Multivariate
Statistical Analysis

Statistics and Actuarial Science

中国人民大学统计与精算系列教材

多元统计分析

田茂再　编著

中国人民大学出版社

Multivariate
Statistical Analysis

Statistics and Actuarial Science

中国人民大学统计与精算系列教材

多元统计分析

田茂再　编著

中国人民大学出版社
· 北京 ·

图书在版编目（CIP）数据

多元统计分析 / 田茂再编著. — 北京 ：中国人民大学出版社，2017.4
中国人民大学统计与精算系列教材
ISBN 978-7-300-23935-4

Ⅰ.①多… Ⅱ.①田… 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（2017）第 016810 号

中国人民大学统计与精算系列教材
多元统计分析
田茂再 编著
Duoyuan Tongji Fenxi

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	**邮政编码**	100080
电　　话	010－62511242（总编室）	010－62511770（质管部）	
	010－82501766（邮购部）	010－62514148（门市部）	
	010－62515195（发行公司）	010－62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京密兴印刷有限公司		
规　　格	185 mm×260 mm　16 开本	**版　　次**	2017 年 4 月第 1 版
印　　张	13 插页 1	**印　　次**	2017 年 4 月第 1 次印刷
字　　数	328 000	**定　　价**	38.00 元

前　言

具有复杂多变量特征的数据普遍存在于现实世界中。忽视数据的这一多维度性质而进行逐个变元分析, 无异于盲人摸象, 因为这常常会使传统的统计分析方法效果不佳, 甚至失效。多元数据分析是统计科学中发展较早的一个分支, 在各学科领域越来越受重视, 许多成果较为成熟, 但相关的研究依然显得非常活跃, 以至于文献中出现了各种各样的称谓。

目前介绍经典多元统计分析方法的专著、教材不少, 但缺少能兼顾经典多元统计分析理论与现代多元统计分析理论及方法的书, 更谈不上反映复杂多元数据分析国际前沿研究的专著了。本书朝着这个目标努力, 用严格的数学语言对多元统计分析的现代面貌做较为详细的介绍。

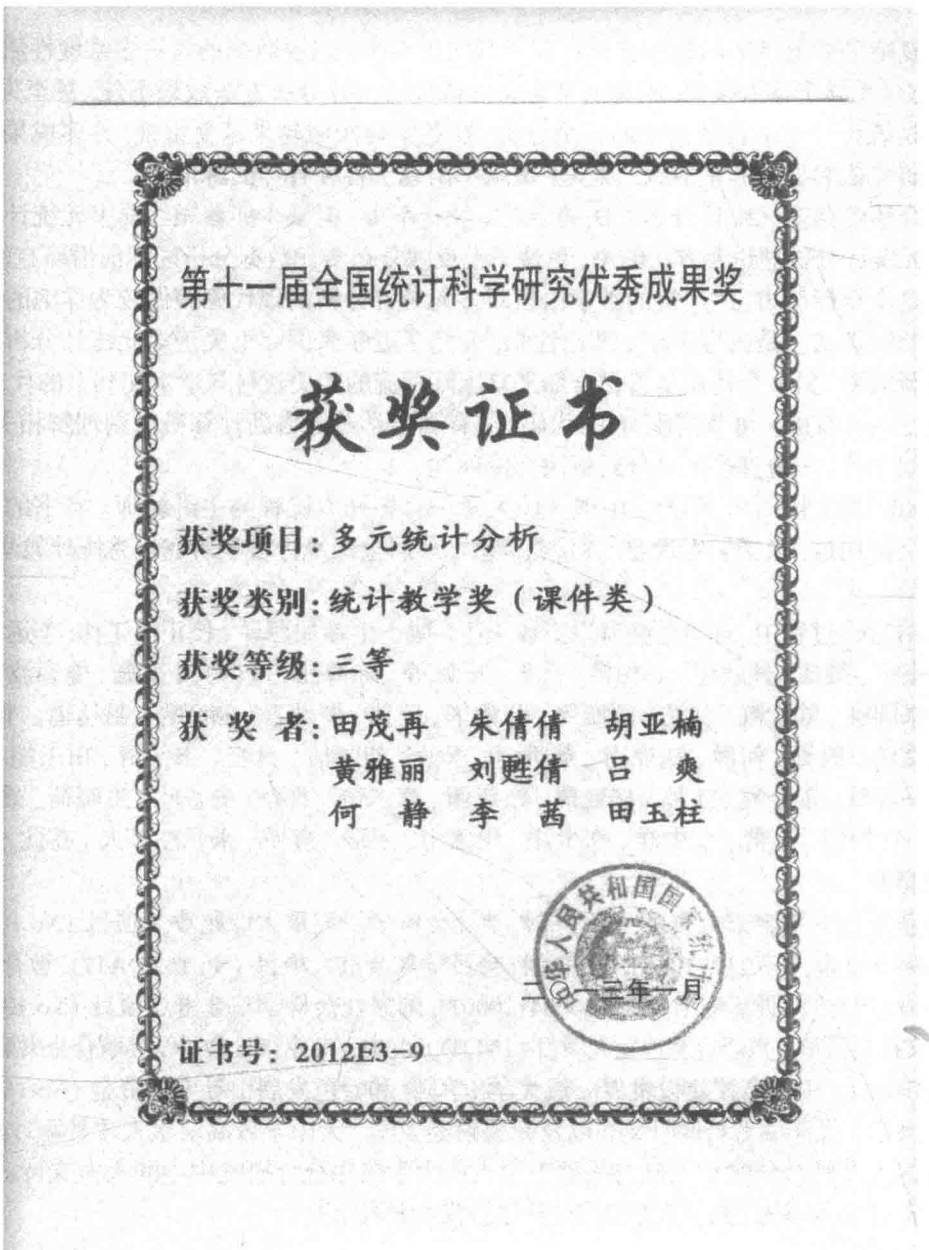
本书的特点之一是内容全新、理论性强, 介绍了近年来国际上关于多元统计分析理论与方法的许多新成果。另一个特点是素材全部来自国际一流的相关教材与学术期刊上的代表性文章, 信息量大、内容权威。每节都配有 R 代码, 这有利于学习者借助计算机深刻理解相关内容, 因为可以一边学习, 一边直接上手分析解决实际问题。

