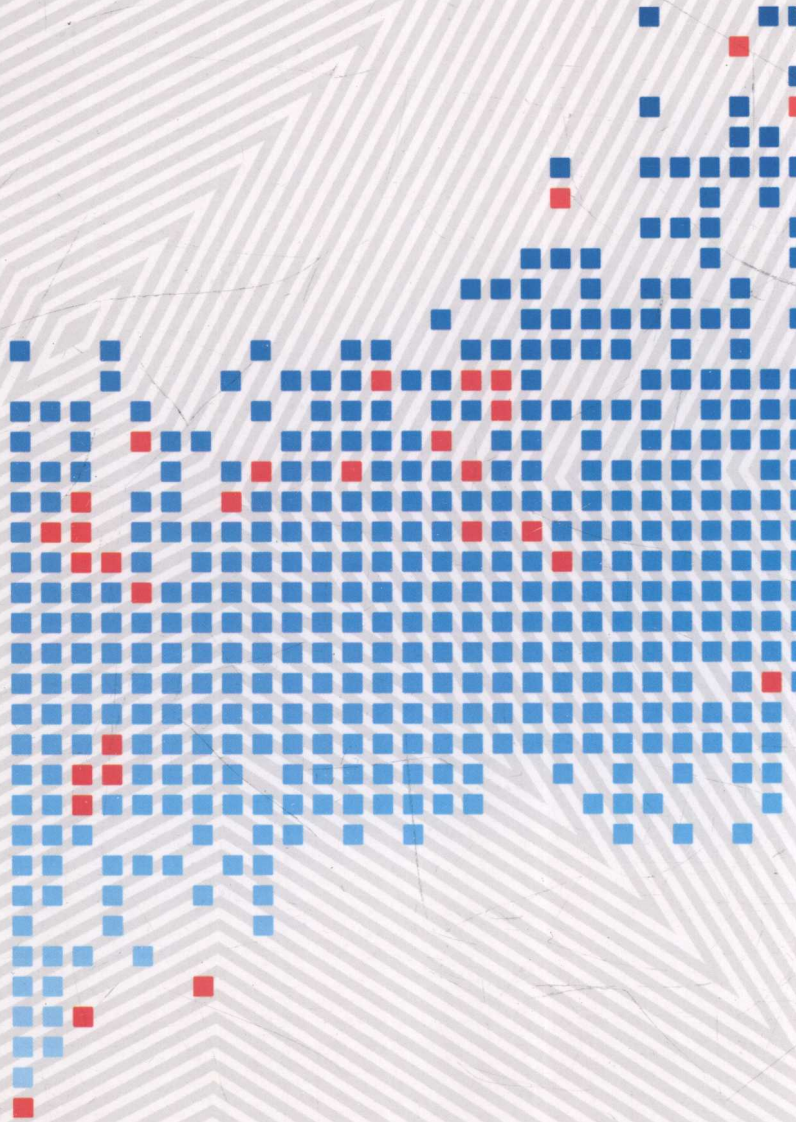
Principle and Application of
Multivariate Analysis

多元分析原理及应用

李永江　刘春雨　著

Statistics



经济科学出版社
Economic Science Press

Principle and Application of
Multivariate Analysis

多元分析原理及应用

李永江　刘春雨　著

Statistics

经济科学出版社
Economic Science Press

图书在版编目（CIP）数据

多元分析原理及应用/李永江，刘春雨著. —北京：
经济科学出版社，2016.2
ISBN 978-7-5141-6606-4

Ⅰ.①多… Ⅱ.①李…②刘… Ⅲ.①多元分析-研究
Ⅳ.①O212.4

中国版本图书馆 CIP 数据核字（2016）第 030106 号

责任编辑：刘　莎
责任校对：刘　昕
责任印制：邱　天

多元分析原理及应用

李永江　刘春雨　著
经济科学出版社出版、发行　新华书店经销
社址：北京市海淀区阜成路甲 28 号　邮编：100142
总编部电话：010-88191217　发行部电话：010-88191522
网址：www.esp.com.cn
电子邮件：esp@esp.com.cn
天猫网店：经济科学出版社旗舰店
网址：http://jjkxcbs.tmall.com
北京密兴印刷有限公司印装
710×1000　16 开　14.75 印张　300000 字
2016 年 3 月第 1 版　2016 年 3 月第 1 次印刷
ISBN 978-7-5141-6606-4　定价：52.00 元
（图书出现印装问题，本社负责调换。电话：010-88191502）

