研究生数学教材

高等工程数学讲义

南京理工大学高等工程数学编写组

前　言

“双一流”建设是新时代我国高等教育的重大改革举措，是高等学校进入新时期迈入新高度的重要机遇。随着研究生招生数量的提高，势必带来研究生教育的各项改革措施。针对目前研究生培养工作的特点（有全日制学术学位，全日制专业学位，非全日制学位），为迎合国家“供给侧”改革的战略形势，进行研究生数学课程教学改革势在必行。

在理工科专业（非数学专业）研究生中统一开设必修的基础数学课程“高等工程数学”是研究生基础教学的重要改革措施。教学内容在注重基本数学理论基础上，以基本数学方法为重点。通过该课程的教学，使研究生掌握矩阵分析，优化方法，方程求解，数据拟合，统计分析等方面的知识，能够熟练应用这些数学方法解决学科研究中面临的相关问题。 本书适合为工科高校研究生教材， 也可为理科或管理等学科的研究生、教师、有关研究者作教学与研究参考书。

全书共八章，第一章由范金华编写，第二章由李宝成编写，第三章由王海侠编写，第四章由饶玲编写，第五章1-4节由严涛编写，5-7节由朱元国编写，第六章由刘红毅编写，第七章由张军编写，第八章1-3节由陆中胜编写，4-5节由侯传志编写。全书由朱元国统稿。由于我们水平有限，书中难免有疏漏之处，敬请广大专家及读者批评指正。

高等工程数学编写组

2018年7月于南京理工大学

常用符号

$\boldsymbol{x}$, $\boldsymbol{y}$	向量
A, B	矩阵
$\boldsymbol{C}$	复数域
$\boldsymbol{R}$	实数域
$\boldsymbol{C}^n$	n 维复向量集合
$\boldsymbol{R}^n$	n 维实向量集合
$\boldsymbol{C}^{m\times n}$	m 行 n 列复矩阵组成的集合
$\boldsymbol{C}_r^{m\times n}$	m 行 n 列秩为 r 的复矩阵组成的集合
$\boldsymbol{R}^{m\times n}$	m 行 n 列实矩阵组成的集合
$\overline{A}$	矩阵 A 的共轭
A^T	矩阵 A 的转置
A^H	矩阵 A 的共轭转置
0	数零，零向量，零矩阵
I	单位矩阵
J	方阵的 Jordan 标准型
rank$A$	方阵 A 的秩
det$A$	方阵 A 的行列式
tr$A$	方阵 A 的迹（A 的主对角线上元素之和）
cond(A)	方阵 A 的条件数
$\rho(A)$	方阵 A 的谱半径
$\|A\|$	矩阵 A 的范数
diag$(a_1, a_2, \cdots, a_n)$	以 $a_1, a_2, \cdots, a_n$ 为对角元素的 $n\times n$ 对角矩阵

目　录

第 一 章　距离与范数

距离空间是实直线$\boldsymbol{R}$的推广，它在一般分析中的地位犹如高等数学中的实直线$\boldsymbol{R}$。将实直线$\boldsymbol{R}$推广到一般的距离空间，有利于对于一般问题的统一处理。

本章主要首先介绍距离空间的概念、性质，在此基础上将实直线$\boldsymbol{R}$上一些性质和概念推广到距离空间。 在线性空间中，我们引入一种类似于距离的概念——范数。以范数为基础，我们将介绍带有范数的线性空间（赋范线性空间）的性质以及一些特殊的赋范线性空间。 利用向量范数与矩阵范数的概念及其性质，介绍矩阵的谱半径，条件数，以及它们在矩阵级数， 求解线性方程组的误差估计中的应用。

1.1　距离空间、极限与连续性

Euclid空间$\boldsymbol{R}^n$是由所有形如$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^T$ (下面简写为$\boldsymbol{x}=(x_i)$) 的n维向量构成，它是三维空间的一种自然推广。 类似于三维空间中$\boldsymbol{R}^n$中Euclid距离的定义，$\boldsymbol{R}^n$中任何两点$\boldsymbol{x}=(x_i),\boldsymbol{y}=(y_i)$的距离定义为

$$d(\boldsymbol{x},\boldsymbol{y})=(\sum_{i=1}^{n}(x_i-y_i)^2)^{1/2}.$$

在实际中，除了空间的距离，其他的距离也很多，例如大家的学识差异、收入差距等都是一种距离。为了研究一类事物或集合之间的差异， 或进一步研究集合上的"函数"性质，我们需要在该集合元素之间引入距离的概念。

定义 1.1 设X是一非空集合，对X中任何两元素$\boldsymbol{x},\boldsymbol{y}$, 按某一法则对应唯一的实数$d(\boldsymbol{x},\boldsymbol{y})$，而且满足下列三条性质:

(1) （非负性）$d(\boldsymbol{x},\boldsymbol{y})\geq 0$, $d(x,y)=0$当且仅当$x=y$;

(2) （对称性）$d(\boldsymbol{x},\boldsymbol{y})=d(\boldsymbol{y},\boldsymbol{x})$;;

(3) （三角不等式）$d(\boldsymbol{x},\boldsymbol{y})\leq d(\boldsymbol{x},\boldsymbol{z})+d(\boldsymbol{y},\boldsymbol{z})$.对所有$\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in X$, 成立;

则称$d(\boldsymbol{x},\boldsymbol{y})$为$\boldsymbol{x},\boldsymbol{y}$之间的距离， 并称$X$是以$d$为距离的距离空间，记作$(X,d)$。

通常，在距离给出并不引起混淆下，我们简记(X,d)为X。 X中的元素称为X中的点。下面举一些距离空间的例子，验证所引入的距离满足定义 1.1中的三条性质，留作习题。

例 1.1 设X是任意非空集合，对X中任何两个元素$\boldsymbol{x},\boldsymbol{y}$， 令

$$d(\boldsymbol{x},\boldsymbol{y})=\begin{cases}1, & \boldsymbol{x}\neq\boldsymbol{y},\\ 0, & \boldsymbol{x}=\boldsymbol{y}.\end{cases}$$

则(X,d)成为一个距离空间，称空间(X,d)为离散距离空间。这种距离是粗糙的，它只能将X中不同的元素区分开，但是不能衡量不同元素之间差异的大小。

例 1.2 设 $X=\boldsymbol{R}^n$，对 $\boldsymbol{x}=(x_i),\boldsymbol{y}=(y_i)\in\boldsymbol{R}^n$，分别定义

$$d_1(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{n}|x_i-y_i|,\ d_2(\boldsymbol{x},\boldsymbol{y})=\{\sum_{i=1}^{n}(x_i-y_i)^2\}^{1/2},\ d_\infty(\boldsymbol{x},\boldsymbol{y})=\max_i|x_i-y_i|.$$

则$(\boldsymbol{R}^n,d_1),(\boldsymbol{R}^n,d_2),(\boldsymbol{R}^n,d_\infty)$都是距离空间。

例 1.3 对$1\le p<+\infty$,记

$$l^p=\{\boldsymbol{x}:\boldsymbol{x}=(x_1,x_2,\cdots,x_i,\cdots)^T,\sum_{i=1}^{n}|x_i|^p<+\infty\},$$

对 $\boldsymbol{x}=(x_1,x_2,\cdots,x_i,\cdots)^T$，$\boldsymbol{y}=(y_1,y_2,\cdots,y_i,\cdots)^T\in l^p$，定义

$$d_p(\boldsymbol{x},\boldsymbol{y})=\left\{\sum_{i=1}^{+\infty}|x_i-y_i|^p\right\}^{1/p},$$

则(l^p,d_p)是距离空间。

例 1.4 记

$$l^\infty:=\{\boldsymbol{x}:\boldsymbol{x}=(x_1,x_2,\cdots,x_i,\cdots)^T,\sup_i|x_i|<+\infty\},$$

对$\boldsymbol{x}=(x_1,x_2,\cdots,x_i,\cdots)^T$，$\boldsymbol{y}=(y_1,y_2,\cdots,y_i,\cdots)^T\in l^\infty$，定义

$$d_\infty(\boldsymbol{x},\boldsymbol{y}):=\sup_i|x_i-y_i|,$$

则(l^∞,d_∞)是距离空间。

类似于实数集$\boldsymbol{R}$上数列的极限，下面介绍一般距离空间上极限以及其相关性质。

定义 1.2 设$\{\boldsymbol{x}_n\}_{n=1}^{\infty}$是距离空间$(X,d)$中点列，$\boldsymbol{x}_0$是$X$中确定的点。若对任何给定的正数$\varepsilon$，总存在自然数$N$，当$n>N$时成立$d(\boldsymbol{x}_n,\boldsymbol{x}_0)<\varepsilon$，则称点列$\{\boldsymbol{x}_n\}_{n=1}^{\infty}$收敛于$\boldsymbol{x}_0$，或者说$\boldsymbol{x}_0$是$\{\boldsymbol{x}_n\}_{n=1}^{\infty}$的极限，记作

$$\lim_{n\to\infty}\boldsymbol{x}_n=\boldsymbol{x}_0,$$

或者 $\boldsymbol{x}_n\to\boldsymbol{x}_0\ (n\to+\infty)$。

利用距离三角不等式性质，不难发现极限具有下面性质，证明留给读者。

性质 1.1 (1) 若距离空间(X,d)中点列$\{\boldsymbol{x}_n\}$和$\{\boldsymbol{y}_n\}$满足

$$\lim_{n\to\infty}\boldsymbol{x}_n=\boldsymbol{x}_0,\lim_{n\to\infty}\boldsymbol{y}_n=\boldsymbol{y}_0,$$

则

$$\lim_{n\to\infty} d(\boldsymbol{x}_n, \boldsymbol{y}_n) = d(\boldsymbol{x}_0, \boldsymbol{y}_0);$$

(2) 若距离空间(X,d)中点列$\{x_n\}$收敛，则极限是唯一的。

例 1.5 $(\boldsymbol{R}^n, d_2)$中点列$\{\boldsymbol{x}^{(m)}\}_{m=1}^{+\infty} = \{(x_1^{(m)}, x_2^{(m)}, \cdots, x_n^{(m)})\}_{m=1}^{+\infty}$收敛于

$$\boldsymbol{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \cdots, x_n^{(0)})$$

的充要条件是对任何$1 \le i \le n$ 有$x_i^{(m)} \to x_i^{(0)}(m \to \infty)$, 即按坐标分量收敛。

证明: "$\Longrightarrow$" 对任何$1 \le i \le n$, 由于

$$|x_i^{(m)} - x_i^{(0)}| \le \left\{\sum_{i=1}^{n}(x_i^{(m)} - x_i^{(0)})^2\right\}^{1/2} = d_2(x^{(m)}, x^{(0)}),$$