自 2007 年以来, 本书作者在中国人民大学一直担任本课程的主讲教师。本书的大部分材料在课堂上试用过, 深受学生欢迎, 并荣获 "第十一届全国统计科学研究优秀统计教学奖 (课件类)"。

在本书写作过程中, 自始至终有以下硕士生、博士生参加翻译、校正等工作: 李远、周朋朋、范洁瑜、张宁、戴成、钱政超、石恒泽、周健、安姝静、陈博钰、范博文、范燕、姜春波、马维华、苏宇楠、张圆圆、陈彦靓、郭洁、康雁飞、荣耀华、王伟、罗幼喜、储昭霁、封达道、李兆媛、司世景、夏文涛、熊巍、何静、胡亚南、黄雅丽、李茜、刘甦倩、吕爽、朱倩倩、田玉柱、梁晓琳、马春桃、马绰欣、孟令宾、王榛、杨亚琦、张亚丽、李二倩、罗静、史普欣、王晓荷、袁梦、吴延科、晏振、谷梅川、李蕾、李少洋、李少洋、田鑫涛、王珊、袁博、张元杰等人。在此, 对他们表示衷心的感谢!

本书获得以下基金部分资助: 教育部哲学社会科学研究重大课题攻关项目 (No.15JZD015), 国家自然科学基金 (No.11271368), 北京市社会科学基金重大项目 (No.15ZDA17), 教育部高等学校博士学科点专项科研基金 (No.20130004110007), 国家社会科学基金重点项目 (No.13AZD064), 教育部人文社会科学重点研究基地重大项目 (15JJD910001), 北京市社会科学界联合会决策咨询课题 (No.2016010021), 中央高校建设世界一流大学 (学科) 和特色发展引导专项资金 (No.15XNL008), 兰州财经大学 "飞天学者特聘计划" 以及新疆财经大学 "天山学者高层次人才特聘计划"。特别感谢教育部人文社会科学重点研究基地中国人民大学应用统计研究中心的大力支持。

由于作者水平有限, 错误在所难免, 甚望读者批评指正!



第十一届全国统计科学研究优秀成果奖

获奖证书

获奖项目：多元统计分析

获奖类别：统计教学奖（课件类）

获奖等级：三等

获 奖 者：田茂再　朱倩倩　胡亚楠
黄雅丽　刘甦倩　吕　爽
何　静　李　茜　田玉柱

二〇一三年一月

证书号：2012E3-9

第十一届全国统计科学研究优秀统计教学奖(课件类)

目　录

第 1 章　多元分布

1.1　分布和密度函数

本章从多元随机向量 (random vector) 入手。随机向量是由多个随机变量组成的。

令 $X=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 为一随机向量, 则 X 的累积分布函数 (cdf) 定义为:

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

对于连续型的随机变量 X, 存在非负的概率密度函数 f, 使得对于任意实数 x, 有

$$F(x)=\int_{-\infty}^{x}f(u)\mathrm{d}u \tag{1.1}$$

而且

$$\int_{-\infty}^{\infty}f(u)\mathrm{d}u=1$$

需要注意的是, 本书大多数如式 (1.1) 所示的积分都是多维的。例如, 积分 $\int_{-\infty}^{x}f(u)\mathrm{d}u$ 表示

$$\int_{-\infty}^{x_p}\cdots\int_{-\infty}^{x_2}\int_{-\infty}^{x_1}f(u_1,u_2,\cdots,u_p)\mathrm{d}u_1\mathrm{d}u_2\cdots\mathrm{d}u_p$$

此外, 对累积分布函数 F 求导即得 X 的密度函数

$$f(x)=\frac{\partial^p F(x)}{\partial x_1\partial x_2\cdots\partial x_p}$$

对于离散型的随机变量 X, 随机变量的取值是有限可数的, 或者说集中在点集 $\{c_j\}_{j\in J}$, 则事件 $\{X\in D\}$ 发生的概率即为所有可能值发生的概率之和, 即

$$P(X\in D)=\sum_{\{j:c_j\in D\}}P(X=c_j)$$

如果把随机向量 X 分解成 $X=(X_1,X_2)^{\mathrm{T}}$, 其中, $X_1\in\mathbb{R}^k,X_2\in\mathbb{R}^{p-k}$, 则称函数

$$F_{X_1}(x_1)=P(X_1\leqslant x_1)=F(x_{11},x_{12},\cdots,x_{1k},\infty,\cdots,\infty) \tag{1.2}$$