前　言

多元分析是多元统计分析的简称，是数理统计近几十年来发展起来的最重要的分支之一，学界的很多学者也将其纳入“现代分析技术”的重要组成部分。它在社会经济、工农业、生物医药、环保、气象等各个领域都有着广泛的应用。近年来，随着计算技术和信息技术的高度发展，尤其是大数据时代到来的今天，这门学科越来越展现出其无穷的魅力。

时代在前进，学科在发展，读者的需求也在发生变化。为了与时俱进，满足大数据时代的需要，在多年教学科研积累的基础上又通过近几年广泛而深入的调研与科学的计算分析，我们不失时机地写出了这本书。撰写本书的指导思想主要基于以下三点：首先，假定读者已经具备线性代数以及一元分析（初级数理统计）的基本知识；其次，读者可以是大学里统计学及其相关专业的高年级本科生或研究生，也可以是能用到多元分析技术的实际工作者；最后，这些读者通过阅读本书，会比较轻松地理解并掌握多元分析的基本原理，而他们注重的并非原理本身而是其实际应用。我们认为这是一个庞大的读者群，因此希望本书能够很好地满足这些读者们的需要，并希望读者能在使用过程中提出宝贵意见。

与国内外同类书籍相比，本书具有以下鲜明的特点：一是系统性，本书系统地阐述了多元分析这一新兴学科的起源与发展、曲折与完善、理论与实践的八十年历程。二是全面性，本书全面地阐述了多元分析这一新兴学科所包含的各种统计分析方法的基本原理，而且每一章都配有习题并在书后附有参考答案。三是实用性，本书在全面阐述多元分析较为深奥的基本原理的基础上，深入浅出地讲述了各种统计分析方法的理解方法和计算步骤，真正做到了易学易用。四是操作性，本书指出了所述的各种统计分析方法的计算机实现途径。五是前瞻性，本书通过案例的形式，向读者展示了多元分析在学科的应用方面将具有广阔而美好的发展空间。

青岛理工大学　李永江
广东东软学院　刘春雨
2016 年 3 月

目　录

上篇　绪　论　篇

中篇　原　理　篇

下篇　应　用　篇

上篇　绪　论　篇

一、多元分析的起源及发展

多元分析是多元统计分析的简称，起源于20世纪初。1928年维希特（Wishart）发表论文《多元正态总体样本协差阵的精确分布》，被学术界认为是多元分析的开端。20世纪30年代，费雪（R. A. Fisher）、霍特林（H. Hotelling）、罗伊（S. N. Roy）、许宝碌等人做了一系列奠基性工作，使多元分析在理论上得到了迅速的发展。40年代，在心理学、教育学、生物学等方面有不少的应用，但由于计算量大，使其发展受到影响，甚至停滞了相当长的时间。50年代中期，随着电子计算机的出现和发展，使多元分析方法在地质学、气象学、医学、社会学等方面得到广泛的应用。60年代，通过应用和实践又完善和发展了理论，由于新理论、新方法的不断涌现又促使它的应用范围更加扩大。到了70年代，多元分析在我国才受到各个领域的极大关注，四十余年来，我国在多元分析的理论研究和应用上也取得了很多显著成绩，有些研究工作已达到国际水平。

二、多元分析的基本思想

在经济、管理等许多领域，我们常常需要同时观测多个指标。比如，要了解一个国家或地区的经济发展状况，需要观测人均国民收入、人均第一产业产值、人均第二产业产值、人均第三产业产值、人均可支配收入、人均消费水平等。在多元分析中，我们总是把以下三个概念等同起来的：指标、变量、随机变量（实际上多元分析所涉及的指标、变量均是带有很强的随机性的）。如何对多个随机变量的观测数据进行有效的分析和研究呢？一种做法是把多个随机变量分开，一次处理一个地去分析研究；另一种做法是把多个随机变量作为一个整体（即随机向量）进行分析研究。显然前者的做法有时是有效的，但一般来说，由于变量

多，避免不了变量之间的相关性，如果分开处理不仅会丢失很多信息，往往也不容易取得好的研究结果。而后一种做法通常可以用多元分析的方法来解决，通过对多个随机变量观测数据的分析，来研究变量之间的相互关系以及揭示这些变量内在的变化规律。如果说一元统计分析是研究一个随机变量统计规律的学科，那么多元统计分析则是研究多个随机变量之间相互依赖关系以及内在统计规律性的一门统计学科。

