对应分析数学模型及其应用

陶凤梅　韩　燕
刘　洪　杨毅恒　编著

科学出版社
www.sciencep.com

对应分析数学模型及其应用

陶凤梅　韩　燕
刘　洪　杨毅恒　编著

科学出版社
北京

内 容 简 介

本书以Guttman的内部一致性准则作为对应分析的基本数学模型，介绍了若干种与之等价的数学模型，讨论了对应分析与主成分分析之间的关系．对于有序数据和多维表数据，介绍了对应分析的具体算法．书中以专门一章介绍了对应分析的变量选择方法(逐步对应分析)，并以若干应用实例证明了它的功效．

本书既可作为统计学专业本科生和研究生的教学参考书，又可为与应用统计有关各领域的科研工作者和工程技术人员提供参考．

图书在版编目（CIP）数据

对应分析数学模型及其应用 / 陶凤梅等编著．—北京：科学出版社，2008
ISBN 978-7-03-021628-1

Ⅰ.对… Ⅱ.陶… Ⅲ.系统分析-数学模型 Ⅳ.N945.12

中国版本图书馆CIP数据核字（2008）第050648号

责任编辑：陈玉琢 房 阳／责任校对：陈玉凤
责任印制：赵德静／封面设计：王 浩

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

双青印刷厂印刷

科学出版社发行 各地新华书店经销

*

2008年6月第 一 版 开本：B5(720×1000)
2008年6月第一次印刷 印张：8 1/4
印数：1—3 000 字数：151 000

定价：28.00元

（如有印装质量问题，我社负责调换<长虹>）

前　言

人们在自然科学和社会科学的许多研究工作中，常常需要分析处理含有多个变量(包括定量变量、定性变量)的数据问题，研究探索多元关系的复杂性. 对应分析作为多元统计分析的一个重要内容，自 20 世纪 70 年代传入我国以来，被广泛应用于地质、医学、环保、生物、经济等各个领域，取得了很多有实际意义的成果. 国内大多有关文献所采用的数学模型都源于法国学者 Benzécri，我们感到这个模型不太直观，令初学者难以理解.

研读 Nishisato 关于对偶标度的书后发现，自 20 世纪 30 年代以来,有许多统计学家分别从不同角度独立地研究了这类数学模型和计算准则，各自建立了一种新的统计方法，并冠以不同的名字，但这些数学模型基本等价，计算结果基本一致. 本书参考 Nishisato 的书，用 Guttman 的内部一致性准则作为对应分析的基本数学模型，介绍了若干等价模型，讨论了对应分析与主成分分析的关系，针对在实际应用中有重要意义的有序数据和多维表数据，介绍了对应分析的具体算法. 用专门一章介绍了对应分析的变量选择方法(逐步对应分析)，在不同领域中的应用实例证明这个方法是成功的、有效的.

本书的完成是作者们多年来友好合作的结果. 这个过程得到了我们的老师夏立显教授的热情指导，没有他的支持，本书难以成稿，同时吉林大学范继璋教授以及其他的同仁也给予了热情支持和帮助，作者在此谨向他们表示衷心的谢意；同时感谢科学出版社的同志对本书出版所做的工作.

作　者

2007 年 7 月

目　录

绪　　论

0.1　对应分析所研究的问题

作为一种特殊的多元统计分析方法, 对应分析所研究的原始数据为如下形式的矩阵:

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix},$$

其中, x_{ij} 表示第 i 个样品 $(i=1,2,\cdots,n)$ 在第 j 个变量 $(j=1,2,\cdots,m)$ 上的观测值. 这些数据一般是定性的, 也可以是定量的. 假定 $x_{ij}\geqslant 0$, 且

$$f_i=\sum_{j=1}^{m}x_{ij}>0,\quad g_j=\sum_{i=1}^{n}x_{ij}>0,\quad i=1,2,\cdots,n;\quad j=1,2,\cdots,m.$$

例如, n 个样品可以是在不同地点采集的标本, 变量可以是它们的 m 种化学元素的含量 (定量变量). 再如, n 个样品也可以是不同的地质单元, 变量是刻画这些单元的 m 个地质特征 (可以是构造、岩性等定性变量).

我们的目的是求诸变量的加权向量 $\boldsymbol{a}=(a_1,a_2,\cdots,a_m)'$, 使得由此算得的平均样品得分, 即诸变量的线性组合

$$v_i=\sum_{j=1}^{m}\frac{x_{ij}a_j}{f_i},\quad i=1,2,\cdots,n \tag{0.1}$$

满足一定的准则: 各分量间有尽可能大的偏差.

