大学毕业生分配方案、血液成分、药剂配方、以及地质学中的岩石化学全分析结果、岩石的矿物组成、沉积物粒度数据、化石种属比例等等，这些都是成分数据。因此，成分数据的研究对人类认识社会、认识自然至关重要。

成分数据最简单的应用就是在描述性统计分析中，应用饼图来表示在某一事物中各成分所占的比重。例如，为了研究中国的三次产业结构，可以分别统计第一产业、第二产业和第三产业的国民生产总值、并绘制成饼图，在饼图中可以分别表现每一年中三次产业所占国民生产总值的比重，而这些份额的总和应该等于1。更进一步的，如果我们还收集了三次产业的就业人口比重数据，就可以分析三次产业的生产总值结构对就业结构的影响关系，等等。

与普通数据相比，一方面，从原始绝对数据计算得到的比例结构，即成分数据，能够更进一步揭示绝对数据背后的相对信息；另一方面，成分数据更适于分析整体的各部分比例关系，能够对比例结构数据进行整体性综合分析，深入挖掘其中的相互关联和制约关系。

另外，在很多数据分析方法的理论研究和模型建立过程中，成分数据也起着举足轻重的作用。例如，在主成分分析中，所有主成分的贡献率构成一个成分数据；在Fisher判别分析中，组间离差阵和组内离差阵的全体特征值的和是一个常数(由数据表的样本和变量数所

决定),因此也可以转化为成分数据进行建模预测。

然而,与普通数据不同,成分数据要求各成分分量非负且总和为1,称为定和约束条件,这一约束条件给有关对成分数据的统计分析带来诸多困难。早在1897年,Karl Pearson在一篇讨论伪相关的权威性文章中第一次指出:试图解释在分母及分子中含有公共部分的那些比例之间的相关性是很危险的。这意味着,研究作为整体的各部分的相互关系的成分数据分析将充满困难。历史证明他是对的;在以后的年代里,可以说直到今天,还没有哪个数据分析领域像成分数据的统计分析那样充满了概念的混淆及方法的滥用。在实际的成分数据分析中,定和限制常常被有意或无意地忽略,为不带限制条件的数据而设计的"传统"统计方法经常被不适当地滥用,造成了灾难性的后果。

成分数据这类分布在有限区域内的带"定和"约束条件的数据,又进一步带来了闭合效应及参数分布上的许多困难。由于成分数据的闭合效应,使得成分数据的协方差矩阵具有明显的负偏性,截然不同于对开放数据的协方差矩阵的经典解释。实验结果显示,对原始数据进行成分化或闭合化后,原始数据与成分数据的统计量(相关系数、方差)不一致,存在很大的差异。在原始数据中弱相关的变量,在成分数据中却呈现强的负相关。而且,在成分数据分析中,很难找到一个适合的参数分布形式。但是常规统计方法多基于正态分布假设,而正态分布的随机变量的取值范围为$(-\infty,+\infty)$,即实数空间,在实数空间上可行的和正态分布相应的一系列统计方法到了局限得多的单形空间就不适用了。结果,成分数据的特殊约束使得用于普通数据的数理统计方法如主成分分析、因子分析等在用于成分数据分析过程中就得出错误的结论。因此,可以看出,不认识这些困难,盲目地套用为无约束条件数据而设计的经典统计法方法来研究成分数据是完全错误的。

可见,成分数据作为一类非常特殊的数据类型,其在约束条件、协方差结构、分布的复杂性等多方面的研究都存在相当的难度。在很多实际领域的应用中,涉及到成分数据的众多不规范做法不断遭到质疑和否定。因此,进一步探讨有关成分数据分析领域的科学方

法确是十分必要并且迫切的。

针对上述问题，上世纪70年代，英国统计学家Aitchison就开始研究成分数据统计分析中的问题。他认为如同带约束条件的方向数据有一套独特的统计分析方法一样，成分数据也应该有自己的一套统计方法。根据成分分量的比值不受闭合效应的影响、比值的对数常常服从正态分布的特点，他提出用成分分量比值的对数，即“对数比(logratio)”，来作为研究成分数据的基本单元，从而发展了一整套适用于成分数据的统计分析方法，解决了应用统计学中一类特殊领域中的众多难题。由于他在1982年发表的论文《The Statistical Analysis of Compositional Data》有很高的被引用率，他被授予英国皇家统计学会1988年研究奖章；这显示了统计学界对他的贡献的承认。

本书正是在这样的应用背景和研究基础上选题的。针对成分数据的特殊性质，在已建立的成分数据分析的基本理论的基础上，拓展并创新成分数据的多元统计分析及建模技术，进一步完善成分数据的理论方法体系，从深度和广度上增强模型的应用价值和解决实际问题的能力。