为边际累积分布函数。而 $F=F(x)$ 称为联合累积分布函数。对于连续型的随机变量 X, 边际密度函数可以通过对联合密度函数关于其他变量求积分得到

$$f_{X_1}(x_1)=\int_{-\infty}^{\infty}f(x_1,x_2)\mathrm{d}x_2 \tag{1.3}$$

在给定 $X_1=x_1$ 的条件下, 随机变量 X_2 的条件密度函数为:

$$f(x_2|x_1)=\frac{f(x_1,x_2)}{f_{X_1}(x_1)} \tag{1.4}$$

【例 1.1】 随机向量 (X_1,X_2) 的概率密度为:

$$f(x_1,x_2)=\begin{cases}\dfrac{1}{3}x_1+\dfrac{2}{3}x_2, & 0\leqslant x_1,x_2\leqslant 1\\ 0, & \text{其他}\end{cases}$$

求: (1) 边缘概率密度; (2) 条件概率密度。

解:

$$f_{X_1}(x_1)=\int f(x_1,x_2)\mathrm{d}x_2=\int_0^1\left(\frac{1}{3}x_1+\frac{2}{3}x_2\right)\mathrm{d}x_2=\frac{1}{3}x_1+\frac{1}{3}$$

$$f_{X_2}(x_2)=\int f(x_1,x_2)\mathrm{d}x_1=\int_0^1\left(\frac{1}{3}x_1+\frac{2}{3}x_2\right)\mathrm{d}x_1=\frac{2}{3}x_2+\frac{1}{6}$$

$$f(x_2|x_1)=\frac{f(x_1,x_2)}{f_{X_1}(x_1)}=\frac{\frac{1}{3}x_1+\frac{2}{3}x_2}{\frac{1}{3}x_1+\frac{1}{3}}$$

$$f(x_1|x_2)=\frac{f(x_1,x_2)}{f_{X_2}(x_2)}=\frac{\frac{1}{3}x_1+\frac{2}{3}x_2}{\frac{2}{3}x_2+\frac{1}{6}}$$

值得注意的是, 虽然 X_1,X_2 的联合密度函数是关于 x_1,x_2 的线性函数, 但是它们各自的条件密度函数却是非线性的。

两个随机变量的独立性定义如下:

【定义 1.1】 随机变量 X_1 和 X_2 相互独立的充要条件是 $f(x_1,x_2)=f_{X_1}(x_1)f_{X_2}(x_2)$。

也就是说, 如果 X_1 和 X_2 是相互独立的, 那么它们的条件密度函数就等于边缘密度函数, 即 $f(x_1|x_2)=f_{X_1}(x_1), f(x_2|x_1)=f_{X_2}(x_2)$。

另外, 需要强调的是, 不同的联合分布函数可能有相同的边缘分布函数。

【例 1.2】 随机变量 (X_1,X_2) 的概率密度为:

(1) $f(x_1,x_2)=1, 0<x_1,x_2<1$。

(2) $f(x_1,x_2)=1+\alpha(2x_1-1)(2x_2-1),\ 0<x_1,x_2<1,\ -1\leqslant\alpha\leqslant 1$。

分别计算 (1) 和 (2) 中 X_1,X_2 的边缘分布函数。

解:

(1) $f_{X_1}(x_1)=1,\ f_{X_2}(x_2)=1$

(2) $f_{X_1}(x_1)=\int_0^1[1+\alpha(2x_1-1)(2x_2-1)]\mathrm{d}x_2=[x_2+\alpha(2x_1-1)(x_2^2-x_2)]_0^1=1$

$$f_{X_2}(x_2)=\int_0^1[1+\alpha(2x_1-1)(2x_2-1)]\mathrm{d}x_1=[x_1+\alpha(2x_2-1)(x_1^2-x_1)]_0^1=1$$

可见, 不同的联合密度函数得到了相同的边缘密度函数。

1.2 矩和特征函数

1.2.1 矩 — 期望和协方差阵

设随机向量 X 的密度函数为 $f(x)$, 则 X 的期望为:

$$EX=\begin{pmatrix}EX_1\\ \vdots\\ EX_p\end{pmatrix}=\int xf(x)\mathrm{d}x=\begin{pmatrix}\int x_1f(x)\mathrm{d}x\\ \vdots\\ \int x_pf(x)\mathrm{d}x\end{pmatrix}=\mu \tag{1.5}$$

相应地

$$E(\alpha X+\beta Y)=\alpha EX+\beta EY \tag{1.6}$$

如果一个 $q\times p$ 维的矩阵 $\mathcal{A}$ 是实数矩阵, 则有

$$E(\mathcal{A}X)=\mathcal{A}EX \tag{1.7}$$

在给定随机向量 X 和 Y 相互独立的条件下, 有

$$E(XY^{\mathrm{T}})=EXEY^{\mathrm{T}} \tag{1.8}$$

记矩阵

$$\mathrm{Var}(X)=\Sigma=E[(X-\mu)(X-\mu)^{\mathrm{T}}] \tag{1.9}$$

为协方差阵。如果向量 X 的均值为 μ, 协方差阵为 Σ, 则记为 $X\sim(\mu,\Sigma)$。

设 $X\sim(\mu,\Sigma_{XX}),Y\sim(\nu,\Sigma_{YY})$, 则 X 和 Y 的协方差阵为:

$$\Sigma_{XY}=\mathrm{Cov}(X,Y)=E[(X-\mu)(Y-\nu)^{\mathrm{T}}] \tag{1.10}$$

也可记为:

$$\mathrm{Cov}(X,Y)=E(XY^{\mathrm{T}})-\mu\nu^{\mathrm{T}}=E(XY^{\mathrm{T}})-EXEY^{\mathrm{T}} \tag{1.11}$$