这类问题有重要的实际意义和广泛的应用价值. 在前面所述的问题中, 这相当于将原有的 m 个变量 (多指标问题) 线性地组合成单变量, 便于将各样品很好地区分开来, 这类似于主成分分析, 但这里的对应分析可以处理各种定性变量. 为此, 这里举两个简单的实际例子.

例 0.1　假定 6 支足球队的联赛结果如表 0.1 和表 0.2 所示. 表中所列的比赛胜负结果是客观存在的事实, 就是所要研究的原始数据 (这里 n=6, m=3), 而各队的实际得分取决于足协规定的对各种结果的加权. 由此算出的各队得分结果有两个

缺点：①没有将各队优劣充分显示出来，有些队得分相同，只好再以净胜球数排序，这加进了人为因素；为了鼓励进攻，若只按进球数排序，可能得到不同的结果；②对各种结果采用不同的加权向量，如 $\boldsymbol{a}_1=(3,1,0)'$ 和 $\boldsymbol{a}_2=(1,0,-1)'$，算出的各队得分就可能有不同的排序结果.

表 0.1　6 支球队的联赛结果 (1)

成绩 / 球队	胜 (3 分)	平 (1 分)	负 (0 分)	得分
1	3	0	2	9
5	2	3	0	9
3	1	4	0	7
6	2	1	2	7
2	2	0	3	6
4	0	2	3	2

表 0.2　6 支球队的联赛结果 (2)

成绩 / 球队	胜 (1 分)	平 (0 分)	负 (−1 分)	得分
5	2	3	0	2
1	3	0	2	1
3	1	4	0	1
6	2	1	2	0
2	2	0	3	−1
4	0	2	3	−3

用对应分析方法可以根据实际比赛结果算出 “最佳” 的，而不是人为设定的加权向量 $\boldsymbol{a}=(a_1,a_2,a_3)'$，以使由此算出的各队得分有尽可能大的差异，最终得到客观上合理的排序.

例 0.2　询问 10 位顾客对 3 种商品的评价，分为好，一般，不好三个档次. 结果如表 0.3 所示，表中数据为每种商品获得各种评价的顾客人数. 前两种商品都得到了 10 位顾客的评价，而第 3 种商品仅得到 9 位顾客的评价，有一位顾客没有表态. 这里的原始数据组成一个 3×3 阶矩阵，我们的目的是要对三种评价 (变量) 给出加权向量 $\boldsymbol{a}=(a_1,a_2,a_3)'$，由此算得 3 种商品的得分，以便对它们的优劣给出综合的排序.

表 0.3 10 位顾客对 3 种商品的评价

商品 \ 档次	好	一般	不好	$\sum$
1	1	3	6	10
2	3	5	2	10
3	6	3	0	9
$\sum$	10	11	8	29

0.2 定性资料

前面已经说过, 对应分析的一个重要功效是可以处理各种定性资料 (categorical data). 这种资料的特点是可用少数有限个数值或字母表示测试结果, 就这些数值所给出的信息来说, 定性资料分成如下两种类型:

1) 名义的 (nominal) 定性资料

它的不同数值仅表示状态的差异, 而没有量的概念或顺序的含意. 例如,

性别: (1) 男, (2) 女;

职业: (1) 工人, (2) 农民, (3) 军人, (4) 教师;

岩性: (1) 灰岩, (2) 砂岩, (3) 页岩.

2) 有序的 (ordinal) 定性资料

它的不同数值大小不仅表示状态的差异, 而且反映了一定的顺序关系. 例如,

年龄: (1) 老年, (2) 中年, (3) 青年;

家庭年收入: (1) 高, (2) 中, (3) 低;

对某客观对象的评价: (1) 很好, (2) 好, (3) 一般, (4) 坏, (5) 很坏.

有些定性资料 (如名义资料及对某客观对象的评价) 在本质上是自然形成的, 它们来自于对定性变量的观测, 只能取有限个数值. 而有些定性资料实质上来自于对定量变量的观测, 它的数值代表定量变量的不同区段, 如年龄和家庭收入等. 用定性资料研究定量变量, 实质上是用模糊的定性概念刻画那些比较困难而又不必精确计量的定量变量. 这样做的优点是数据资料便于提取和储存, 但要以信息的损失为代价. 对应分析可以帮助我们解决这个矛盾, 从有限的定性资料中计算提取出关于原始定量变量的大量信息.