1.2 成分数据分析方法概述

有关成分数据的研究工作，可以分为“理论”和“方法”两方面内容。理论是方法研究的基础，对成分数据多元分析方法的研究同样是建立在成分数据分析的理论基础之上的。下面就对这两方面作一简要概述。

1.2.1 成分数据分析的理论基础

成分数据的概念最早来自于1866年Ferrers的工作。1897年，Pearson在一篇讨论伪相关的权威性文章中指出：“在实际的成分数据分析中，定和限制常常被有意或无意地忽略，一些为不带限制条件的数据而设计的统计方法经常被不适当地滥用，从而造成灾难性的后果。”之后，Quenouille(1953，1959)发表了关于成分数据的统计模型研究。1986年，Aitchison发表了首部系统研究成分数据的论著《成分数据统计分析》，该书对成分数据的单形空间、协方差结构、降

维技术以及成分数据的逻辑正态分布等理论方法进行了深入的研究。

用数学方法来描述，一个成分数据是指任意非负的 D 维向量 $X=(x_1,\cdots,x_D)\in R^D$，向量 X 中的 D 个分量的取值满足以下约束条件

$$\sum_{j=1}^{D} x_j = 1,\quad 0 \leqslant x_j \leqslant 1$$

上式又被称为定和限制，它是成分数据的基本特征。

成分数据分量的取值范围是[0,1]，D 元成分数据所张的空间在数学上称为“单形(simplex)”空间，其维数是 $D-1$。例如，三元成分数据($D=3$)所张的空间是边长为 1 的等边三角形，即岩石学中常用的三元图，其维数为 2。但是，常规统计方法大多是基于正态分布假设，其随机变量取值范围是$(-\infty,+\infty)$，即实数空间。这些统计方法对于单形空间中的成分数据是不适用的，这就像常规统计方法不适用于方向数据一样。对于单形空间中随机变量的分布形式，过去所知的仅是 Dirichlet 分布族，其密度函数呈复杂的形式。它代表的是一种完全独立(包含任何形式的独立性)的数据结构，其相关系数全部是负数，可看作是由完全独立、具有相同标度的伽玛分布的基形成的成分，而不能用于研究成分数据中大量存在的相关性质。

针对成分数据分析中所存在的困难，Aitchison 等人在成分数据统计方法的研究中发现了一条可以联系原始数据和成分数据的线索，这就是 Aitchison 定义的对数比变换(logratio 变换)。对数比变换使成分数据从原来的 D 维空间被降低到一个 $D-1$ 维空间，消除完全线性相关性，但非对称变换使得变量的物理含义发生变化。为此，Aitchison 提出中心化的对数比变换(对称的 logratio 变换)。中心化对数比变换使变换后的各分量仍保持对称性，所建模型的可解释性就更强，但中心化对数比变换并不能消除变量的完全相关性。

Aitchison 提出对数比变换的主要论据是：①若要从成分数据中得到基的信息，必须抓住完全不受闭合效应影响的特征，组分的比值就是这种特征，比值的对数亦然；②对数比的取值范围是整个实数空

间，而不是单形空间。如果成分数据是由很多次相互独立的非负摄动形成，则根据中心极限定理，其对数比将近似地服从正态分布，而此时成分数据则服从单形空间中的加法逻辑正态分布（additive logistic normal distribution）。就这样，对数比变换同时解决了成分数据统计分析中的闭合效应和统计分布的问题。

由于变换后数据消除了约束条件的限制，可以自由取值，并且易于再实施反变换而得到原始的成分数据，因此，这就使得对成分数据进行建模分析和预测的研究成为可能。同时，成分数据分布问题的解决，也使得能够对成分数据进行参数统计建模以及假设检验，验证模型的合理有效性，从理论上进一步完善成分数据分析的方法体系。

在对数比变换的基础上，Aitchison 提出建立基于对数比变换的成分数据的协方差结构，有效解决了采用基于原始协方差结构研究成分变异性的实质所带来的混乱。同时，如同在普通数据的多元分析方法中，如主成分分析、线形回归分析、判别分析等，模型的建立与求解都是以数据表的协方差矩阵为基础进行的，在成分数据的多元分析中，依然如此。成分数据的协方差结构，特别是中心化对数比协方差矩阵，是建立成分数据多元分析模型的基础。已有的成分数据对数衬度主分量分析（Aitchison，1986），对数衬度偏最小二乘回归分析（Hinkle，Rayens，1994），模型的建立与求解都是以中心化对数比协方差矩阵为直接工具而实现的。在本书中，我们在有关成分数据的回归分析、成分数据的偏最小二乘通径分析的研究中，也遵循这一基础理论，采用了基于协方差结构的建模方法，所得结果更符合成分数据单形空间的理论性质要求。