所以当$d_2(\boldsymbol{x}^{(m)}, \boldsymbol{x}^{(0)}) \to 0(m \to +\infty)$时，一定有$d_2(x_i^{(m)}, x_i^{(0)}) \to 0\ (m \to +\infty)$，即$x_i^{(m)} \to x_i^{(0)}\ (m \to +\infty)$。

"$\Longleftarrow$" 由于

$$d_2(\boldsymbol{x}^{(m)}, \boldsymbol{x}^{(0)}) = \left\{\sum_{i=1}^{n}(x_i^{(m)} - x_i^{(0)})^2\right\}^{1/2} \le \sum_{i=1}^{n}|x_i^{(m)} - x_i^{(0)}|,$$

所以若对任何 $1 \le i \le n$有$x_i^{(m)} \to x_i^{(0)}\ (m \to \infty)$， 则$d_2(\boldsymbol{x}^{(m)}, \boldsymbol{x}^{(0)}) \to 0\ (m \to +\infty)$。证毕。

例 1.6 距离空间(l^p, d_p)中点列$\{\boldsymbol{x}^{(m)}\}_{m=1}^{+\infty} = \{(x_1^{(m)}, x_2^{(m)}, \cdots, x_i^{(m)}, \cdots)\}_{m=1}^{+\infty}$收敛于点$\boldsymbol{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \cdots, x_i^{(0)}, \cdots)$，类似于上例的讨论可以得到对任何$i \in \mathbf{N}$， 都有$x_i^{(m)} \to x_i^{(0)}\ (m \to +\infty)$。即在距离空间$(l^p, d_p)$中按距离收敛蕴含着按坐标分量收敛。但反之不成立。对$\forall m \in \mathbf{N}$, 分别令

$$\boldsymbol{x}^{(m)} = (\underbrace{(1/m)^{1/p}, \cdots, (1/m)^{1/p}}_{m个}, 0, 0, \cdots), \boldsymbol{x}^{(0)} = (0, 0, \cdots),$$

则不难验证，对$m \in \mathbf{N}$, $\boldsymbol{x}^{(m)}, \boldsymbol{x}^{(0)} \in l^p$且 $x_i^{(m)} \to x_i^{(0)} = 0\ (m \to +\infty)$。但是

$$d_p(\boldsymbol{x}^{(m)}, \boldsymbol{x}^{(0)}) = \left\{\sum_{i=1}^{+\infty}|x_i^{(m)}|^p\right\}^{1/p} = \left\{\sum_{i=1}^{m}\frac{1}{m}\right\}^{1/p} = 1,$$

故在距离空间(l^p, d_p)中按坐标分量收敛不能保证距离收敛。

下面介绍实数区间上的函数概念推广到一般距离空间上，并介绍连续性概念。

定义 1.3 设(X, d_X)和(Y, d_Y)是两个距离空间，T是X到Y的的一个映射，$\boldsymbol{x}_0 \in X$，若对$\varepsilon > 0$，存在$\delta > 0$，当 $d_X(\boldsymbol{x}, \boldsymbol{x}_0) < \delta$时，有$d_Y(T\boldsymbol{x}, T\boldsymbol{x}_0) < \varepsilon$，则称$T$在$x_0$连续；若$T$在$X$上的每一点都连续，则称$T$为$X$上的连续映射。

例 1.7 设(X,d)是距离空间，x_0是X上一定点，对$x \in X$，定义映射$f(x)=d(x,x_0)$，则通过定义不难验证$f(x)$是X到$\boldsymbol{R}$的连续映射（函数）。

连续可以利用极限进行描述，下面的定理反映了连续和极限之间的联系，证明留做习题。

定理 1.1 设X,Y是两个距离空间，$T:X\to Y$，$x_0\in X$，则下列两个命题是等价的。

(1) T在x_0连续;

(2) 对X中任意点列$\{x_n\}$，若$x_n\to x_0\ (n\to\infty)$，则$Tx_n\to Tx_0\ (n\to\infty)$。

1.2 距离空间的可分性、完备性与紧性

在实数空间$\boldsymbol{R}$中，任何实数都可以用有理数去逼近，即有理数在$\boldsymbol{R}$中稠密；$\boldsymbol{R}$中任何Cauchy序列的极限是实数，即$\boldsymbol{R}$是完备的；此外$\boldsymbol{R}$中，任何有界数列必有收敛子列，即$\boldsymbol{R}$是紧的。实数空间$\boldsymbol{R}$的这三个良好性质，对研究$\boldsymbol{R}$上的函数性质起到了关键的作用。本节，我们将在一般距离空间中引入类似概念并研究其性质。

1.2.1 可数集

为了叙述距离空间的一般性质和概念，我们在本段对集合元素数量“多少”进行数学严格的描述。在集合元素有限时，描述集合元素数量是自然的；但是当集合数量无限时，如何定量描述集合元素数量，以及比较两个无穷多元素集合之间的“多少”关系是无穷集合理论中的重点。19世纪70年代，德国数学家Cantor开创了无穷集合理论，很好地解决了无穷集合元素个数的衡量问题。在学习本段之后，你会惊讶地发现所有自然数和所有有理数居然“一样多”。

定义 1.4 设A,B是两个集合，如果存在A到B的一一映射(既是单射也是满射)，则称A与B对等，记为$A\sim B$。

显然两个有限集（元素个数有限）对等的充要条件是它们的元素个数相同，从而有限集不可能与它的真子集对等，但是无限集（元素个数无限）却完全不一样。

例 1.8 令$A=\{$全体偶数$\}$，$\boldsymbol{Z}$表示整数集，取$T(x)=2x$，令

$$T:A\longrightarrow \boldsymbol{Z},\quad Tx=2x,$$

则T是A到$\boldsymbol{Z}$的一一映射，从而$A\sim\boldsymbol{Z}$。

更进一步的，我们可以得到无限集下面的性质定理。

定理 1.2 任何无限集必与它的某真子集对等。

证明：设A为一个无限集，取出一个元素$a_1 \in A$，由于A为无限集，则$A - \{a_1\} \neq \emptyset$。同样办法，在$A$中可以依次取出一列互异 元素$a_1, a_2, \cdots, a_n, \cdots$，记集合$B = A - \{a_1, a_2, \cdots, a_n, \cdots\}$, $C = A - \{a_2, \cdots, a_n, \cdots\}$。 定义映射$T : A \to C$如下

$$Ta = \begin{cases} a, & a \in B, \\ a_{k+1}, & a = a_k \in B \end{cases}$$

则T是A到C的一一映射，从而$A \sim C$。 证毕。

定义 1.5 和自然数集$\mathbf{N}$对等的集合称为可数集， 不能和$\mathbf{N}$对等的无限集称为不可数集。

由定义可知，若A是可数集，则A全体元素可表成无穷序列的形式，即

$$A = \{x_1, x_2, \cdots, x_n, \cdots\}.$$

如整集$Z = \{0, 1, -1, 2, -2, \cdots, -n, n, \cdots\}$是可数集。下面的定理说明可数集是“最小”的无限集。

定理 1.3 任意无限集必含有一个可数子集。

证明：可数子集的构造方法和定理1.2证明构造类似，详细过程在此省略。

定理 1.4 有限或可数个可数集的并仍然是可数集；可数个有限集的并（若为无限集）也是可数集。

证明：不失一般性，我们只证明可数个可数集的并情形，其他情形类似。假设$\{A_n : n \in \mathbf{N}\}$为一列可数集，并假定 $A_n \cap A_m = \emptyset (n \neq m)$, 于是有

$$\begin{aligned} A_1 &= \{a_{11}, a_{12}, a_{13}, a_{14}, \cdots\} \\ &\quad \nearrow \ \nearrow \ \nearrow \\ A_2 &= \{a_{21}, a_{22}, a_{23}, a_{24}, \cdots\} \\ &\quad \nearrow \ \nearrow \ \nearrow \\ A_3 &= \{a_{31}, a_{32}, a_{33}, a_{34}, \cdots\} \\ &\quad \cdots\cdots\cdots \end{aligned}$$

依照对角线原则（箭头所示）可把$\cup_{n=1}^{\infty} A_n$中全部元素排成列如下形式

$$a_{11}, a_{21}, a_{12}, a_{31}, a_{22}, a_{13}, \cdots$$

所以$\cup_{n=1}^{\infty} A_n$是可数集。 证毕。

由这个定理可以得到下面一个重要的事实。

定理 1.5 有理数集$\mathbf{Q}$是可数集。

证明：由于每个有理数可以表示成既约分数p/q，其中p,q为整数，且规定$q \in \mathbf{N}$。对每个$q \in \mathbf{N}$, 记$A_q = \{p/q : p \in Z\}$。 则$\mathbf{Q} = \cup_{q=1}^{\infty} A_q$为可数个可数集的并，故$\mathbf{Q}$为可数集。 证毕。

并不是所有的无限集都是可数集，下面举一个不可数集的例子

例 1.9 点集$(0,1) = \{x \in \boldsymbol{R} : 0 < x < 1\}$ 是不可数集。

证明：假设$(0,1)$是可数集，则$(0,1)$中全体数可以排成一列$x_1, x_2, \cdots$。将每个x_n用十进制小数表示，则有

$$x_1 = 0.t_{11}t_{12}t_{13}\cdots$$

$$x_2 = 0.t_{21}t_{22}t_{23}\cdots$$

$$\cdots\cdots\cdots$$

其中所有的t_{ij}都是从$0,1,2,\cdots 9$中取值，并且对每个i，数列$\{t_{ij} : j = 1,2,\cdots\}$应有无限个不为0，即小数$0.32000\cdots$应改写为$0.31999\cdots$。

作十进制小数$x = 0.b_1b_2\cdots$，当$t_{ii} = 1$时，令$b_i = 2$；而当$t_{ii} \neq 1$时，令$b_i = 1$。显然 $x \in (0,1)$，且由于对每个n，$b_n \neq t_{nn}$，故$x \neq x_n$，这与假设矛盾，从而$(0,1)$是不可数集。 证毕。

1.2.2 距离空间的可分性

定义 1.6 设(X,d)是一距离空间，A,B都是X的子集。若对$\forall \boldsymbol{y} \in B$ 以及$\forall \varepsilon > 0$，都存在 $\boldsymbol{x} \in A$ 使得 $d(\boldsymbol{x},\boldsymbol{y}) < \varepsilon$， 则称$A$在$B$中稠密。

我们考察距离空间X是否具有某种性质时，往往先在它的稠密子集上考察，然后通过极限过程得到X相应的性质，这是引入稠密这个概念的原因。 利用稠密，我们引入距离空间另外一个概念。