我们通常称 $\mu=EX$ 是 X 的一阶矩, $E(XX^{\mathrm{T}})=\{E(X_iX_j)\}$ 是 X 的二阶矩。

1. 协方差阵 $\Sigma=\mathbf{Var(X)}$ 的性质

$$\Sigma=(\sigma_{X_iX_j}),\sigma_{X_iX_j}=\mathrm{Cov}(X_i,X_j),\sigma_{X_iX_i}=\mathrm{Var}(X_i) \tag{1.12}$$

$$\Sigma=E(XX^{\mathrm{T}})-\mu\mu^{\mathrm{T}} \tag{1.13}$$

$$\Sigma\geqslant 0 \tag{1.14}$$

2. 方差和协方差的性质

$$\mathrm{Var}(a^{\mathrm{T}}X)=a^{\mathrm{T}}\mathrm{Var}(X)a=\sum a_ia_j\sigma_{X_iX_j} \tag{1.15}$$

$$\mathrm{Var}(\mathcal{A}X+b)=\mathcal{A}\mathrm{Var}(X)\mathcal{A}^{\mathrm{T}} \tag{1.16}$$

$$\mathrm{Cov}(X+Y,Z)=\mathrm{Cov}(X,Z)+\mathrm{Cov}(Y,Z) \tag{1.17}$$

$$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Cov}(X,Y)+\mathrm{Cov}(Y,X)+\mathrm{Var}(Y) \tag{1.18}$$

$$\mathrm{Cov}(\mathcal{A}X,\mathcal{B}Y)=\mathcal{A}\mathrm{Cov}(X,Y)\mathcal{B}^{\mathrm{T}} \tag{1.19}$$

【例 1.3】 (续例 1.1) $f(x_1,x_2)=\dfrac{1}{3}x_1+\dfrac{2}{3}x_2,\ 0\leqslant x_1,x_2\leqslant 1$, 求均值向量和协方差矩阵。

解: 均值向量 $\mu=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}$, 其中

$$\begin{aligned}\mu_1&=\iint x_1f(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2=\int_0^1\int_0^1x_1\left(\frac{1}{3}x_1+\frac{2}{3}x_2\right)\mathrm{d}x_1\mathrm{d}x_2\\&=\int_0^1x_1\left(\frac{1}{3}x_1+\frac{1}{3}\right)\mathrm{d}x_1=\frac{1}{3}\left[\frac{x_1^3}{3}\right]_0^1+\frac{1}{3}\left[\frac{x_1^2}{2}\right]_0^1\\&=\frac{1}{9}+\frac{1}{6}=\frac{5}{18}\end{aligned}$$

$$\mu_2 = \int\int x_2 f(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 = \int_0^1\int_0^1 x_2\left(\frac{1}{3}x_1 + \frac{2}{3}x_2\right)\mathrm{d}x_1\mathrm{d}x_2$$
$$= \int_0^1 x_2\left(\frac{2}{3}x_2 + \frac{1}{6}\right)\mathrm{d}x_2 = \frac{2}{3}\left[\frac{x_2^3}{3}\right]_0^1 + \frac{1}{6}\left[\frac{x_2^2}{2}\right]_0^1$$
$$= \frac{2}{9} + \frac{1}{12} = \frac{11}{36}$$

协方差阵 $\Sigma = \begin{pmatrix} \sigma_{X_1X_1} & \sigma_{X_1X_2} \\ \sigma_{X_2X_1} & \sigma_{X_2X_2} \end{pmatrix}$, 其中, $\sigma_{X_1X_1} = EX_1^2 - \mu_1^2$, $\sigma_{X_2X_2} = EX_2^2 - \mu_2^2$, $\sigma_{X_1X_2} = E(X_1X_2) - \mu_1\mu_2$。

$$EX_1^2 = \int_0^1\int_0^1 x_1^2\left(\frac{1}{3}x_1 + \frac{2}{3}x_2\right)\mathrm{d}x_1\mathrm{d}x_2 = \frac{1}{3}\left[\frac{x_1^4}{4}\right]_0^1 + \frac{1}{3}\left[\frac{x_1^3}{3}\right]_0^1 = \frac{7}{36}$$

$$EX_2^2 = \int_0^1\int_0^1 x_2^2\left(\frac{1}{3}x_1 + \frac{2}{3}x_2\right)\mathrm{d}x_1\mathrm{d}x_2 = \frac{2}{3}\left[\frac{x_2^4}{4}\right]_0^1 + \frac{1}{6}\left[\frac{x_2^3}{3}\right]_0^1 = \frac{2}{9}$$

$$E(X_1X_2) = \int_0^1\int_0^1 x_1x_2\left(\frac{1}{3}x_1 + \frac{2}{3}x_2\right)\mathrm{d}x_1\mathrm{d}x_2 = \int_0^1\left(\frac{1}{9}x_2 + \frac{1}{3}x_2^2\right)\mathrm{d}x_2 = \frac{1}{6}$$