针对不同的实际问题, 常把定性资料整理成不同形式的数表, 以构成前段所述的原始数据矩阵 $\boldsymbol{X} = [x_{ij}]_{n\times m}$, 这里列出如下的几种常见的形式.

0.2.1 列联表

表 0.4 列出了两个定性随机变量 X 和 Y 的各种可能状态组合而成的矩形表.

对 $n..$ 个对象观测在两个变量上的取值, 诸网格上的数据 n_{ij} 表示观测结果为 $X=i(i=1,2,\cdots,r), Y=j(j=1,2,\cdots,c)$ 的频数, 这样的数据表就是一个 $r\times c$ 的二维列联表. 例如, X 表示学生的 r 个可能的原籍, 即 $n..$ 个学生来自 r 个地区中的一个, Y 表示学生属于 c 个可能的学院, 每个学生属于其中的一个. n_{ij} 表示来自第 i 地区, 属于第 j 学院的学生人数, $n..$ 表示受调查的学生总数.

这里仅考虑观测对象的两种属性 (X, Y), 若考虑 $n\geqslant 2$ 种属性, 就可以得到相应的 n 维列联表 (表 0.4).

表 0.4　$r\times c$ 二维列联表

X \ Y	1	2	$\cdots$	c	$\sum$
1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\cdot}$
2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r\cdot}$
$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot c}$	$n..$

$$
\begin{aligned}
n_{i\cdot} &= \sum_{j=1}^{c} n_{ij}, \quad i=1,2,\cdots,r,\\
n_{\cdot j} &= \sum_{i=1}^{r} n_{ij}, \quad j=1,2,\cdots,c,\\
n.. &= \sum_{i=1}^{r} n_{i\cdot} = \sum_{j=1}^{c} n_{\cdot j}\ .
\end{aligned}
$$

0.2.2　反映频数表

这是一种研究对象和反映类型构成的交叉表. 例如, 设从 5 个学院 a,b,c,d,e 中各自随机地抽取若干学生, 对某种商品进行评价, 学生的回答分为 4 个档次: ①不好; ②一般; ③好; ④很好. 5 个学院的学生在 4 个档次上的反映频数 (学生人数) 如表 0.5 所示. 由此可以研究各学院学生对这种商品的喜好倾向.

0.2.3　项目、类目反映表

这是数量化理论[2] 中所处理的原始数据表, 这里仅用数值例子给予解释. 表 0.6 中列出了 50 个研究对象, 对于每个对象考察它在 3 个项目上的反映, 而这 3 个项目分别含有 2,3,3 个类目. 每个研究对象在每个项目中有且仅有一个类目上有反映 (取值为 1), 在其余类目上没有反映 (取值为 0), 这说明每个项目所含的各类目都是互斥的. 由于每个项目中的各类目上的反映数据具有相关性, 可以从每个项目中删

去一个类目, 以得到一个更加紧凑的数据表 0.7.

表 0.5 反映频数表

学院 \ 类型	(1)	(2)	(3)	(4)	$\sum$
a	5	4	32	10	51
b	0	3	14	30	47
c	20	23	5	3	51
d	2	35	10	1	48
e	4	22	23	7	56
$\sum$	31	87	84	51	253

表 0.6 项目类目反映表

对象 \ 项目 类目	1		2			3			$\sum$
	(1)	(2)	(1)	(2)	(3)	(1)	(2)	(3)	
1	0	1	0	1	0	0	0	1	3
2	0	1	1	0	0	1	0	0	3
3	1	0	0	0	1	0	1	0	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	1	0	1	0	0	0	0	1	3
$\sum$	30	20	18	19	13	15	16	19	150

表 0.7 紧凑的项目类目反映表

对象 \ 项目 类目	1	2		3		$\sum$
	(1)	(1)	(2)	(1)	(2)	
1	0	0	1	0	0	1
2	0	1	0	1	0	2
3	1	0	0	0	1	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	1	1	0	0	0	2
$\sum$	30	18	19	15	16	98

表 0.7 启示我们, 可以不考虑项目, 直接研究诸对象在若干类目上的反映. 表 0.8 给出了 7 个研究对象在 5 个类目上的反映, 这种表在实际上更灵活, 有时更有用处.