定义 1.7 如果度量空间X存在一个可数的稠密子集，那么称X是可分的。

由于有理数集$\mathbf{Q}$是可数集，并且$\mathbf{Q}$在实数集$\boldsymbol{R}$中是稠密的，从而可以得到下面性质。

例 1.10 $\boldsymbol{R}^n$是可分的。

并不是所有的距离空间都是可分的，下面给出一个不可分的空间。

例 1.11 (l^{∞}, d_{∞})是不可分的。

证明：假设l^∞是可分的，则存在可数稠密子集，记为A。在l^∞中构造以下一个子集B，

$$B=\{\boldsymbol{x}=(x_1,x_2,\cdots,x_n,\cdots)^T:x_n\in\mathbf{N},0\le x_n\le 9\}.$$

集合B与区间$[0,1]$可以建立如下一一对应，

$$T:B\leftrightarrow[0,1]:T\boldsymbol{x}=T\{(x_1,x_2,\cdots,x_n,\cdots)\}=0.x_1x_2\cdots.$$

由于$[0,1]$是不可数的，从而B是不可数的。

由于A在l^∞中稠密，从而$\bigcup\limits_{a\in A}B(a,\frac{1}{4})=l^\infty$，其中$B(a,\frac{1}{4})$表示以$a$为中心，$\frac{1}{4}$为半径的球。由于$A$可数，$B$不可数，所以至少存在B中两个不同的点$\boldsymbol{x},\boldsymbol{y}$落在某一个球$B(a_0,\frac{1}{4})$内。从$B$的元素构造可得$d_\infty(\boldsymbol{x},\boldsymbol{y})\ge 1$，但是

$$d_\infty(\boldsymbol{x},\boldsymbol{y})\le d_\infty(\boldsymbol{x},a_0)+d_\infty(\boldsymbol{y},a_0)<\frac{1}{4}+\frac{1}{4}<\frac{1}{2},$$

矛盾，所以l^∞不可数。 证毕。

例 1.12 *X是一个不可数集，在X上定义离散距离d (见例1.1)，类似上例证明方法，可以证明 (X,d) 不可分。*

1.2.3 距离空间的完备性

一个实数序列是否收敛等价于该数列是否是Cauchy序列，但是该结论在一般的度量空间中是不成立的，其主要原因在于实数空间具有一定良好的性质——完备性。为此，我们需要将完备性引入到距离空间中。

定义 1.8 *设$\{\boldsymbol{x}_n\}_{n=1}^\infty$是距离空间$(X,d)$中的点列，若对任何$\varepsilon>0$，都存在$N\in\mathbf{N}$，使得当$n,m>N$时有$d(\boldsymbol{x}_m,\boldsymbol{x}_n)<\varepsilon$，则称$\{\boldsymbol{x}_n\}_{n=1}^\infty$为Cauchy列。如果$X$中任何Cauchy列都在$X$中收敛，则称 X是完备的。*

下面分别举几个完备和不完备的例子。

例 1.13 *$(\boldsymbol{R}^n,d_2)$是完备的。*

证明：设$\{\boldsymbol{x}^{(k)}=(x_1^{(k)},x_2^{(k)},\cdots x_n^{(k)})\}_{k=1}^\infty$是$\boldsymbol{R}^n$中Cauchy列，则对$\varepsilon>0$，存在$N\in\mathbf{N}$，当$k_n,\ k_m>N$，有$d(\boldsymbol{x}^{(k_n)},\boldsymbol{x}^{(k_m)})<\varepsilon$。分量形成的数列$\{x_i^{(k)}\}_{k=1}^\infty$显然也是Cauchy 列，因为

$$|x_i^{(k_n)}-x_i^{(k_m)}|\le d(\boldsymbol{x}^{(k_n)},\boldsymbol{x}^{(k_m)})<\varepsilon.$$

因此，利用实数空间的完备性，存在实数 x_i 使得 $x_i^{(k)}\to x_i\ (k\to\infty)$。记 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$，则 $x\in\boldsymbol{R}^n$，且 $\boldsymbol{x}^{(k)}\to\boldsymbol{x}(k\to\infty)$。 证毕。

例 1.14 *l^∞是完备的。*

证明：设$\{\boldsymbol{x}^{(m)}=(x_1^{(m)},x_2^{(m)},\cdots x_n^{(m)},\cdots)\}_{m=1}^{\infty}$是$l^{\infty}$中Cauchy列，则对$\varepsilon>0$，都存在$N\in\mathbf{N}$，当$n,\ m>N$，有

$$d(\boldsymbol{x}^{(n)},\boldsymbol{x}^{(m)})=\sup_{j\in\mathbf{N}}|x_j^{(n)}-x_j^{(m)}|<\varepsilon.$$

类似于$\boldsymbol{R}^n$完备性的讨论，对每个分量$j\in\mathbf{N}$，$\{x_j^{(m)}\}_{m=1}^{\infty}$是实数空间的Cauchy列，从而存在$x_j\in\boldsymbol{R}$使得$x_j^{(m)}\to x_j\ (m\to\infty)$。令$\boldsymbol{x}=(x_1,x_2,\cdots x_n,\cdots)$。下面分别证$\boldsymbol{x}\in l^{\infty}$以及$\boldsymbol{x}^{(m)}\to\boldsymbol{x}\ (m\to\infty)$。

对任何$j\in\mathbf{N}$，在$|x_j^{(n)}-x_j^{(m)}|<\varepsilon$中令$m\to\infty$得：$|x_j^{(n)}-x_j|<\varepsilon$。因为$\boldsymbol{x}^{(n)}\in l^{\infty}$，从而存在$M>0$使得对任何$j\in\mathbf{N}$都有$|x_j^{(n)}|<M$。利用三角不等式得，对任何$j\in\mathbf{N}$，

$$|x_j|\le|x_j^{(n)}-x_j|+|x_j^{(n)}|\le\varepsilon+M.$$

即$\boldsymbol{x}\in l^{\infty}$。

由于对任何$j\in\mathbf{N}$，都成立$|x_j^{(n)}-x_j|\le\varepsilon$，可得对一切$n>N$，成立

$$d(\boldsymbol{x}^{(n)},\boldsymbol{x})=\sup_{i\in\mathbf{N}}|x_j^{(n)}-x_j|\le\varepsilon.$$

所以$\boldsymbol{x}^{(m)}\to\boldsymbol{x}(m\to\infty)$，因此$l^{\infty}$是完备的。 证毕。

例 1.15 *有理数空间是不完备的，因为有理数的Cauchy序列有可能极限是无理数，即有理数的Cauchy列在有理数空间中是不收敛的。*

不完备的距离空间是大量存在的，但是每一个不完备的空间都可以扩大成完备的距离空间。比如有理数空间$\mathbf{Q}$是不完备的，但是我们可以将有理数中加入新元素（无理数），使它成为新的完备距离空间$\boldsymbol{R}$，并且$\mathbf{Q}$在$\boldsymbol{R}$中稠密。事实上上述有理数完备化的过程对一般的不完备距离空间也可以类似地去做。下面介绍一下完备化的结果，具体证明在此略去。

定义 1.9 *设(X,d)，$(\widetilde{X},\tilde{d})$是两个距离空间。若存在$X$到$\widetilde{X}$上的一一映射$T$，使得对任何$x,\ y\in X$都有$\tilde{d}(Tx,Ty)=d(x,y)$。则称距离空间$(X,d)$与$(\widetilde{X},\tilde{d})$等距同构。映射$T$称为等距同构映射。*

定义 1.10 *设(X,d)，$(\widetilde{X},\tilde{d})$是两个距离空间，$(\widetilde{X},\tilde{d})$是完备的。若$\widetilde{X}$中含有一稠密子集$X_0$，并且$(X_0,\tilde{d})$与$(X,d)$等距同构。则称距离空间$(\widetilde{X},\tilde{d})$是$(X,d)$的完备化空间。*

定理 1.6 *在等距同构意义下，每一个距离空间(X,d)都有唯一的完备化空间$(\widetilde{X},\tilde{d})$。*

1.2.4 距离空间的列紧性

实数空间上每一个有界无限集至少有一个聚点，这被称作实数的聚点原理。但是在一般的距离空间中没有上述类似的结论。对于一般的距离空间，我们将介绍相

应的概念，以及解决距离空间什么样的情况下具有类似于实数的聚点原理。

首先介绍度量空间集合的一些基本概念。

定义 1.11 设(X,d)是一距离空间，$x_0 \in X$，对$r>0$，称集合$\{x \in X : d(x,x_0) < r\}$为以$x_0$的$r$邻域，记为$B(x_0,r)$；称集合$\{x \in X : d(x,x_0) \leq r\}$为以$x_0$中心，$r$为半径的闭球，记为$\overline{B}(x_0,r)$。$A \subset X$，若存在$x_0$的$r$邻域$B(x_0,r) \subset A$，则称$x_0$是$A$的内点；如果$x_0$的任何邻域$B(x_0,r)$都含有$A-\{x_0\}$的点，即$A \cap (A-\{x_0\}) \neq \emptyset$，则称$x_0$是$A$的聚点；如果$A$中每一点都是$A$的内点，则称$A$为开集；如果$A$中每一聚点都属于$A$，则称$A$为闭集；称$A$与$A$的所有聚点集的并集为$A$的闭包，记做$\overline{A}$。

通常规定，全集X和空集$\emptyset$既是开集又是闭集。根据定义，不难证明下面一些关于集合的性质。证明较简单，留给读者。

性质 1.2 设X是距离空间，X中开集具有如下性质

(1)任何多个开集的并都是开集；

(2)有限多个开集的交是开集。

性质 1.3 设X是距离空间，$A \subset X$，A为闭集的充要条件为$A=\overline{A}$。

性质 1.4 设X是距离空间，X中开集具有如下性质

(1)任何多个闭集的交都是闭集；

(2)有限多个闭集的并是闭集。

性质 1.5 设X是距离空间，$A \subset X$。若A是开集，则$X-A$是闭集；反之，若A是闭集，则$X-A$是开集。

性质 1.6 设X,Y是两个距离空间，$T: X \to Y$是一个映射，则下列命题是等价的。

(1) T是连续映射；

(2) 对于Y中任何开集B，$T^{-1}(B)$是X中的开集；

(3)对于Y中任何闭集B，$T^{-1}(B)$是X中的闭集；

在上述集合的一些概念基础上，下面介绍紧的相关概念和性质。

定义 1.12 设A是距离空间X的一个子集，如果A中任何点列都含有子列收敛到X中的点，则称A是列紧的；如果A中任何点列都含有子列收敛到A中的点，则称A是紧的。若X是紧的，则称X为紧的距离空间。

通过定义，可以较简单验证下面例子以及定理，具体证明留作习题。

例 1.16 (1) 任何距离空间上有限集是紧的；

(2) $A=[a,b] \subset \boldsymbol{R}$，$a,b \in \boldsymbol{R}$，则$A$是$\boldsymbol{R}$中的紧集；