1.2.2 条件期望

$$E(X_2|x_1) = \int x_2 f(x_2|x_1)\mathrm{d}x_2 \tag{1.20}$$

$$E(X_1|x_2) = \int x_1 f(x_1|x_2)\mathrm{d}x_1 \tag{1.21}$$

$E(X_2|x_1)$ 表示给定 $X_1 = x_1$ 的条件下 X_2 的条件密度函数。同样, 在给定 $X_1 = x_1$ 的条件下, 我们可以定义 X_2 的离散程度如下:

$$\mathrm{Var}(X_2|X_1 = x_1) = E(X_2X_2^{\mathrm{T}}|X_1 = x_1) - E(X_2|X_1 = x_1)E(X_2^{\mathrm{T}}|X_1 = x_1) \tag{1.22}$$

有了条件协方差阵, 可以定义条件相关系数 (即偏相关系数) 如下:

$$\rho_{X_2X_3|X_1=x_1} = \frac{\mathrm{Cov}(X_2, X_3|X_1 = x_1)}{\sqrt{\mathrm{Var}(X_2|X_1 = x_1)\mathrm{Var}(X_3|X_1 = x_1)}} \tag{1.23}$$

条件期望的性质

因为 $E(X_2|X_1 = x_1)$ 是关于 x_1 的函数, 不妨定义随机变量 $h(X_1) = E(X_2|X_1)$, $g(X_1) = \mathrm{Var}(X_2|X_1)$, 则有以下性质:

$$E(X_2) = E\{E(X_2|X_1)\} \tag{1.24}$$

$$\mathrm{Var}(X_2) = E\{\mathrm{Var}(X_2|X_1)\} + \mathrm{Var}\{E(X_2|X_1)\} \tag{1.25}$$

【例 1.4】 随机变量 (X_1, X_2) 的概率密度为:

$$f(x_1, x_2) = 2\mathrm{e}^{-\frac{x_2}{x_1}},\ 0 < x_1 < 1, x_2 > 0$$

易算得

$$f(x_1) = 2x_1, 0 < x_1 < 1; E(X_1) = \frac{2}{3}; \mathrm{Var}(X_1) = \frac{1}{18}$$

$$f(x_2|x_1) = \frac{1}{x_1}\mathrm{e}^{-\frac{x_2}{x_1}}, x_2 > 0; E(X_2|X_1) = X_1; \mathrm{Var}(X_2|X_1) = X_1^2$$

$$E(X_2) = E[E(X_2|X_1)] = E(X_1) = \frac{2}{3}$$

$$\mathrm{Var}(X_2) = E[\mathrm{Var}(X_2|X_1)] + \mathrm{Var}[E(X_2|X_1)] = E(X_1^2) + \mathrm{Var}(X_1) = \frac{5}{9}$$

条件期望 $E(X_2|X_1)$ 可以看作 X_1 的函数 $h(X_1)$ (即 X_2 关于 X_1 的回归函数), 也可以理解为通过 X_1 的函数来表示 X_2 的条件近似值, 误差项表示为:

$$U = X_2 - E(X_2|X_1)$$

【定理 1.1】 设 $X_1 \in \mathbb{R}^k, X_2 \in \mathbb{R}^{p-k}, U = X_2 - E(X_2|X_1)$, 则有如下结论:

(1) $E(U) = 0$;

(2) $E(X_2|X_1)$ 是用 X_1 的函数 $h(X_1)$ 的形式表示的 X_2 的最佳近似, 最佳是指均方误差 (MSE) 最小, 其中

$$MSE(h) = E\{[X_2 - h(X_1)]^{\mathrm{T}}[X_2 - h(X_1)]\}$$

1.2.3　特征函数

设一随机向量 $X \in \mathbb{R}^p$ 的概率密度函数为 $f(x)$, 其特征函数定义为:

$$\varphi_X(t) = E(\mathrm{e}^{\mathrm{i}t^{\mathrm{T}}X}) = \int \mathrm{e}^{\mathrm{i}t^{\mathrm{T}}x} f(x)\mathrm{d}x,\ t \in \mathbb{R}^p$$

其中, i 是单位复数, $\mathrm{i}^2 = -1$, 特征函数具有如下性质:

$$\varphi_X(0) = 1, |\varphi_X(t)| \leqslant 1 \tag{1.26}$$

(1) 如果 φ 是绝对可积的, 积分 $\int_{-\infty}^{\infty} |\varphi(x)|\mathrm{d}x$ 存在并且是有限的, 则

$$f(x) = \frac{1}{(2\pi)^p} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}t^{\mathrm{T}}x} \varphi_X(t)\mathrm{d}t \tag{1.27}$$

(2) 如果 $X = (X_1, X_2, \cdots, X_p)^{\mathrm{T}}$, 则对于 $t = (t_1, t_2, \cdots, t_p)^{\mathrm{T}}$, 有

$$\varphi_{X_1}(t_1) = \varphi_X(t_1, 0, \cdots, 0), \cdots, \varphi_{X_p}(t_p) = \varphi_X(0, \cdots, 0, t_p) \tag{1.28}$$

(3) 如果 $X_1, X_2, \cdots, X_p$ 是相互独立的随机变量, 则对于 $t = (t_1, t_2, \cdots, t_p)^{\mathrm{T}}$, 有

$$\varphi_X(t) = \varphi_{X_1}(t_1)\varphi_{X_2}(t_2)\cdots\varphi_{X_p}(t_p) \tag{1.29}$$