0.2.4 秩顺序表

一个典型的秩顺序问题是要求 N 个人对 n 个对象进行排序. 例如; 要求 N=6 位顾客对 n=3 种商品排序, 结果可构成一个 6×3 阶矩阵, 如表 0.9 所示, 它给出了

这个排序问题的完整的信息. 若不考虑每位顾客的具体排序, 只考虑 3 种商品所得到的排序结果, 可统计每种商品得到各种排序的频数, 由表 0.9 可算得简单的商品 - 排序表, 如表 0.10 所示.

表 0.8　类目反映表

对象 \ 类型	1	2	3	4	5	$\sum$
1	1	0	0	1	0	2
2	1	1	0	0	1	3
3	0	1	1	0	0	2
4	0	0	1	1	0	2
5	1	0	1	0	1	3
6	1	0	0	0	1	2
7	0	1	0	1	0	2
$\sum$	4	3	3	3	3	16

表 0.9　6 位顾客对 3 种商品的排序

顾客 \ 商品	1	2	3
1	2	1	3
2	3	1	2
3	1	2	3
4	3	2	1
5	1	3	2
6	1	2	3
$\sum$	11	11	14

表 0.10　商品 - 排序表

商品 \ 排序	1	2	3	$\sum$
1	3	1	2	6
2	2	3	1	6
3	1	2	3	6
$\sum$	6	6	6	

表 0.10 和表 0.3 的数据似乎有相似之处, 但却有本质的差异. 表 0.3 所列的是顾客对每种商品的评价, 允许不同商品得到一位顾客相同的评价; 表 0.10 来自表 0.9, 要求每位顾客对不同商品给出不同的排序. 例如, 某位顾客对 3 种商品可以有 (1,1,2) 的评价, 却不允许有 (1,1,2) 的排序.

当被排序的对象数 n 较大时, 要求每人给出 n 个对象的完整排序是困难的. 一个简化的方法是只要求每个人对每两个对象给出一个比较排序, 由此得到成对比较数据. 例如, 设有 n=4 种商品, 它们共可构成 $\frac{1}{2}n(n-1)=6$ 对, 要求 8 个顾客就每对商品进行比较, 可将每对商品 (e_j,e_k), $1\leqslant j<k\leqslant 4$ 看成一个类目, 若某个顾客认为 $e_j<e_k$, 就确定该顾客在这类目上的反映为 1; 否则, 若该顾客认为 $e_j>e_k$, 就将这反映确定为 0. 由此得到一个类目反映表 0.11.

表 0.11 成对比较数据的类目反映表

商品对 / 顾客	(e_1,e_2)	(e_1,e_3)	(e_1,e_4)	(e_2,e_3)	(e_2,e_4)	(e_3,e_4)
1	1	1	0	1	0	1
2	0	1	0	0	1	0
3	1	0	1	0	1	1
4	1	1	1	1	0	1
5	0	1	0	0	1	1
6	0	0	0	1	0	1
7	1	1	0	0	1	0
8	0	1	1	0	0	1

0.2.5 多维表

当研究多个定性变量的观测资料时, 将前边介绍的数表适当地进行组合, 就可以得到多维数据表.

例如, 在前边讨论过的反映频数表 0.5 的基础上, 还要求学生对商品的价格进行评价, 结果可能是: (5) 便宜, (6) 适中, (7) 贵. 可得形如表 0.12 的多维表. 若根据每个学院学生对商品质量和价格的反映频数, 构成一个 3×4 二维列联表, 可得由 5 个列联表组成的多维表 0.13.

表 0.12 多维反映频数表

类型 / 学院	质量				价格			$\sum$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
a	5	4	32	10	21	16	14	102
b	0	3	14	30	10	21	16	94
c	20	23	5	3	23	20	8	102
d	2	35	10	1	21	15	12	96
e	4	22	23	7	22	18	16	112
$\sum$	31	87	84	51	97	90	66	306

表 0.13　多个列联表的组合

学院 \ 质量 / 价格		(1)	(2)	(3)	(4)
	(5)	3	1	15	2
a	(6)	1	2	10	3
	(7)	1	1	7	5
⋮			⋮		
	(5)	1	10	8	3
e	(6)	2	4	10	2
	(7)	1	8	5	2

0.3　简单的历史回顾

由于对应分析的问题简单明确, 它的实际背景可以涉及自然科学和社会科学的许多领域, 从 20 世纪 30 年代到 70 年代, 许多著名的统计学家都参与反复地研究它的数学模型和计算准则, 各自独立地建立了一种新的统计方法, 并冠以不同的名字: 互平均方法 (the method of reciprocal averages), 可加得分 (additive scoring), 恰当得分 (appropreate scoring), 典型得分 (canonical scoring), Guttman 加权 (Guttman weighting), 定性资料的主成分分析 (principal component analysis of qualitative data), 最优标度 (optimal scaling), 林知已夫的数量化理论 (Hayashi's theory of quantification), 协同线性回归 (simultaneous linear regression), 双标度 (biplot), 对应因子分析 (correspondence factor analysis), 对应分析 (correspondence analysis), 对偶标度 (dual scaling) 等. 这些方法的名字虽然不同, 但其最优化准则基本等价, 计算结果基本一致, 这是数学史上的一个罕见的事实.