(3) $\boldsymbol{R}$是非紧的，因为点列$\{n\}_{n=1}^{\infty}$不含任何收敛子列；

(4) $\boldsymbol{R}^n$上集合A是紧集当且仅当A是有界闭集。

定理 1.7 (1)如果A是列紧的，则$\overline{A}$是紧的；

(2) 列紧集的子集也是列紧集。

在$\boldsymbol{R}^n$空间上，有界闭集(紧集)上连续函数能取到最大值和最小值，在一般的距离空间上也有类似结论。

定理 1.8 设A是距离空间X中一紧集，f是定义在A上连续函数，那么f是有界的，并且上下确界可达。

证明：先证f有界。若不然，则存在 $\boldsymbol{x}_n \in A$，使$\lim\limits_{n\to\infty}|f(\boldsymbol{x}_n)| = +\infty$。由于$A$是紧集，从而存在子列（不妨仍记为$\boldsymbol{x}_n$）在$A$中收敛，即存在 $\boldsymbol{x}_0 \in A$使得$\lim\limits_{n\to\infty}\boldsymbol{x}_n = \boldsymbol{x}_0$。由于$f$在$A$上连续，所以有$f(\boldsymbol{x}_0) = \lim\limits_{n\to\infty} f(x_n) = +\infty$，这与$f$在$A$上有定义矛盾，从而$f$在$A$上有界。

再证f可以达到上下确界，下面只就达到上确界的情形给予证明，下确界的情形类似。记$\beta = \sup\limits_{\boldsymbol{x}\in A} f(\boldsymbol{x})$，则存在点列$\boldsymbol{x}_n \in A$使得 $f(\boldsymbol{x}_n) > \beta - \frac{1}{n}$,由于$A$是紧集，从而存在$x_n$的子列（不妨仍记为$\boldsymbol{x}_n$）以及$x_0 \in A$使得$\lim\limits_{n\to\infty}\boldsymbol{x}_n = \boldsymbol{x}_0$。由$f$在$A$上连续，所以有$\beta \geq f(\boldsymbol{x}_0) = \lim\limits_{n\to\infty} f(\boldsymbol{x}_n) \geq \beta$，从而$f(\boldsymbol{x}_0) = \beta$。 证毕。

1.3 压缩映射原理

作为完备距离空间，本节将介绍压缩映射原理。压缩映射原理是求解代数方程、微分方程、积分方程以及数值分析中迭代算法收敛的理论依据，是数学和工程计算中非常常用的方法。

定义 1.13 设X是度量空间，映射$T : X \to X$，如果对$\boldsymbol{x}_0 \in T$成立$T\boldsymbol{x}_0 = \boldsymbol{x}_0$，则称$\boldsymbol{x}_0$是$T$的不动点。

定义 1.14 设(X, d)是度量空间，映射$T : X \to X$，如果存在$a \in (0, 1)$，对任何$\boldsymbol{x}, \boldsymbol{y} \in X$满足$d(T\boldsymbol{x}, T\boldsymbol{y}) \leq a(\boldsymbol{x}, \boldsymbol{y})$，则称$T$为压缩映射。

定理 1.9 (压缩映射原理) 设(X, d)是完备度量空间，$T : X \to X$是压缩映射，那么T有唯一不动点。

证明：任取$\boldsymbol{x}_0 \in X$，作迭代序列$\boldsymbol{x}_n = T\boldsymbol{x}_{n-1}(n \geq 1)$。下面证明$\{\boldsymbol{x}_n\}$是收敛的，因为$X$是完备空间，只需证明 $\{\boldsymbol{x}_n\}$是Cauchy列。由于

$$(\boldsymbol{x}_2, \boldsymbol{x}_1) = d(T\boldsymbol{x}_1, T\boldsymbol{x}_0) \leq \alpha(\boldsymbol{x}_1, \boldsymbol{x}_0);$$

$$d(\boldsymbol{x}_3, \boldsymbol{x}_2) = d(T\boldsymbol{x}_2, T\boldsymbol{x}_1) \leq \alpha(\boldsymbol{x}_2, \boldsymbol{x}_1) \leq \alpha^2(\boldsymbol{x}_1, \boldsymbol{x}_0);$$

.....

$$d(\boldsymbol{x}_n, \boldsymbol{x}_{n-1}) \leq \alpha^{n-1}(\boldsymbol{x}_1, \boldsymbol{x}_0).$$

所以对任何自然数n和p，有

$$\begin{aligned} d(\boldsymbol{x}_{n+p}, \boldsymbol{x}_n) &\leq d(\boldsymbol{x}_{n+p}, \boldsymbol{x}_{n+p-1}) + d(\boldsymbol{x}_{n+p-1}, \boldsymbol{x}_{n+p-2}) + \cdots + d(\boldsymbol{x}_{n+1}, \boldsymbol{x}_n) \\ &\leq (\alpha^{n+p-1} + \alpha^{n+p-2} + \cdots + \alpha^n) d(\boldsymbol{x}_1, \boldsymbol{x}_0) \\ &= \frac{\alpha^n}{1-\alpha} d(\boldsymbol{x}_1, \boldsymbol{x}_0). \end{aligned}$$

由此可见 $\lim\limits_{n\to\infty} d(x_{n+p}, x_n) = 0$，从而$\{x_n\}$是Cauchy列。因此存在$x_* \in X$使得 $\lim\limits_{n\to\infty} x_n = x_*$。

下面证明x_*是T的不动点。事实上

$$\begin{aligned} d(T\boldsymbol{x}_*, \boldsymbol{x}_*) &\leq d(T\boldsymbol{x}_*, T\boldsymbol{x}_n) + d(T\boldsymbol{x}_n, \boldsymbol{x}_*) \\ &\leq \alpha d(\boldsymbol{x}_*, \boldsymbol{x}_n) + d(\boldsymbol{x}_{n+1}, \boldsymbol{x}_*) \\ &\to 0 \quad (n \to \infty). \end{aligned}$$

即$d(T\boldsymbol{x}_*, \boldsymbol{x}_*) = 0$，所以$T\boldsymbol{x}_* = \boldsymbol{x}_*$。

最后证明唯一性。假设存在另外一个不动点$\boldsymbol{y}_*$，则

$$d(\boldsymbol{x}_*, \boldsymbol{y}_*) = d(T\boldsymbol{x}_*, T\boldsymbol{y}_*) \leq \alpha d(\boldsymbol{x}_*, \boldsymbol{y}_*),$$

所以$d(\boldsymbol{x}_*, \boldsymbol{y}_*) = 0$，即$\boldsymbol{x}_* = \boldsymbol{y}_*$。 证毕。

在实际应用中，有时T本身不是压缩映射，但是T复合若干次后的映射T^n是压缩映射，这时T仍然有唯一的不动点，具体如以下定理。

定理 1.10 *设X是完备度量空间，映射$T : X \to X$。如果存在某个自然数n使得T^n是压缩映射，则T有唯一不动点。这里 T^n是T的n次复合。*

证明：由于T^n是压缩映射，所以存在唯一不动点$\boldsymbol{x}_*$，即$T^n\boldsymbol{x}_* = \boldsymbol{x}_*$。由于

$$T^n(T\boldsymbol{x}_*) = T^{n+1}\boldsymbol{x}_* = T(T^n\boldsymbol{x}_*) = T\boldsymbol{x}_*,$$

所以$T\boldsymbol{x}_*$仍是T^n的不动点，由压缩映射不动点的唯一性得：$T\boldsymbol{x}_* = \boldsymbol{x}_*$，即$\boldsymbol{x}_*$是$T$的不动点。

若$\boldsymbol{y}_*$也是T的不动点，则

$$T^n\boldsymbol{y}_* = T^{n-1}(T\boldsymbol{y}_*) = T^{n-1}\boldsymbol{y}_* = \cdots = \boldsymbol{y}_*,$$

所以$\boldsymbol{y}_*$也是T^n的不动点，由T^n不动点的唯一性得：$\boldsymbol{x}_* = \boldsymbol{y}_*$。 证毕。

下面分别通过代数方程、微分方程来说明压缩映射原理的应用。

例 1.17 线性代数方程$Ax=b$都可以表示成如下形式

$$\boldsymbol{x}=\boldsymbol{C}\boldsymbol{x}+\boldsymbol{D} \tag{1.1}$$

其中$\boldsymbol{C}=(c_{ij})_{n\times n}$，$\boldsymbol{D}=(d_1,d_2,\dots,d_n)^{\mathrm{T}}$。如果矩阵$\boldsymbol{C}$满足条件

$$\sum_{j=1}^{n}|c_{ij}|<1\ (i=1,2,\dots,n)$$

则式(1.1)存在唯一解，且此解可由迭代求得。

证明：取$X=\boldsymbol{R}^n$，定义度量为

$$\rho(\xi,\eta)=\max_{1\le i\le n}|a_i-b_i|$$
$$\xi=(a_1,a_2,\dots,a_n)^{\mathrm{T}},\ \eta=(b_1,b_2,\dots,b_n)^{\mathrm{T}}$$

构造映射$T:X\to X$为$T\boldsymbol{x}=\boldsymbol{C}\boldsymbol{x}+\boldsymbol{D}$，那么方程(1.1)的解等价于映射$T$的不动点。

对于$\boldsymbol{x}=(x_1,x_2,\dots,x_n)^{\mathrm{T}},\boldsymbol{y}=(y_1,y_2,\dots,y_n)^{\mathrm{T}}$，由于

$$\begin{aligned}\rho(Tx,Ty)&=\max_{1\le i\le n}\left|\sum_{j=1}^{n}(c_{ij}x_j+d_j)-\sum_{j=1}^{n}(c_{ij}y_j+d_j)\right|\\&=\max_{1\le i\le n}\left|c_{ij}(x_j-y_j)\right|\le\max_{1\le i\le n}|c_{ij}|\rho(x,y)\end{aligned}$$

记$a=\max_{1\le i\le n}\sum_{j=1}^{n}|c_{ij}|$，由条件$a<1$，因此$T$是压缩映像，于是$T$有唯一不动点，所以方程(1.1)有唯一解，且此解可由如下迭代序列

$$\boldsymbol{x}^{(k)}=\boldsymbol{C}\boldsymbol{x}^{(k-1)}+\boldsymbol{D}$$

近似计算求得。

证毕。

例 1.18 考察如下常微分方程的初值问题：

$$\begin{cases}\frac{\mathrm{d}y}{\mathrm{d}x}=f(x,y)\\ y(x_0)=y_0.\end{cases} \tag{1.2}$$