三、多元分析的基本应用

从以下列举的实际问题中，我们不仅可以看到多元分析能解决哪些不同类型的实际问题，而且还可以看到多元分析是具有非常广阔和深入的应用前景的。

在社会学中，可以根据人口密度、人均收入、人均支出、居住面积、成人识字率、城市绿化覆盖率等这些指标的数据对31个省市自治区利用聚类分析进行分类，再根据分类结果对社会情况进行综合评价。在经济学中，可根据人均国民收入、人均农业产值、人均消费水平等多种指标利用判别分析判定一个国家或地区的经济发展程度所属类型。在管理学中，对若干地区的众多指标如固定资产投资额、银行贷款余额、职工工资总额、全员劳动生产率、产品成本降低率、资金利税率、万元产值能耗等等进行主成分分析和因子分析，将这些具有错综复杂关系的指标综合成几个较少的因子，既有利于对问题进行分析和解释，又能便于抓住主要矛盾。在教育学中，如何对高考考生成绩做因素分析？学生入学后的考试成绩和入学考试各门成绩有何相关关系？在地质学中，如何根据演示标本的多种特征来判别地层的地质年代，是有矿还是无矿，是铜矿还是铁矿等。考古学中，根据挖掘出来的人头盖骨的高、宽等特征来判断其是男或女；根据挖掘的动物牙齿的有关测试指标，判别它是属于哪一类动物牙齿、是哪一个时代的？气象学中，根据气象站的观测记录，利用气象指标气温、气压、湿度等之间关系对降雨量做预报。

当然，多元分析的方法还可以应用到其他生产和科研的各个领域，在此就不一一列举，请读者详见本书的下篇（应用篇）。

中篇　原　理　篇

第一章

多元正态分布

一元正态分布在统计学的理论和实际应用方面都有着重要的地位。同样，在多元分析中，多元正态分布也占有十分重要的位置。原因有三：一是许多实际问题中的随机向量服从或近似服从正态分布；二是有关正态分布的理论最完善；三是许多统计分析方法都是在正态分布假定下进行的。

第一节　基本概念

这里给出的基本概念并非仅限于多元正态分布，是适合于多元分析的全部理论的。

一、随机向量

在多元分析中，仍将所研究对象的全体称为总体。如果构成总体的个体具有 p 个需要观测的指标，我们称这样的总体为 p 维（元）总体。因为 p 维总体中随机抽取一个个体，其 p 个指标观测值是不能事先精确知道的，所以 p 维总体可用一个 p 维随机向量来表示。

定义 1.1　设 X_1、X_2、…、X_p 为 p 个随机变量，由它们组成的向量称作随机向量，记作 $X=(X_1, X_2, \cdots, X_p)'$。

二、概率分布

描述随机变量的最基本工具是分布函数，类似地描述随机向量的最基本工具还是分布函数。区别在于前者是一元函数，而后者是多元函数。

定义 1.2　设 $X=(X_1, X_2, \cdots, X_p)'$ 是 p 维随机向量，称

$$F(x) \triangleq F(x_1, x_2, \cdots, x_p) = P(X_1 \leqslant x_1, X_2 \leqslant x_2, \cdots, X_p \leqslant x_p)$$

为 X 的分布函数，式中 $x = (x_1, x_2, \cdots, x_p) \in R^p$，并记 $X \sim F(x)$。

定义 1.3　设 $X = (X_1, X_2, \cdots, X_p)'$ 是 p 维随机向量，若

$$P(X = x_k) = p_k (k = 1, 2, \cdots) \text{ 且满足 } p_1 + p_2 + \cdots = 1$$

则称 X 为离散型随机向量，并称 $P(X = x_k) = p_k (k = 1, 2, \cdots)$ 为 X 的概率分布，值得注意的是，这里的 x_k 为 p 维向量。

定义 1.4　设 $X \sim F(x) = F(x_1, x_2, \cdots, x_p)$，若存在非负可积函数 $f(\cdot)$ 使得

$$F(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} f(t_1, t_2, \cdots, t_p)\, dt_1 dt_2 \cdots dt_p$$