(4) 如果 $X_1, X_2, \cdots, X_p$ 是相互独立的随机变量, 则对于 $t \in \mathbb{R}$, 有

$$\varphi_{X_1+X_2+\cdots+X_p}(t) = \varphi_{X_1}(t_1)\varphi_{X_2}(t_2)\cdots\varphi_{X_p}(t_p) \tag{1.30}$$

【例 1.5】 随机变量 X 服从标准正态分布, 则其概率密度函数为:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

特征函数为:

$$\begin{aligned}\varphi_X(t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}tx}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}(x^2-2\mathrm{i}tx+\mathrm{i}^2t^2)\right\}\exp\left\{\frac{1}{2}\mathrm{i}^2t^2\right\}\mathrm{d}x \\ &= \exp\left(-\frac{t^2}{2}\right)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(x-\mathrm{i}t)^2}{2}\right\}\mathrm{d}x \\ &= \exp\left(-\frac{t^2}{2}\right)\end{aligned}$$

式中, $\mathrm{i}^2=-1$, $\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(x-\mathrm{i}t)^2}{2}\right\}\mathrm{d}x=1$。

表 1-1 给出了几个常见分布的特征函数。

表 1-1　　常见分布的特征函数

分布	密度函数	特征函数
$U(a,b)$	$f(x)=I(x\in[a,b])/(b-a)$	$\varphi_X(t)=(\mathrm{e}^{\mathrm{i}bt}-\mathrm{e}^{\mathrm{i}at})/[(b-a)\mathrm{i}t]$
$N_1(\mu,\sigma^2)$	$f(x)=(2\pi\sigma^2)^{-1/2}\exp\{-(x-\mu)^2/2\sigma^2\}$	$\varphi_X(t)=\mathrm{e}^{\mathrm{i}\mu t-\sigma^2t^2/2}$
$\chi^2(n)$	$f(x)=I(x>0)x^{n/2-1}\mathrm{e}^{-x/2}/\{\Gamma(n/2)2^{n/2}\}$	$\varphi_X(t)=(1-2\mathrm{i}t)^{-n/2}$
$N_p(\mu,\Sigma)$	$f(x)=\lvert 2\pi\Sigma\rvert^{-1/2}\exp\{-(x-\mu)^{\mathrm{T}}\Sigma(x-\mu)/2\}$	$\varphi_X(t)=\mathrm{e}^{\mathrm{i}t^{\mathrm{T}}\mu-t^{\mathrm{T}}\Sigma t/2}$

【定理 1.2】(Cramer-Wold) $X\in\mathbb{R}^p$ 的分布完全取决于其所有的 $t^{\mathrm{T}}X(t\in\mathbb{R}^p)$ 的一维分布的集合。

也就是说, 给定 p 个线性组合 $\sum_{j=1}^{p}t_jX_j=t^{\mathrm{T}}X(t=(t_1,t_2,\cdots,t_p)^{\mathrm{T}})$ 的分布, 可以求得 $X\in\mathbb{R}^p$ 的分布。

设一维随机变量 X 的密度函数为 f, 存在 k 阶矩, $m_k=\int x^kf(x)\mathrm{d}x$。矩函数可以用来描述变量的分布特征, 譬如, 一维的正态分布函数完全由其一、二阶矩决定, $\mu=m_1,\sigma^2=m_2-m_1^2$。

1.3 多元随机向量变换

设随机变量 X 的密度函数为 $f_X(x)$, 那么 $Y=3X$ 的密度函数如何求解? 如果 $X=(X_1,X_2,X_3)^{\mathrm{T}}$, 那么 $Y=\begin{pmatrix}3X_1\\X_1-4X_2\\X_3\end{pmatrix}$ 的密度函数又如何求解呢?

考虑一种特殊的情形:

$$X=u(Y) \tag{1.31}$$

其中, $u:\mathbb{R}^p\to\mathbb{R}^p$ 是一对一的变换。定义 u 的雅可比矩阵为:

$$\mathcal{J}=\left(\frac{\partial x_i}{\partial y_i}\right)=\left(\frac{\partial u_i(y)}{\partial y_i}\right)$$

记雅可比行列式的绝对值为 $\mathrm{abs}(|\mathcal{J}|)$, 则 Y 的密度函数为:

$$f_Y(y) = \mathrm{abs}(|\mathcal{J}|) \cdot f_X\{u(y)\} \tag{1.32}$$

由此, 前面提出的问题得到解答, 设

$$(x_1, x_2, \cdots, x_p)^{\mathrm{T}} = u(y_1, y_2, \cdots, y_p) = \frac{1}{3}(y_1, y_2, \cdots, y_p)^{\mathrm{T}}$$

有

$$\mathcal{J} = \begin{pmatrix} 1/3 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/3 \end{pmatrix}$$

则 $\mathrm{abs}(|\mathcal{J}|) = \left(\frac{1}{3}\right)^p$, 故 Y 的密度函数 $f_Y(y) = \frac{1}{3^p} f_X\left(\frac{y}{3}\right)$。