如果$f(x,y)$在$\boldsymbol{R}^2$上连续，且关于第二元y满足Lipschitz条件，即

$$|f(x,y_1)-f(x,y_2)|\le L|y_1-y_2|$$

这里$L>0$是常数，则方程(1.2)在$[x_0-\delta,x_0+\delta]$上有唯一解($\delta<\frac{1}{L}$)。

证明：方程(1.2)的解等价于如下方程

$$y(x)=y_0+\int_{x_0}^{x}f(t,y(t))\mathrm{d}t \tag{1.3}$$

的解。取连续函数空间$C[x_0-\delta,x_0+\delta]$（其上的距离以及完备性见习题9），定义其上的映射

$$T:C[x_0-\delta,x_0+\delta]\to C[x_0-\delta,x_0+\delta]$$

为

$$(Ty)(x)=y_0+\int_{x_0}^{x}f(t,y(t))\mathrm{d}t$$

则积分方程(1.3)的解等价于T的不动点。对任意两个连续函数$y_1(x),y_2(x)\in C[x_0-\delta,x_0+\delta]$，由于

$$\begin{aligned}\rho(Ty_1,Ty_2)&=\max_{x\in[x_0-\delta,x_0+\delta]}\left|\int_{x_0}^{x}[f(t,y_1(t))-f(t,y_2(t))]\mathrm{d}t\right|\\&\le\max_{x\in[x_0-\delta,x_0+\delta]}\int_{x_0}^{x}|f(t,y_1(t))-f(t,y_2(t))|\mathrm{d}t\\&\le\max_{x\in[x_0-\delta,x_0+\delta]}L\int_{x_0}^{x}|y_1(t)-y_2(t)|\mathrm{d}t\le\delta L\rho(y_1,y_2),\end{aligned}$$

令$a=L\delta$，则$a<1$，故T是压缩映射，从而T有唯一不动点，即积分方程(1.3)有唯一解，从而微分方程(1.2) 在$[x_0-\delta,x_0+\delta]$上有唯一解。 证毕。

1.4 范数与赋范空间，Banach空间

前面通过在集合上引入距离进而定义了距离空间，并且在距离空间上研究了点列的收敛以及映射的性质。 在线性空间中，由于元素之间可以进行线性运算，在其上一类比较特殊的距离将被引入。

1.4.1 范数与线性赋范空间

定义 1.15 设X是复数域$\boldsymbol{C}$上线性空间，$\mathbf{0}$为X的零元素， 若对X中每一个元素$\boldsymbol{x}$，按照一个法则对应一个实数$\|\boldsymbol{x}\|$满足

(1) $\|\boldsymbol{x}\|\ge 0$且$\|\boldsymbol{x}\|=0$当且仅当$\boldsymbol{x}=\mathbf{0}$;
(2) $\|\boldsymbol{x}+\boldsymbol{y}\|\le\|\boldsymbol{x}\|+\|\boldsymbol{y}\|$ $(\boldsymbol{x},\boldsymbol{y}\in X)$;
(3) $\|\alpha\boldsymbol{x}\|=|\alpha|\|\boldsymbol{x}\|$ $(\alpha\in\boldsymbol{C},\boldsymbol{x}\in X)$。

则称$\|x\|$为x的范数，X称为以$\|\cdot\|$为范数的赋范线性空间。

对于赋范线性空间，通过公式

$$d(\boldsymbol{x},\boldsymbol{y})=\|\boldsymbol{x}-\boldsymbol{y}\|$$

可以定义元素之间的距离，容易验证$d(\boldsymbol{x},\boldsymbol{y})$满足距离的三个条件，因而$X$是一个距离空间。从而距离空间上诸多概念都可以定义在赋范线性空间中，例如开集、闭集、收敛性、完备性、可分性、列紧性等。

在赋范线性空间X中，若$d(\boldsymbol{x}_n,\boldsymbol{x})=\|\boldsymbol{x}_n-\boldsymbol{x}\|\to 0\ (n\to\infty)$，记为$\lim\limits_{n\to\infty}\boldsymbol{x}_n=\boldsymbol{x}$或记为$\boldsymbol{x}_n\to\boldsymbol{x}\ (n\to\infty)$。

完备的赋范线性空间称为Banach空间。

例 1.19 *$\boldsymbol{C}^n$与l^p $(p>1)$分别在以下范数下是赋范线性空间，而且是Banach空间，*

$$\|\boldsymbol{x}\|=(\sum_{i=1}^{n}x_i^2)^{\frac{1}{2}},\boldsymbol{x}=(x_1,x_2,\cdots,x_n)\in\boldsymbol{C}^n,$$

$$\|\boldsymbol{x}\|=(\sum_{i=1}^{\infty}|x_i|^p)^{\frac{1}{p}},\boldsymbol{x}=(x_1,x_2,\cdots,x_n,\cdots)\in l^p.$$

1.4.2 赋范线性空间的性质

赋范线性空间上范数可以导出距离，从而它具有距离空间的一般性质。此外由于它是线性空间，赋范线性空间还具有其他一些性质，下面将做一些介绍。

定理 1.11 *设X是赋范线性空间，$\{\boldsymbol{x}_n\}\subset X$，$\boldsymbol{x}\in X$，若$\lim\limits_{n\to\infty}\boldsymbol{x}_n=\boldsymbol{x}$，则$\{\|\boldsymbol{x}_n\|\}$是有界数列。*

证明：因$\lim\limits_{n\to\infty}\|\boldsymbol{x}_n-\boldsymbol{x}\|=0$，所以对$\varepsilon=1$，存在自然数$N$，当$n>N$时，$\|\boldsymbol{x}_n-\boldsymbol{x}\|<1$，于是

$$\|\boldsymbol{x}_n\|\le\|\boldsymbol{x}\|+\|\boldsymbol{x}-\boldsymbol{x}_n\|<1+\|\boldsymbol{x}\|.$$

令$M=\max\{\|\boldsymbol{x}_1\|,\|\boldsymbol{x}_2\|,\cdots,\|\boldsymbol{x}_N\|,1+\|\boldsymbol{x}\|\}$，则对一切$n$成立$\|\boldsymbol{x}_n\|\le M$，即$\{\|\boldsymbol{x}_n\|\}$有界。 证毕。

定理 1.12 *设X是定义在$\boldsymbol{C}$上赋范线性空间，$\{\boldsymbol{x}_n\}$，$\{\boldsymbol{y}_n\}$为X中两点列，$\{\alpha_n\}\subset\boldsymbol{C}$满足*

$$\lim_{n\to\infty}\boldsymbol{x}_n=\boldsymbol{x},\lim_{n\to\infty}\boldsymbol{y}_n=\boldsymbol{y},\lim_{n\to\infty}\alpha_n=\alpha,(\boldsymbol{x},\boldsymbol{y}\in X,\alpha\in\boldsymbol{C}),$$

则 (1) $\lim\limits_{n\to\infty}\boldsymbol{x}_n+\boldsymbol{y}_n=\boldsymbol{x}+\boldsymbol{y}$，(2) $\lim\limits_{n\to\infty}\alpha_n\boldsymbol{x}_n=\alpha\boldsymbol{x}$。

证明：(1) 由

$$\|\boldsymbol{x}_n+\boldsymbol{y}_n-\boldsymbol{x}-\boldsymbol{y}\|\le\|\boldsymbol{x}_n-\boldsymbol{x}\|+\|\boldsymbol{y}_n-\boldsymbol{y}\|,$$

得$\lim\limits_{n\to\infty}\boldsymbol{x}_n+\boldsymbol{y}_n=\boldsymbol{x}+\boldsymbol{y}$。

(2)由$\lim\limits_{n\to\infty}\boldsymbol{x}_n=\boldsymbol{x}$得$\{\|\boldsymbol{x}_n\|\}$有界，即存在$M>0$使得$\|\boldsymbol{x}_n\|\le M$，于是

$$\begin{aligned}\|\alpha_n\boldsymbol{x}_n-\alpha\boldsymbol{x}\| &= \|\alpha_n\boldsymbol{x}_n-\alpha\boldsymbol{x}_n+\alpha\boldsymbol{x}_n-\alpha\boldsymbol{x}\|\\ &\le \|\alpha_n\boldsymbol{x}_n-\alpha\boldsymbol{x}_n\|+\|\alpha\boldsymbol{x}_n-\alpha\boldsymbol{x}\|\\ &\le |\alpha_n-\alpha|\|\boldsymbol{x}_n\|+|\alpha|\|\boldsymbol{x}_n-\boldsymbol{x}\|\\ &\le M|\alpha_n-\alpha|+|\alpha|\|\boldsymbol{x}_n-\boldsymbol{x}\|\to 0(n\to\infty).\end{aligned}$$

所以 $\lim\limits_{n\to\infty}\alpha_n\boldsymbol{x}_n=\alpha\boldsymbol{x}$。 证毕。

在同一个线性空间上可以定义不同的范数，它们之间的关系可以有“强弱”之分，具体描述以及刻画如下。

定义 1.16 *设X是一线性空间，$\|\cdot\|_1$和$\|\cdot\|_2$是X上两个范数，如果*

$$\lim_{n\to\infty}\|\boldsymbol{x}_n\|_1=0\Longrightarrow\lim_{n\to\infty}\|\boldsymbol{x}_n\|_2=0,$$

则称$\|\cdot\|_1$强于$\|\cdot\|_2$；如果

$$\lim_{n\to\infty}\|\boldsymbol{x}_n\|_1=0\Longleftrightarrow\lim_{n\to\infty}\|\boldsymbol{x}_n\|_2=0,$$

则称$\|\cdot\|_1$等价于$\|\cdot\|_2$。

定理 1.13 *设X是一线性空间，$\|\cdot\|_1$和$\|\cdot\|_2$是X上两个范数，$\|\cdot\|_1$强于$\|\cdot\|_2$的充要条件为存在常数$\lambda>0$使得$\|\boldsymbol{x}\|_2\le\lambda\|\boldsymbol{x}\|_1$。*

证明：根据$\|\cdot\|_1$强于$\|\cdot\|_2$的定义，充分性显然成立。下面证明必要性。

假设定理中的不等式不成立，即对任何$n\in\mathbf{N}$，存在$\boldsymbol{x}_n\in X$使得$\|\boldsymbol{x}_n\|_2>n\|\boldsymbol{x}\|_1$。令$\boldsymbol{y}_n=\frac{1}{\|\boldsymbol{x}_n\|_2}\boldsymbol{x}_n$，则

$$\lim_{n\to\infty}\|\boldsymbol{y}_n\|_1=\lim_{n\to\infty}\frac{\|\boldsymbol{x}_n\|_1}{\|\boldsymbol{x}_n\|_2}<\lim_{n\to\infty}\frac{1}{n}=0,\ \lim_{n\to\infty}\|\boldsymbol{y}_n\|_2=1.$$