对一切 $x \in R^p$ 成立，则称 X 为连续型随机向量，并称 $f(\cdot)$ 为密度函数。

【注】　一个 p 元函数 $f(\cdot)$ 能成为某个随机向量的密度函数的充要条件是

$f(x) \geqslant 0$，且 $\forall x \in R^p$ 有 $\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1, x_2, \cdots, x_p)\, dx_1 dx_2 \cdots dx_p = 1$。

三、边际分布

定义 1.5　设 $X = (X_1, X_2, \cdots, X_p)'$ 是 p 维随机向量，称由它的 q 个分量组成的子向量 $X^{(i)} = (X_{i_1}, X_{i_2}, \cdots, X_{i_q})'$ 的分布为 X 的边缘（边际）分布，相对地把 X 的分布称为联合分布。通过变换 X 中各分量的次序，总可假定 $X^{(1)}$ 正好是 X 的前 q 个分量（这种假定会为我们的描述带来方便），而其余 $p - q$ 个分量为 $X^{(2)}$，即

$X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}_{p-q}^{q}$，相应的取值也可分为两部分 $x = \begin{bmatrix} x^{(1)} \\ x^{(2)} \end{bmatrix}$。

当 X 的分布函数是 $F(x_1, x_2, \cdots, x_p)$ 时，$X^{(1)}$ 的分布函数（边际分布）为

$$\begin{aligned} F_1(x_1, x_2, \cdots, x_q) &= P(X_1 \leqslant x_1, X_2 \leqslant x_2, \cdots, X_q \leqslant x_q) \\ &= P(X_1 \leqslant x_1, \cdots, X_q \leqslant x_q, X_{q+1} \leqslant +\infty, \cdots, X_p \leqslant +\infty) \\ &= F(x_1, x_2, \cdots, x_q, +\infty, \cdots, +\infty)。 \end{aligned}$$

当 X 有密度函数 $f(x_1, x_2, \cdots, x_p)$ 时，则 $X^{(1)}$ 也有密度函数（边际密度）

$$f_1(x_1, x_2, \cdots, x_q) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1, x_2, \cdots, x_p)\, dx_{q+1} dx_{q+2} dx_p。$$

类似地，$X^{(2)}$ 的分布函数及密度函数分别为

$$F_2(x_{q+1}, x_{q+2}, \cdots, x_p) = F(+\infty, \cdots, +\infty, x_{q+1}, x_{q+2}, \cdots, x_p),$$

$$f_2(x_{q+1}, x_{q+2}, \cdots, x_p) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1, x_2, \cdots, x_p)\, dx_1 dx_2 \cdots dx_q。$$

四、条件分布

定义 1.6 设 $X=(X_1, X_2, \cdots, X_p)'$是连续型随机向量，称

$$f(x_1, x_2, \cdots, x_q \mid x_{q+1}, x_{q+2}, \cdots, x_p)=\frac{f(x_1, x_2, \cdots, x_p)}{f_2(x_{q+1}, x_{q+2}, \cdots, x_p)}$$

是 $X^{(2)}=(X_{q+1}, X_{q+2}, \cdots, X_p)$ 条件下 $X^{(1)}=(X_1, X_2, \cdots, X_q)$ 的条件概率密度（函数）。

五、独立性

定义 1.7 若对于一切 x 和 y 都有

$$P(X\leqslant x, Y\leqslant y)=P(X\leqslant x)P(Y\leqslant y)$$

则称随机向量 X 与 Y 相互独立。

【注 1】 这里的 X、Y、x、y 都是向量，而 X 与 Y 的维数可以相同也可以不同，但 X 与 x 的维数必相同，Y 与 y 的维数也必相同。

【注 2】 若 $F(x, y)$ 为 $(X, Y)'$的联合分布函数，$F_X(x)$ 和 $F_Y(y)$ 分别为 X 和 Y 的分布函数（即边际分布函数），则 X 与 Y 独立当且仅当 $F(x, y)=F_X(x)\ F_Y(y)$。

【注 3】 若 $(X, Y)'$有密度函数 $f(x, y)$，则 X 与 Y 独立当且仅当 $f(x, y)=f_X(x)\ f_Y(y)$ 或 $f(x\mid y)=f_X(x)$。