上述情形是 $Y = \mathcal{A}X + b$ 的一个特例。其中, 矩阵 $\mathcal{A}$ 是非奇异阵。由 $Y = \mathcal{A}X + b$ 逆变换得 $X = \mathcal{A}^{-1}(Y - b)$, 则有 $\mathcal{J} = \mathcal{A}^{-1}$, 故

$$f_Y(y) = \mathrm{abs}(|\mathcal{A}|^{-1}) f_X\{\mathcal{A}^{-1}(y - b)\} \tag{1.33}$$

【例 1.6】 随机变量 $X = (X_1, X_2)$ 的概率密度函数为 $f_X(x) = f_X(x_1, x_2)$, 有 $\mathcal{A} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, $b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $Y = \mathcal{A}X + b = \begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix}$。求随机变量 Y 的密度函数。

解:

$$|\mathcal{A}| = -2,\ \mathrm{abs}(|\mathcal{A}|^{-1}) = \frac{1}{2},\ \mathcal{A}^{-1} = -\frac{1}{2}\begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix}$$

故

$$\begin{aligned} f_Y(y) &= \mathrm{abs}(|\mathcal{A}|^{-1}) \cdot f_X(\mathcal{A}^{-1} y) \\ &= \frac{1}{2} f_X\left\{\frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right\} \\ &= \frac{1}{2} f_X\left\{\frac{1}{2}(y_1 + y_2), \frac{1}{2}(y_1 - y_2)\right\} \end{aligned}$$

1.4　多元正态分布

设多元正态分布的均值向量为 μ, 协方差矩阵为 Σ, 则密度函数为:

$$f(x) = |2\pi\Sigma|^{-1/2} \exp\{-\frac{1}{2}(x - \mu)^{\mathrm{T}} \Sigma^{-1}(x - \mu)\} \tag{1.34}$$

记作 $X \sim N_p(\mu, \Sigma)$。

多元正态分布 $X \sim N_p(\mu, \Sigma)$ 与多元标准正态分布 $N_p(0, \mathcal{I}_p)$ 有怎样的关系呢? 其实, 通过前面部分介绍的线性变换即可转换得到, 定理如下。

【定理 1.3】 $X \sim N_p(\mu, \Sigma), Y = \Sigma^{-1/2}(X - \mu)$, 即马氏变换, 则 $Y \sim N_p(0, \mathcal{I}_p)$。

也就是说, 元素 $Y_j \in \mathbb{R}$ 是相互独立且服从一维标准正态分布 $N(0, 1)$ 的变量。

证明: 记 $(X - \mu)^{\mathrm{T}} \Sigma^{-1}(X - \mu) = Y^{\mathrm{T}} Y$, 运用式 (1.33), 给定 $\mathcal{J} = \Sigma^{1/2}$, 则有

$$f_Y(y) = (2\pi)^{-p/2} \exp\left(-\frac{1}{2} y^{\mathrm{T}} y\right) \tag{1.35}$$

根据式 (1.34), 上式即为 $N_p(0, \mathcal{I}_p)$ 的密度函数。

不难发现, 上述的马氏变换实际上产生的是随机变量 $Y = (Y_1, Y_2, \cdots, Y_p)^{\mathrm{T}}$, 它由相互独立、服从一维标准正态分布的元素 $Y_j \sim N_1(0,1)$ 所组成, 因为

$$\begin{aligned} f_Y(y) &= \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} y^{\mathrm{T}} y\right) \\ &= \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} y_j^2\right) \\ &= \prod_{j=1}^{p} f_{Y_j}(y_j) \end{aligned}$$

$f_{Y_j}(y)$ 是标准正态密度函数 $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} y^2\right)$。由此可以得到, $E(Y) = 0, \mathrm{Var}(Y) = \mathcal{I}_p$。

反过来, 由一系列服从 $N_p(0, \mathcal{I}_p)$ 分布的变量如何产生服从 $N_p(\mu, \Sigma)$ 分布的变量呢? 我们用如下的逆线性变换

$$X = \Sigma^{1/2} Y + \mu \tag{1.36}$$

由式 (1.36) 和 (1.16) 可以验得 $E(X) = \mu, \mathrm{Var}(X) = \Sigma$。

【定理 1.4】 设 $X \sim N_p(\mu, \Sigma), \mathcal{A}(p \times p), c \in \mathbb{R}^p$, 其中, $\mathcal{A}$ 是非奇异阵, 则 $Y = \mathcal{A}X + c$ 服从 p 元正态分布, 即

$$Y \sim N_p(\mathcal{A}\mu + c, \mathcal{A}\Sigma\mathcal{A}^{\mathrm{T}}) \tag{1.37}$$

【定理 1.5】 如果 $X \sim N_p(\mu, \Sigma)$, 那么变量 $U = (X - \mu)^{\mathrm{T}} \Sigma^{-1} (X - \mu)$ 服从自由度为 p 的卡方分布, 即 $U \sim \chi_p^2$。

【定理 1.6】 多元正态分布 $N_p(\mu, \Sigma)$ 的特征函数为:

$$\varphi_X(t) = \exp(\mathrm{i} t^{\mathrm{T}} \mu - \frac{1}{2} t^{\mathrm{T}} \Sigma t) \tag{1.38}$$