这与$\|\cdot\|_1$强于$\|\cdot\|_2$矛盾。 证毕。

从定理的结论可知，两个范数等价当且仅当存在常数$\lambda\ge\mu>0$使得

$$\lambda\|\boldsymbol{x}\|_1\le\|\boldsymbol{x}\|_2\le\mu\|\boldsymbol{x}\|_1.$$

1.4.3 有限维赋范线性空间

当赋范线性空间的维数有限时，赋范线性空间将具有一些良好的性质，下面做一些介绍。

引理 1.1 *设X是一个n维实赋范线性空间，$\{e_1,e_2,\cdots,e_n\}$是X的一个基，则存在正数M及M'，$M'\ge M$，使得对一切$\boldsymbol{x}=\sum\limits_{k=1}^{n}x_ke_k\in X$，下列不等式成立*

$$M\|\boldsymbol{x}\|\le\left(\sum_{k=1}^{n}|x_k|^2\right)^{\frac{1}{2}}\le M'\|\boldsymbol{x}\|. \tag{1.4}$$

证明：对$\boldsymbol{x} \in X$, 有

$$\begin{aligned}\|\boldsymbol{x}\| = \left\|\sum_{k=1}^{n} x_k e_k\right\| &\leq \sum_{k=1}^{n} \|e_k\||x_k| \\ &\leq \left(\sum_{k=1}^{n} \|e_k\|^2\right)^{\frac{1}{2}} \left(\sum_{k=1}^{n} x_k^2\right)^{\frac{1}{2}} \\ &= m\left(\sum_{k=1}^{n} |x_k|^2\right)^{\frac{1}{2}}, \text{其中 } m = \left(\sum_{k=1}^{n} \|e_k\|^2\right)^{\frac{1}{2}}.\end{aligned}$$

再任取$\boldsymbol{y} = \sum_{k=1}^{n} y_k e_k \in X$，由上面不等式得：

$$\|\boldsymbol{x} - \boldsymbol{y}\| \leq m\left(\sum_{k=1}^{n} |x_k - y_k|^2\right)^{\frac{1}{2}}$$

于是便有

$$|\|\boldsymbol{x}\| - \|\boldsymbol{y}\|| \leq \|\boldsymbol{x} - \boldsymbol{y}\| \leq m\left(\sum_{k=1}^{n} |x_k - y_k|^2\right)^{\frac{1}{2}}.$$

将$(x_1, x_2, \cdots, x_n) = \boldsymbol{x}$，$(y_1, y_2, \cdots, y_n) = \boldsymbol{y}$看成$\boldsymbol{R}^n$的中点，令

$$f(\boldsymbol{x}) = \|\boldsymbol{x}\|, \quad \left(\boldsymbol{x} = \sum_{k=1}^{n} x_k e_k\right),$$

则

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq m\|\boldsymbol{x} - \boldsymbol{y}\|.$$

所以f是$\boldsymbol{R}^n$到$\boldsymbol{R}$的连续函数，取$\boldsymbol{R}^n$单位球面

$$S = \left\{\boldsymbol{x} = (x_1, x_2, \cdots, x_n) : \left(\sum_{k=1}^{n} |x_k|^2\right)^{\frac{1}{2}} = 1\right\}.$$

则S是$\boldsymbol{R}^n$中的有界闭集且不含零元素，故$f(\boldsymbol{x})$在S中任一点处都不为零且上、下确界可达，从而存在下确界$m' > 0$，使$|f(\boldsymbol{x})| \geq m'$或$\|\boldsymbol{x}\| \geq m'$。

任取$\boldsymbol{x} = \sum_{k=1}^{n} x_k e_k$，且$\boldsymbol{x} \neq \boldsymbol{0}$，则$(\sum_{k=1}^{n} |x_k|^2)^{\frac{1}{2}} \neq 0$. 令

$$\boldsymbol{x}' = \frac{\sum_{k=1}^{n} x_k e_k}{\left(\sum_{k=1}^{n} |x_k|^2\right)^{\frac{1}{2}}} = \frac{x}{\left(\sum_{k=1}^{n} |x_k|^2\right)^{\frac{1}{2}}},$$

$$\|\boldsymbol{x}'\| = f(\boldsymbol{x}') \geq m',$$

即

$$\|\boldsymbol{x}\| = \left(\sum_{k=1}^{n}|x_k|^2\right)^{\frac{1}{2}} \|x'\| \geq m'\left(\sum_{k=1}^{n}|x_k|^2\right)^{\frac{1}{2}}. \tag{1.5}$$

分别在式(1.4)和(1.5)中令$M=\frac{1}{m}$、$M'=\frac{1}{m'}$，则不等式得证。 证毕。

定理 1.14 设X是m维实线性空间，$\|.\|_1$与$\|.\|_2$是X上定义的两个范数，则$\|.\|_1$与$\|.\|_2$等价。

证明：由引理1.1知，存在常数$M_1' \geq M_1 > 0$及$M_2' \geq M_2 > 0$满足

$$M_1\left(\sum_{k=1}^{n}|x_k|^2\right)^{\frac{1}{2}} \leq \|\boldsymbol{x}\|_1 \leq M_1'\left(\sum_{k=1}^{n}|x_k|^2\right)^{\frac{1}{2}},$$

$$M_2\left(\sum_{k=1}^{n}|x_k|^2\right)^{\frac{1}{2}} \leq \|\boldsymbol{x}\|_2 \leq M_2'\left(\sum_{k=1}^{n}|x_k|^2\right)^{\frac{1}{2}}$$

这里$\boldsymbol{x}=\sum_{k=1}^{n}x_k e_k$。组合以上两个不等式有

$$\frac{M_1}{M_2'}\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_1 \leq \frac{M_1'}{M_2}\|\boldsymbol{x}\|_2.$$

证毕。

定理 1.15 任意n维实赋范线性空间X必与R^n同构且同胚。

证明：任取 X 中一个基 $\{e_1, e_2, \cdots, e_n\}$，设 $\boldsymbol{x}=\sum_{k=1}^{n}x_k e_k \in X$，且 $\xi=(x_1, x_2, \cdots, x_n)$ 看成 $\boldsymbol{R}^n$ 中点，作 $\boldsymbol{R}^n$ 到 X 上的同构映射 $T:\xi \to \boldsymbol{x}$，则由定理1.14可知，T 是连续的且 T^{-1} 也是连续的，故 X 与 $\boldsymbol{R}^n$ 同胚。 证毕。

由该定理可知，任意一个n维实赋范线性空间与$\boldsymbol{R}^n$空间的代数结构和分析性质一致。因此在只讨论有限维实赋范线性空间的代数性质与分析性质时只需研究空间$\boldsymbol{R}^n$代数和分析性质即可。

注 1.1 本节所有结论对有限维复线性空间及相应的 $\boldsymbol{C}^n$ 均成立。

1.5 Hilbert空间，正交系

上一节主要是将$\boldsymbol{R}^n$中模长推广到线性空间上的范数。在$\boldsymbol{R}^n$上，向量之间可以讨论夹角，而这对只有模长概念的赋范线性空间是做不到。$\boldsymbol{R}^n$上向量之间的夹角是通过向量的内积来描述的，本节将对一些空间引入内积以及讨论可以定义内积的距离空间性质。

1.5.1 内积的一般概念

在R^3中，设$\boldsymbol{x}=(x_1,x_2,x_3)$，$\boldsymbol{y}=(y_1,y_2,y_3)$，则他们的内积定义为$\langle \boldsymbol{x},\boldsymbol{y}\rangle=\sum_{i=1}^{3}x_iy_i$，不难发现$\|\boldsymbol{x}\|=\sqrt{\langle \boldsymbol{x},\boldsymbol{x}\rangle}$。容易得到，内积具有如下一些性质。

(1) $\langle \boldsymbol{x},\boldsymbol{y}\rangle\geq 0,\forall \boldsymbol{x}\in \boldsymbol{R}^3$，且$\langle \boldsymbol{x},\boldsymbol{x}\rangle=0\Longleftrightarrow \boldsymbol{x}=\boldsymbol{0}$;

(2) $\langle \boldsymbol{x},\boldsymbol{y}\rangle=\langle \boldsymbol{y},\boldsymbol{x}\rangle,\forall \boldsymbol{x},\boldsymbol{y}\in \boldsymbol{R}^3$;

(3) $\langle \boldsymbol{x}_1+\boldsymbol{x}_2,\boldsymbol{y}\rangle=\langle \boldsymbol{x}_1,\boldsymbol{y}\rangle+\langle\ \boldsymbol{x}_2,\boldsymbol{y}\rangle,\forall \boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\in \boldsymbol{R}^3$;

(4) $\langle \lambda\boldsymbol{x},\boldsymbol{y}\rangle=\lambda\langle \boldsymbol{x},\boldsymbol{y}\rangle,\forall \boldsymbol{x}\in \boldsymbol{R}^3,\forall\lambda\in \boldsymbol{R}$.

由此，我们引入内积的概念以及内积空间。

定义 1.17 设X为$\boldsymbol{C}$上的线性空间，若对任何两个元素（也称为向量）$\boldsymbol{x},\boldsymbol{y}\in X$，有唯一$\boldsymbol{C}$中的数与之对应，记为$\langle \boldsymbol{x},\boldsymbol{y}\rangle$，并且满足下面性质：

(1) $\langle \boldsymbol{x},\boldsymbol{y}\rangle\geq 0,\forall \boldsymbol{x}\in X$，且$\langle \boldsymbol{x},\boldsymbol{x}\rangle=0\Longleftrightarrow \boldsymbol{x}=\boldsymbol{0}$;

(2) $\langle \boldsymbol{x},\boldsymbol{y}\rangle=\overline{\langle \boldsymbol{y},\boldsymbol{x}\rangle},\forall \boldsymbol{x},\boldsymbol{y}\in X$;

(3) $\langle \boldsymbol{x}_1+\boldsymbol{x}_2,\boldsymbol{y}\rangle=\langle \boldsymbol{x}_1,\boldsymbol{y}\rangle+\langle \boldsymbol{x}_2,\boldsymbol{y}\rangle,\forall \boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}\in X$;

(4) $\langle \lambda\boldsymbol{x},\boldsymbol{y}\rangle=\lambda\langle \boldsymbol{x},\boldsymbol{y}\rangle,\forall \boldsymbol{x}\in X,\forall\lambda\in \boldsymbol{C}$;

则称$\langle \boldsymbol{x},\boldsymbol{y}\rangle$为$\boldsymbol{x}$与$\boldsymbol{y}$的内积，有了内积的线性空间称作内积空间，称$X$为复内积空间，若取$\boldsymbol{C}$为$\boldsymbol{R}$，则称$X$为实内积空间。