【注 4】 若 X_1、X_2、$\cdots$、X_n 为随机向量，则 X_1、X_2、$\cdots$、X_n 相互独立当且仅当对 1、2、$\cdots$、n 的任意子序列 n_1、n_2、$\cdots$、n_k 都有

$$F(x_{n_1}, x_{n_2}, \cdots, x_{n_k})=F_{n_1}(x_{n_1})\ F_{n_2}(x_{n_2})\ \cdots F_{n_k}(x_{n_k})$$

其中 $F_{n_i}(x_{n_i})$ 表示随机向量 $X_{n_i}(i=1, 2, \cdots, k)$ 的边际分布函数。

【注 5】 若 X_1、X_2、$\cdots$、X_n 为连续型随机向量，则 X_1、X_2、$\cdots$、X_n 相互独立当且仅当对 1、2、$\cdots$、n 的任意子序列 n_1、n_2、$\cdots$、n_k 都有

$$f(x_{n_1}, x_{n_2}, \cdots, x_{n_k})=f_{n_1}(x_{n_1})\ f_{n_2}(x_{n_2})\ \cdots f_{n_k}(x_{n_k})$$

其中 $f_{n_i}(x_{n_i})$ 表示随机向量 $X_{n_i}(i=1, 2, \cdots, k)$ 的边际分布密度。

六、随机向量的数字特征（矩）

1. 数学期望

定义 1.8 设 $X=\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1q} \\ X_{21} & X_{22} & \cdots & X_{2q} \\ \vdots & \vdots & & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pq} \end{bmatrix}$是由随机变量构成的随机矩阵，定义

X 的数学期望为

$$E(X)=\begin{bmatrix} E(X_{11}) & E(X_{12}) & \cdots & E(X_{1q}) \\ E(X_{21}) & E(X_{22}) & \cdots & E(X_{2q}) \\ \vdots & \vdots & & \vdots \\ E(X_{p1}) & E(X_{p2}) & \cdots & E(X_{pq}) \end{bmatrix}。$$

特别当 $q=1$ 时便可得到随机向量 $X=(X_1, X_2, \cdots, X_p)'$ 的数学期望

$$E(X)=(E(X_1), E(X_2), \cdots, E(X_p))'。$$

由此定义容易推出以下四条基本性质。

性质 1 设 a 为常数，则

$$E(aX)=aE(X)。$$

性质 2 设 A、B、C 为常数矩阵，则

$$E(AXB+C)=AE(X)\ B+C。$$

性质 3 若 X、Y 为随机向量，A、B 为适合运算的常数矩阵，则

$$E(AX+BY)=AE(X)+BE\ (Y)。$$

性质 4 设 X_1、X_2、…、X_n 为同阶随机矩阵，则

$$E(X_1+X_2+\cdots+X_n)=E(X_1)+E(X_2)+\cdots+E(X_n)。$$

2. 协方差阵

定义 1.9 设 $X=(X_1, X_2, \cdots, X_p)'$ 和 $Y=(Y_1, Y_2, \cdots, Y_q)'$ 是两个随机向量，定义 X 和 Y 的协方差阵（简称协差阵）为

$$Cov(X, Y)=E((X-E(X)(Y-E(Y)')$$

$$=E\left[\begin{pmatrix} X_1-E(X_1) \\ X_2-E(X_2) \\ \vdots \\ X_p-E(X_p) \end{pmatrix}\ (Y_1-E(Y_1)\quad Y_2-E(Y_2)\quad \cdots\quad Y_q-E(Y_q))\right]$$

$$=\begin{bmatrix} Cov(X_1, Y_1) & Cov(X_1, Y_2) & \cdots & Cov(X_1, Y_q) \\ Cov(X_2, Y_1) & Cov(X_2, Y_2) & \cdots & Cov(X_2, Y_q) \\ \vdots & \vdots & & \vdots \\ Cov(X_p, Y_1) & Cov(X_p, Y_2) & \cdots & Cov(X_p, Y_q) \end{bmatrix}。$$

定义 1.10 随机向量 $X=(X_1, X_2, \cdots, X_p)'$ 的协差阵定义为

$$\sum=D(X)=Var(X)=V(X)$$

$$=\begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_p) \\ Cov(X_2, X_1) & Var(X_2) & \cdots & Cov(X_2, X_p) \\ \vdots & \vdots & & \vdots \\ Cov(X_p, X_1) & Cov(X_p, X_2) & \cdots & Var(X_p) \end{bmatrix}$$