这里将根据 Nishisato 的书 [19] 简述一下对应分析的发展历史.

这套方法从 20 世纪 70 年代开始以对应分析和数量化理论Ⅲ [2] 的形式传入我国, 至今已成为多元统计分析中的一个重要方法, 在各领域中得到了广泛的应用. 本书就采用国内通用的名字 —— 对应分析, 但就方法的实质来说, 有人认为采用对偶标度似乎更为恰当 [19].

对应分析的最早的奠基性工作出现于 20 世纪 30 年代:

(1) Richardson 和 Kuder 在 1933 年首先提出了互平均方法, 包含了对应分析的基本思想: 求变量权 $\boldsymbol{a}=(a_1,\cdots,a_m)'$ 和相应的样品得分向量 $\boldsymbol{V}=\boldsymbol{X}\boldsymbol{a}$, 以降低样品内的方差, 增加样品间的差异. 缺点是他们在计算方法方面存在困难.

(2) Horst 在 1935 年进一步明确了互平均方法的最优化准则, 改进了前者的计算方法, 把它用于二态变量, 以后又把这种方法用于连续变量.

(3) Hirschfeld 在 1935 年提出协同线性回归准则：给定离散的二元随机变量的分布 $p_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,m)$, 求变量值 u_i 和 v_j, 使得双方回归都是线性的. 按这种方式求出的解, 正是对应分析的解.

有趣的是这些工作一直没有引起人们的注意. 在以后的几十年间, 有许多著名统计学家仍然致力于这方面的研究, 独立地提出了许多表面上不同实质上等价的最优化准则和计算方法. 这里仅列出其中的某些重要工作.

Fisher 在 1940 年研究人的眼睛颜色与头发颜色的关系时, 求出关于两个定性变量的两组得分, 所用的方法就是前边所说的互平均, 他还指出, 每组得分是另一组得分的线性回归.

Maung 在 1941 年研究了定性变量二维表的相关性度量问题, 为对应分析提出了三个等价的准则, 即求出行和列的得分, 使能极大化：

(1) 两组得分的典型相关系数.

(2) 两组得分的乘积矩相关系数.

(3) 列间 (或行间) 离差平方和与总离差平方和之比.

同一年, Guttman 在研究多重选择数据时, 用内部一致性 (internal consistency) 作为对应分析的计算准则：求诸变量 (或 response option) 的权, 以使样品内部 (或 within subject) 离差平方和与总离差平方和之比极小化. 变量权确定以后, 再使样品得分与诸变量在该样品上的加权平均成比例. 同样的准则也可用于先计算样品权, 再求变量得分. 由此得到了对应分析变量得分与样品得分的对偶性. 他这个准则与 Maung 的准则 (3) 是等价的, 同时他也独立地得到了准则 (2). 他在 1946 年首次把这套方法用于研究成对比较数据和秩顺序数据, 扩展了对应分析的应用范围.

日本学者林知已夫 (C.Hayashi)[18] 在 20 世纪 50 年代建立了数量化理论, 系统研究了定性数据的数量化方法. 他的数量化理论Ⅲ所用的准则与 Guttman 的内部一致性准则基本一致, 但他极大地推广了 Guttman 的结果, 特别是在成对比较数据的多维数量化方面.

针对 Fisher 和 Maung 的对应分析模型, Williams 1952 年参考判别分析给出了假设检验方法. Lancaster 在 1953 年研究了将 χ^2 统计量用于假设检验的方法.

至此, 对应分析的数学模型和计算方法都以严格的形式建立起来了. 自 60 年代以后, 又有许多著名的统计学家致力于这方面的研究. 研究内容包括：软件开发 (始于 Baker 的工作), 计算方法的创新 (如 Kruskal 在 1965 年将单调回归用于对应分析), 扩大应用范围 (如用于研究成对比较数据, 多维定性数据 [15]), 改善应用效果等.