今后讨论中，不加注明，恒设X为复内积空间。

定理 1.16 设X为内积空间空间，对任何$\boldsymbol{x}\in X$，令$\|\boldsymbol{x}\|=\sqrt{\langle \boldsymbol{x},\boldsymbol{x}\rangle}$，则$\|\boldsymbol{x}\|$是$\boldsymbol{x}$的范数。

证明：由于范数的前两条性质可直接由内积的性质推出，只需验证它满足第三条性质（三角不等式）。事实上

$$\begin{aligned}\|\boldsymbol{x}+\boldsymbol{y}\|^2 &= \langle \boldsymbol{x}+\boldsymbol{y},\boldsymbol{x}+\boldsymbol{y}\rangle=\langle \boldsymbol{x},\boldsymbol{x}\rangle+\langle \boldsymbol{x},\boldsymbol{y}\rangle+\langle \boldsymbol{y},\boldsymbol{x}\rangle+\langle \boldsymbol{y},\boldsymbol{y}\rangle\\ &\leq \|\boldsymbol{x}\|^2+2\|\boldsymbol{x}\|\|\boldsymbol{y}\|+\|\boldsymbol{y}\|^2=(\|\boldsymbol{x}\|+\|\boldsymbol{y}\|)^2.\end{aligned}$$

从而$\|\boldsymbol{x}+\boldsymbol{y}\|\leq\|\boldsymbol{x}\|+\|\boldsymbol{y}\|$。 证毕。

通常我们称$\|\boldsymbol{x}\|=\sqrt{\langle \boldsymbol{x},\boldsymbol{x}\rangle}$为内积导出的范数，于是内积空间按照此范数成为一个赋范线性空间。因此关于赋范线性空间的结论对内积空间都成立。当内积空间X按由内积导出的范数完备时，称X为Hilbert空间。

例 1.20 对于$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$，$\boldsymbol{y}=(y_1,y_2,\cdots,y_n)\in \boldsymbol{C}^n$，定义

$$\langle \boldsymbol{x},\boldsymbol{y}\rangle=\sum_{i=1}^{n}x_i\bar{y}_i,$$

可以验证$\boldsymbol{C}^n$成为一个Hilbert空间。

例 1.21 对$\boldsymbol{x}=(x_1,x_2,\cdots,x_n,\cdots)$，$\boldsymbol{y}=(y_1,y_2,\cdots,y_n,\cdots)\in l^2$，定义

$$\langle \boldsymbol{x},\boldsymbol{y}\rangle=\sum_{i=1}^{n}x_i\bar{y}_i,$$

可以验证l^2成为一个Hilbert空间。

关于内积，以下几条性质是显然的，具体证明留给读者。

(1)内积的连续性。设$\lim\limits_{n\to\infty}\boldsymbol{x}_n=\boldsymbol{x}$，$\lim\limits_{n\to\infty}\boldsymbol{y}_n=\boldsymbol{y}$，则有

$$\lim_{n\to\infty}\langle \boldsymbol{x}_n,\boldsymbol{y}_n\rangle=\langle \boldsymbol{x},\boldsymbol{y}\rangle.$$

(2)极化恒等式。对内积空间X中的元素$\boldsymbol{x}$与$\boldsymbol{y}$，成立

$$\langle \boldsymbol{x},\boldsymbol{y}\rangle=\frac{1}{4}(\|\boldsymbol{x}+\boldsymbol{y}\|^2-\|\boldsymbol{x}-\boldsymbol{y}\|^2+i\|\boldsymbol{x}+i\boldsymbol{y}\|^2-i\|\boldsymbol{x}-i\boldsymbol{y}\|^2).$$

(3)平行四边形法则。对内积空间X中的元素$\boldsymbol{x}$与$\boldsymbol{y}$，成立

$$\|\boldsymbol{x}+\boldsymbol{y}\|^2+\|\boldsymbol{x}-\boldsymbol{y}\|^2=2(\|\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2).$$

1.5.2 正交系

类似于$\boldsymbol{R}^n$两个向量正交概念，对一般的内积空间，我们引入正交的概念如下。

定义 1.18 设X是内积空间，$\boldsymbol{x},\boldsymbol{y}\in X$，若$\langle \boldsymbol{x},\boldsymbol{y}\rangle=0$，则称$\boldsymbol{x}$与$\boldsymbol{y}$正交，记为$\boldsymbol{x}\perp\boldsymbol{y}$；设$\boldsymbol{x}\in X,M\subset X$，若$\boldsymbol{x}$与$M$中每个元素都正交，则称$\boldsymbol{x}$与$M$正交，记为$\boldsymbol{x}\perp M$；若有$N\subset X$，对任何$\boldsymbol{x}\in M,\boldsymbol{y}\in N$，都有$\boldsymbol{x}\perp\boldsymbol{y}$，则称$M$与$N$正交，记为$M\perp N$；对$M\subset X$，记$M^{\perp}=\{\boldsymbol{x}\in X:\boldsymbol{x}\perp M\}$并称$M^{\perp}$为$M$的正交补。

仿照$\boldsymbol{R}^n$中直角坐标系，在内积空间中也可以引入类似的概念。

定义 1.19 设M是内积空间X中不含零元的子集，若M中任何两个元素都正交，则称M为X的一个正交系。又若M中每个元素的范数都是1，则称M为标准正交系。若$\{\boldsymbol{e}_n\}_{n=1}^{\infty}$是内积空间$X$中的一个标准正交系，若对任何$\boldsymbol{x}\in X$，成立$\langle \boldsymbol{x},\boldsymbol{e}_n\rangle=0(n=1,2,\cdots)\Longrightarrow\boldsymbol{x}=\boldsymbol{0}$，则称$\{\boldsymbol{e}_n\}_{n=1}^{\infty}$为完全正交系。

例 1.22 在$\boldsymbol{R}^n$或$\boldsymbol{C}^n$中，

$$\boldsymbol{e}_1=(1,0,\cdots,0),\ \boldsymbol{e}_2=(0,1,0,\cdots,0),\ \cdots,\ \boldsymbol{e}_n=(0,0,\cdots,0,1)$$

是一个标准正交系。

例 1.23 在内积空间l^2中，以下元素列是一个标准正交系

$$\boldsymbol{e}_1=(1,0,\cdots,0),\ \boldsymbol{e}_2=(0,1,0,\cdots,0),\ \cdots,\ \boldsymbol{e}_n=(0,0,\cdots,0,1),\ \cdots$$

其中$\boldsymbol{e}_n=(0,\cdots0,1,0,\cdots)$第$n$个分量为1，其他分量都是0，$n=1,2,\cdots$。

下面的定理提供了判断正交系是完全正交系的方法。

定理 1.17 设$\{e_n\}_{n=1}^{\infty}$是内积空间X中的一个标准正交系，则以下命题等价。

(1) $\{e_n\}_{n=1}^{\infty}$是完全正交系；

(2) 对任何$x \in X$，成立$\|x\|^2 = \sum_{n=1}^{\infty} |c_n|^2$ (Parseval 等式)，其中$c_n = \langle x, e_n \rangle$；

(3) 对任何$x \in X$，有$x = \sum_{n=1}^{\infty} c_n e_n$，其中$c_n = \langle x, e_n \rangle$；

(4) 对任何两个元素$x, y \in X$有

$$\langle x, y \rangle = \sum_{n=1}^{\infty} \langle x, e_n \rangle \cdot \overline{\langle y, e_n \rangle}.$$

证明：(1) $\Longrightarrow$ (2) 需要Hilbert空间的一些准备知识，在此略去证明，有兴趣的读者可以参考一般的泛函分析教材。

(2) $\Longrightarrow$ (3) 任取$x \in X$，对任何$n \in \mathbf{N}$，令$c_n = \langle x, e_n \rangle, s_n = \sum_{k=1}^{n} c_k e_k$。注意到$\{e_n\}_{n=1}^{\infty}$是正交系，对一切$n \in \mathbf{N}$有$\langle x, e_n \rangle = 0$，假设(2)成立，则

$$\|x - s_n\|^2 = \langle x - s_n, x - s_n \rangle = \|x\|^2 - \sum_{k=1}^{n} |c_k|^2 \to 0 \quad (n \to \infty).$$

所以$\lim_{n\to\infty} \sum_{k=1}^{\infty} c_k e_k = x$。

(3) $\Longrightarrow$ (4) 任取$x, y \in X$,令$c_n = \langle x, e_n \rangle, d_n = \langle y, e_n \rangle$，则有

$$x = \lim_{n\to\infty} \sum_{k=1}^{\infty} c_k e_k, y = \lim_{n\to\infty} \sum_{k=1}^{\infty} d_k e_k.$$

于是得：

$$\langle x, y \rangle = \lim_{n\to\infty} \langle \sum_{k=1}^{n} c_k e_k, \sum_{k=1}^{n} d_k e_k \rangle = \lim_{n\to\infty} \sum_{k=1}^{n} c_k \overline{d_k} = \sum_{n=1}^{\infty} c_n \overline{d_n}.$$

(4) $\Longrightarrow$ (1) 任取$x \in X$，若对任何$n \in \mathbf{N}$成立$x \perp e_n$，则由(3)可得：

$$\|x\|^2 = \langle x, x \rangle = \sum_{n=1}^{\infty} \langle x, e_n \rangle \overline{\langle x, e_n \rangle} = 0,$$