线性组合是代数理论体系中的又一基本要素，在众多多元统计方法的研究中，它也是一种常见而重要的表现形式。在成分数据单形空间的理论研究中，Aitchison 给出成分数据的线性组合形式——对数衬度。事实上，成分数据对数衬度也是由成分数据对数比变换的线性组合推导而来，这使得众多基于对数比变换的建模方法，对它们的简单求解，所得结果可以等价的表现为对数衬度的形式。

完备的理论是进行方法研究和创新的基石，可以说，Aitchison 的对数比变换是成分数据分析的根本出发点，由此引出的加法逻辑

正态分布、对数比协方差结构、对数衬度等基本概念，丰富了成分数据的理论内涵，并为进一步解决成分数据分析中的特殊困难、拓展成分数据的多元分析模型，奠定了坚实的科学依据。

目前，对成分数据的研究已经深入到对单形空间的代数—几何体系的研究。在单形空间中定义了内积，并且证明了这个内积空间为 Hilbert 空间(完备)。这样，就可以直接在单形空间中研究有关成分数据的各种分析方法。

1.2.2 成分数据分析方法综述

Aitchison 首次系统地研究了成分数据统计分析方法，详细论证了对数比统计分析的理论推导和计算方法，并用实例阐述了它所能解决的许多有关成分数据的实际问题。例如：①子成分分析问题、降维问题；②两组成分数据之间差异是否显著，即对成分均值相等和协方差相等的假设检验问题；③成分的相关性问题；④成分与外部变量关系的线性模拟问题；⑤成分类别的判别问题；等等。

实际上，将成分数据进行对数比变换以后，就可以使用各种传统(基于正态分布假设)的统计方法。对数比方法的优越性在于：①变换后的数据将满足方法的统计分布假设，因而从统计学的角度看是正确的；②得到的结果将真正代表基的特征，不受闭合效应影响。

Pawlowsky 等(1995)将对数比统计方法与地质统计学方法相结合，提出了区域化成分数据的统计方法；周蒂(1991，1995)将对数比主分量分析和聚类分析方法相结合，改善了海洋沉积物粒度数据的分类。在 Aitchison 的对数比变换中，要求成分数据中不能含有零成分。Aitchison 指出，对零成分可以进行合并成分、用“微迹”(很小的正数)代替零值、分开研究等处理。周蒂(1997)通过实验指出，用“微迹”代替零值后进行对数比主分量分析，样品的群聚表现却显著不同。为了解除对数比变换要求成分数据分量不能为零的限制，王惠文等(2002)提出采用球坐标变换的方法对含有零分量的成分数据进行处理，并进行成分数据的时序预测建模。

1994 年，John Hinkle 和 William Rayens 提出了成分数据的对数衬度偏最小二乘(Logcontrast Partial Least Squares，记为 LCPLS)方法，首次将偏最小二乘回归方法用于成分数据的线性回归

分析中，对中心化对数比变换后的成分数据采用基于协方差结构的单变量偏最小二乘回归算法，实现了模型的简单求解，解决了因变量为一元普通数据，自变量为一元成分数据的建模问题。

国内学者王惠文教授在有关成分数据的多元统计分析建模领域进行了大量的科研工作。在成分数据的回归分析方面，2003、2006年，王惠文等提出将成分数据对称的对数比变换与偏最小二乘回归、偏最小二乘通径分析方法相结合，用以解决一元成分数据、多元成分数据的线性回归问题。本书的研究内容正是在此基础上展开并深入的，主要建模思路沿袭了将成分数据对数比变换与偏最小二乘相关分析技术相结合的做法。同时，本书的理论研究部分主要借鉴了 Aitchison(1986)所提出的成分数据对数衬度主分量分析模型(Logcontrast PCA)，Hinkle 和 Rayens(1995)所提出的对数衬度偏最小二乘回归分析模型(Logcontrast PLS)，在研究过程中特别注意遵循成分数据的基本代数理论性质要求，对已有分析方法进行理论上的深化和完善。

1.3 偏最小二乘分析技术概述

在本书有关成分数据的多元分析方法的研究中，大量工作涉及到应用偏最小二乘相关分析技术来解决成分数据分析中的特殊困难。为此，下面对偏最小二乘回归方法及其相关技术的拓展作一简要概述。

1.3.1 偏最小二乘回归方法

无论在经济管理、社会科学还是在工程技术中，多元线性回归分析都是一种普遍应用的统计分析与预测技术。然而，当自变量存在高度相关时，回归结果会出现许多反常现象，引发一系列应用方面的困难，具体归纳如下。

① 在自变量完全相关的情况下，最小二乘的回归系数完全无法估计。

② 如果自变量之间存在着不完全的共线现象，则回归系数是可以估计的；但是，回归系数的估计方差将随着自变量之间的相关程度的不断增强而迅速扩大。