证明:

$$\begin{aligned} f(x) &= \frac{1}{(2\pi)^p} \int \exp(-\mathrm{i} t^{\mathrm{T}} x + \mathrm{i} t^{\mathrm{T}} \mu - \frac{1}{2} t^{\mathrm{T}} \Sigma t) \mathrm{d}t \\ &= \frac{1}{|2\pi\Sigma^{-1}|^{1/2} |2\pi\Sigma|^{1/2}} \\ &\quad \cdot \int \exp\left[-\frac{1}{2}\{t^{\mathrm{T}} \Sigma t + 2\mathrm{i} t^{\mathrm{T}} (x - \mu) - (x - \mu)^{\mathrm{T}} \Sigma^{-1} (x - \mu)\}\right] \\ &\quad \cdot \exp\left[-\frac{1}{2}\{(x - \mu)^{\mathrm{T}} \Sigma^{-1} (x - \mu)\}\right] \mathrm{d}t \\ &= \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left[-\frac{1}{2}\{(x - \mu)^{\mathrm{T}} \Sigma^{-1} (x - \mu)\}\right] \end{aligned}$$

因为

$$\begin{aligned}
&\int \frac{1}{|2\pi\Sigma^{-1}|^{1/2}} \exp\left[-\frac{1}{2}\{t^{\mathrm{T}}\Sigma t + 2\mathrm{i}t^{\mathrm{T}}(x-\mu) - (x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\}\right] \mathrm{d}t \\
&= \int \frac{1}{|2\pi\Sigma^{-1}|^{1/2}} \exp\left[-\frac{1}{2}\{[t+\mathrm{i}\Sigma^{-1}(x-\mu)]^{\mathrm{T}}\Sigma[t+\mathrm{i}\Sigma^{-1}(x-\mu)]\}\right] \mathrm{d}t \\
&= 1
\end{aligned}$$

同理, 如果 $Y \sim N_p(0, \mathcal{I}_p)$, 那么

$$\begin{aligned}
\varphi_Y(t) &= \exp\left(-\frac{1}{2}t^{\mathrm{T}}\mathcal{I}_p t\right) = \exp\left(-\frac{1}{2}\sum_{i=1}^{p} t_i^2\right) \\
&= \varphi_{Y_1}(t_1)\varphi_{Y_2}(t_2)\cdots\varphi_{Y_p}(t_p)
\end{aligned}$$

这与式 (1.29) 是一致的。

1. 奇异的正态分布

假设协方差阵 Σ 的秩 k 小于 X 的维数 p, 则 X 的奇异密度函数为:

$$f(x) = \frac{(2\pi)^{-k/2}}{(\lambda_1\lambda_2\cdots\lambda_k)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-}(x-\mu)\right\} \tag{1.39}$$

其中

(1) x 位于超平面 $\mathcal{N}^{\mathrm{T}}(x-\mu)=0$ 上, 有 $\mathcal{N}(p\times(p-k)) : \mathcal{N}^{\mathrm{T}}\Sigma = 0, \mathcal{N}^{\mathrm{T}}\mathcal{N} = \mathcal{I}_k$;

(2) Σ^{-} 是 Σ 的广义逆, $\lambda_1, \lambda_2, \cdots, \lambda_k$ 是矩阵 Σ 的非零特征根。

如果 $Y \sim N_k(0, \Lambda_1)$, $\Lambda_1 = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_k)$, 那么存在正交矩阵 $\mathcal{B}(p\times k)$, $\mathcal{B}^{\mathrm{T}}\mathcal{B} = \mathcal{I}_k$, 使得 $X = \mathcal{B}Y + \mu$ 有如式 (1.39) 所示的奇异密度函数。

2. 高斯连接函数

高斯连接函数或称正态连接函数, 即

$$C_\rho(u,v) = \int_{-\infty}^{\Phi_1^{-1}(u)} \int_{-\infty}^{\Phi_2^{-1}(v)} f_\rho(x_1, x_2)\mathrm{d}x_2\mathrm{d}x_1 \tag{1.40}$$

式中, f_ρ 是相关系数为 ρ 的二元正态密度函数; Φ_1, Φ_2 是相对应的一维标准正态分布的边际累积分布函数。

在相关系数为零的情形下, $\rho = 0$, 高斯连接函数变为:

$$\begin{aligned}
C_0(u,v) &= \int_{-\infty}^{\Phi_1^{-1}(u)} f_{X_1}(x_1)\mathrm{d}x_1 \int_{-\infty}^{\Phi_2^{-1}(v)} f_{X_2}(x_2)\mathrm{d}x_2 \\
&= uv \\
&= \Pi(u,v)
\end{aligned}$$

1.5　样本分布和极限定理

在多元统计中, 对多元随机向量 X 进行观测并得到一组样本 $\{x_i\}_{i=1}^n$。在随机抽样的前提下, 这些观测值被认为是独立同分布的的随机变量 $X_1, X_2, \cdots, X_n$ 的一次实现, 其中 X_i 是 p 维随机向量, 表示对总体随机向量 X 的一次重复观测。请注意这里的一些记号的含义: X_i 不是指 X 的第 i 个分量, 而是指 p 维随机向量的第 i 次观测。正是 X_i 提供了第 i 个样本观测值 x_i。

对一个给定的随机样本 $X_1, X_2, \cdots, X_n$, 统计推断的目的是分析总体变量 X 的性质。具体来说主要是分析分布的一些特征, 比如均值、协方差阵等。

推断也可以依据样本 $X_1, X_2, \cdots, X_n$ 的函数展开, 比如统计量样本均值 $\bar{x}$, 样本协方差阵