这期间在理论上的重要进展是搞清了对应分析与其他多元统计方法的关系. 在 McKeon 的专著中, 讨论了对应分析与典型相关分析、因子分析、判别分析的关系. McDonald 提出获取变量加权线性组合的一般过程, 它能包含多重回归加权、典型

变量分析、主成分分析、典型因子分析以及其他的一些著名的模型, 对应分析可作为它的一个特殊情形.

特别值得注意的是法国学者 Benzécri 等的工作 [14]. 他们在 20 世纪六七十年代以法文发表了大量研究论文和著名的专著, 又以数据矩阵的重新标度为基础提出了一种新的数学模型, 首次采用了对应分析 (correspondance analysis) 的名字. 由于他们的工作被大量引用, 对应分析也就成了这类方法的比较通用的名字.

多维标度法 [11] 中的几何思想的引入, 使对应分析理论模型在近年来得到了新的发展. De Leeuw 首先开始系统研究对应分析与多维标度法之间的关系, 并把这两种方法统一到一系列有用的计算机程序中. 此后在 Gifi 等的工作中, 以多元对应分析为出发点, 建立了一类非线性多元分析方法, 将一些经典方法, 如主成分分析, 典型相关分析, 多元回归分析和判别分析等作为特殊 (线性) 情形包含其中, 所有这些特殊情形由多元对应分析中所包含的参数来决定. 这类方法的基本思想是给研究对象以适当的标度 (在低维欧几里得空间中), 使变量取值相似的研究对象有相近的标度, 这与多维标度法是一致的.

0.4　本书的基本内容

第 1 章用 Guttman 的内部一致性准则, 以矩阵分析为工具详细讨论了对应分析的数学模型, 把问题归结为互相对偶的两个特征方程的求解. 这种模型的数学原理比较清楚, 可以很容易地推出对应分析解的对偶性, 在国内的多元分析著作中尚未见到. 特别地, 在论证中提出了对应分析与单因素方差分析的关系, 加深了对这种模型的认识.

第 2 章参考 Nishisato 的书 [19], 讨论了对应分析的另外 5 个等价模型, 从中可以看出它与其他多元分析方法的关系. 特别地, 针对国内文献普遍采用的不便理解的 Benzécri 模型, 给出了它与第 1 章模型的等价性证明. 我们建议, 在对应分析的理论发展和推广应用中, 用便于理解的第 1 章模型代替 Benzécri 模型.

对应分析和主成分分析在许多方面有相似之处, 第 3 章深入讨论了这两种方法之间的关系. 这里用内部一致性准则 [18] 重新讨论了主成分分析的数学模型, 从中可以看出它与对应分析的异同. 特别指出, 在共同性准则之下 [13,16], 两种方法在一定程度上得到了统一.

第 4 章试图从新的角度, 即通过变量选择和样品选择来研究对应分析的数学模型. 我们知道, 对应分析所研究的原始数据是一个 $n \times m$ 阶矩阵 $\boldsymbol{X} = [x_{ij}]_{n\times m}$, 它的含意是: 用 m 个变量刻画 n 个样品, 或用 n 个样品表示 m 个变量. 值得注意的是: 就许多实际问题来说, 在 m 个变量 (n 个样品) 中, 仅有少数变量 (样品) 是重要

的, 剔除那些不重要的变量 (样品), 照样可以很好地刻画 n 个样品 (m 个变量).

这种情形类似于逐步回归分析和逐步判别分析. 回归分析和判别分析都是有因变量的多元统计方法, 自变量选择和剔除的根据是它与因变量的关系, 仅保留那些与因变量关系密切的变量, 剔除那些与因变量关系不密切的变量. 对应分析没有因变量可作为选择或剔除变量的依据. 但由 m 个原始变量 (n 个原始样品) 用对应分析算出的 n 个样品 (m 个变量) 的空间构形, 可以作为选择变量 (样品) 的依据. 于是可以选择那样的一些少数变量 (样品), 用它们算出的 n 个样品 (m 个变量) 的空间构形与用原来的全体变量 (样品) 算出的构形尽可能地相近. 基于这种空间构形的相近程度的计量, 本书介绍了用三个选择变量的准则, 按逐步剔除不重要变量, 或逐步选入重要变量的计算步骤达到选择变量的目的. 四个成功的实际计算实例说明这种变量选择方法是行之有效的. 我们在这里仅讨论了对应分析的变量选择方法, 但由对应分析的对偶性, 这些方法完全可以适用于样品选择, 这里不再赘述. 统称这些方法为逐步对应分析.