即$x = \mathbf{0}$，所以(1)成立。 证毕。

从上面的定理可以看出，内积空间中任何一个向量可以用正交系线性表示，其组合系数有内积直接给出，非常方便。下面介绍一个得到标准正交系的方法。

设$\{x_n\}_{n=1}^{\infty}$是某线性无关序列，通过Gram-Schmidt标准正交化过程可获得一个标准正交系，具体过程如下：

第一步：把x_1标准化，令

$$e_1 = \frac{x_1}{\|x_1\|}.$$

第二步，记$X_1 = span\{e_1\} = span\{x_1\}$，其中$span\{x_1\}$表示由$x_1$的线性组合形成的空间，记$x_2 = \langle x_2, e_1\rangle e_1 + y_2$，则$y_2 \perp e_1$。因为$x_2$与$e_1$线性无关，则$y_2 \neq 0$，令

$$e_2 = \frac{y_2}{\|y_2\|}.$$

容易验证$span\{e_1, e_2\} = span\{x_1, x_2\}$。

于是利用归纳法，记$X_{n-1} = span\{e_1, e_2, \cdots, e_{n-1}\}$，及

$$x_n = \sum_{k=1}^{n-1} \langle x_n, e_k\rangle e_k + y_n.$$

则 $y_n \perp e_k, k = 1, 2, \cdots n-1$，又因$x_n, e_1, e_2, \cdots, e_{n-1}$线性无关，得$y_n \neq 0$，令

$$e_n = \frac{y_n}{\|y_n\|}.$$

则容易验证$span\{e_1, e_2, \cdots, e_n\} = span\{x_1, x_2, \cdots, e_n\}$。

于是以上过程，即得到一个标准正交系$\{e_n\}_{n=1}^{\infty}$。

最后,下面定理介绍可分Hilbert空间的标准正交系。

定理 1.18 *设X是*Hilbert*空间，则*

(1) *若X是可分的，则X必有至多可数的完全标准正交系；*

(2) *若X是无限维的可分空间，则X的每个完全标准正交系都是可数的；此时X与l^2等距同构。*

证明：(1)我们先利用反证法证明结论：若X是可分，则必存在至多可数个线性无关的元素$\{x_k\}$ 使得$\overline{span\{x_k\}} = X$。假设不存在至多可数个线性无关的元素$\{x_k\}$使得$\overline{span\{x_k\}} = X$， 则存在$X$的一组不可数基，记为$\{e_k\}_{k\in I}$，$I$为不可数的指标集。通过Gram-Schmidt标准正交化过程，我们可以得到一个不可数标准正交系， 不妨仍记为记为$\{e_k\}_{k\in I}$。对$\varepsilon_0 = \frac{1}{4}$，由于$X$是可分的，所以存在$X$的至多可数子集$\{x_n\}$使得 $\bigcup_n B(x_n, \frac{1}{4}) = X$。由于$\{x_n\}$可数，$\{e_k\}_{k\in I}$不可数，从而并存在两个向量(记为$e_1, e_2$)以及某个球(记为$B(x_m, \frac{1}{4})$，使得$e_1, e_2 \in B(x_m, \frac{1}{4})$。然而，由于$\{e_k\}_{k\in I}$为标准正交系，所以 $2 = \|e_1 - e_2\|^2$。这与$e_1, e_2 \in B(x_m, \frac{1}{4})$矛盾。

由于X存在至多可数个线性无关元素$\{x_k\}$使得$\overline{span\{x_k\}} = X$，通过Gram-Schmidt标准正交化过程，我们可以得到一个至多可数 的标准正交系。

(2). 由于X为无限维可分空间，由(1)知存在可数标准正交系$\{e_n\}_{n=1}^{\infty}$使得$X = span\{e_n\}$。 假设$\{x_k\}$为X的另一个完全标准正交系，同样由于X为无限维，所以$\{x_k\}$至少为可数集，再用(1)的证明方法，可以证明$\{x_k\}$一定为可数集。

设$\{e_n\}_{n=1}^{\infty}$为X的一个标准正交系，对任何$x \in X$记$c_n = \langle x, e_n \rangle$, $n = 1, 2, \cdots$。定义映射T如下，

$$T : X \longrightarrow l^2, \ Tx = (c_1, c_2, \cdots).$$

由Parseval等式得T是X到l^2的等距映射，从而X与l^2等距同构。 证毕。

1.6 向量范数，矩阵范数及其性质

对于n维复空间C^n中的向量x与y，可以通过向量 $x - y$ 的Euclid长度来描述 x 与 y 之间的距离。对$C^{n\times n}$中的矩阵A与B，如何来度量A与B之间的距离？此外C^n中的向量x与y，除Euclid长度外，有没有其他的距离？本节将利用范数来给出这些问题的解答。

1.6.1 向量范数

在C^n中，第1.4节例1.19已经定义了一种范数，现再定义以下一些范数。

例 1.24 对 $x = (\xi_1, \xi_2, \cdots, \xi_n) \in C^n$，可以分别定义以下范数:

$$\|x\|_1 = \sum_{i=1}^{n} |\xi_i|, \ \ \|x\|_2 = \left(\sum_{i=1}^{n} |\xi_i|^2 \right)^{1/2},$$

$$\|x\|_\infty = \max_{1\le i\le n} |\xi_i|, \ \ \|x\|_p = \left(\sum_{i=1}^{n} |\xi_i|^p \right)^{1/p}.$$

因为

$$\|x\|_\infty \le \|x\|_p \le \sqrt[p]{n}\|x\|_\infty, \tag{1.6}$$

所以 $\|x\|_\infty = \lim\limits_{p\to\infty} \|x\|_p$. 从而对于以上例题，我们只要证明 $\|x\|_p$ 对任何 $p \ge 1$ 是 C^n 上的向量范数就可以。为此，我们先证明以下两个命题。

命题 1.1 (Hölder 不等式) 设 $x = (\xi_1, \xi_2, \cdots, \xi_n)^T$, $y = (\eta_1, \eta_2, \cdots, \eta_n)^T \in C^n$. 则有以下不等式

$$\sum_{i=1}^{n} |\xi_i \eta_i| \le \left(\sum_{i=1}^{n} |\xi_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} |\eta_i|^q \right)^{1/q}, \tag{1.7}$$

其中 $p > 1, q > 1, \dfrac{1}{p} + \dfrac{1}{q} = 1.$

证明：因为函数 $f(z) = z^p$ 在区间 $[0, \infty)$ 上是凸函数，所以对任意 $z_1, z_2, \cdots, z_n \ge 0$ 及 $\alpha_1, \alpha_2, \cdots, \alpha_n \ge 0$, $\alpha_1 + \alpha_2 + \cdots + \alpha_n = 1$, 有

$$f\left(\sum_{i=1}^{n} \alpha_i z_i \right) \le \sum_{i=1}^{n} \alpha_i f(z_i). \tag{1.8}$$

如果 $\boldsymbol{x}=0$ 或 $\boldsymbol{y}=0$，则 Hölder 不等式 (1.7) 显然存立。现设 $\boldsymbol{x}\neq 0$ 及 $\boldsymbol{y}\neq 0$，即 $\xi_1,\xi_2,\cdots,\xi_n$ 不全为零，及 $\eta_1,\eta_2,\cdots,\eta_n$ 不全为零。由于等于零的 ξ_i 及 η_i 可使不等式 (1.7) 两边的相应项省略，所以不妨设 $\xi_1,\xi_2,\cdots,\xi_n$ 全不为零，及 $\eta_1,\eta_2,\cdots,\eta_n$ 全不为零。记

$$z_i=|\xi_i||\eta_i|^{-q/p},\ \alpha_i=\frac{|\eta_i|^q}{\sum\limits_{i=1}^{n}|\eta_i|^q},\quad i=1,2,\cdots,n.$$

代入不等式 (1.8) 可得

$$\left(\frac{\sum\limits_{i=1}^{n}|\xi_i\eta_i|}{\sum\limits_{i=1}^{n}|\eta_i|^q}\right)^p=f\left(\sum_{i=1}^{n}\alpha_i z_i\right)\leq\sum_{i=1}^{n}\alpha_i f(z_i)=\frac{\sum\limits_{i=1}^{n}|\xi_i|^p}{\sum\limits_{i=1}^{n}|\eta_i|^q}.$$

即有

$$\left(\sum_{i=1}^{n}|\xi_i\eta_i|\right)^p\leq\sum_{i=1}^{n}|\xi_i|^p\left(\sum_{i=1}^{n}|\eta_i|^q\right)^{p/q}.$$

可得 Hölder 不等式 (1.7)。 证毕。

注 1.2 当 $p=q=2$ 时，Hölder 不等式 (1.7) 也称为 Cauchy-Schwarz 不等式。

命题 1.2 (Minkowski 不等式) 设 $\boldsymbol{x}=(\xi_1,\xi_2,\cdots,\xi_n)^T$，$\boldsymbol{y}=(\eta_1,\eta_2,\cdots,\eta_n)^T\in\boldsymbol{C}^n$. 则有以下不等式

$$\left(\sum_{i=1}^{n}|\xi_i+\eta_i|^p\right)^{1/p}\leq\left(\sum_{i=1}^{n}|\xi_i|^p\right)^{1/p}+\left(\sum_{i=1}^{n}|\eta_i|^p\right)^{1/p},\tag{1.9}$$

其中 $p\geq 1$.

证明：当 $p=1$ 时，

$$\sum_{i=1}^{n}|\xi_i+\eta_i|\leq\sum_{i=1}^{n}(|\xi_i+|\eta_i|)=\sum_{i=1}^{n}|\xi_i|+\sum_{i=1}^{n}|\eta_i|.$$

即 Minkowski 不等式 (1.9) 成立。当 $p>1$ 时，如果 $\sum\limits_{i=1}^{n}|\xi_i+\eta_i|^p=0$，Minkowski 不等式 (1.9) 显然成立。如果 $\sum\limits_{i=1}^{n}|\xi_i+\eta_i|^p\neq 0$，取 $q>1$ 使 $1/p+1/q=1$. 由 Hölder

不等式 (1.7) 可得

$$\begin{aligned}\sum_{i=1}^{n}|\xi_i+\eta_i|^p &= \sum_{i=1}^{n}|\xi_i+\eta_i||\xi_i+\eta_i|^{p-1}\\ &\le \sum_{i=1}^{n}|\xi_i||\xi_i+\eta_i|^{p-1}+\sum_{i=1}^{n}|\eta_i||\xi_i+\eta_i|^{p-1}\\ &\le \left(\sum_{i=1}^{n}|\xi_i|^p\right)^{1/p}\left(\sum_{i=1}^{n}|\xi_i+\eta_i|^{(p-1)q}\right)^{1/q}\\ &\quad+\left(\sum_{i=1}^{n}|\eta_i|^p\right)^{1/p}\left(\sum_{i=1}^{n}|\xi_i+\eta_i|^{(p-1)q}\right)^{1/q}\\ &= \left[\left(\sum_{i=1}^{n}|\xi_i|^p\right)^{1/p}+\left(\sum_{i=1}^{n}|\eta_i|^p\right)^{1/p}\right]\left(\sum_{i=1}^{n}|\xi_i+\eta_i|^p\right)^{1/q}.\end{aligned}$$

从而可得 Minkowski 不等式 (1.9)。 证毕。

现在我们可以证明例1.24 中 $\|\boldsymbol{x}\|_p$ 对任何 $p\ge 1$ 是 $\boldsymbol{C}^n$ 上的向量范数了。事实上，由 p-范数的定义易知 $\|\boldsymbol{x}\|_p\ge 0$，并且

$$\|\boldsymbol{x}\|_p=0\Longleftrightarrow\sum_{i=1}^{n}|\xi_i|^p=0\Longleftrightarrow\xi_i=0,\ i=1,2\cdots,n\Longleftrightarrow\boldsymbol{x}=0.$$

即非负性得证。对 $\lambda\in\boldsymbol{C}$，有

$$\begin{aligned}\|\lambda\boldsymbol{x}\|_p&=\left(\sum_{i=1}^{n}|\lambda\xi_i|^p\right)^{1/p}=\left(\sum_{i=1}^{n}|\lambda|^p|\xi_i|^p\right)^{1/p}\\ &=|\lambda|\left(\sum_{i=1}^{n}|\xi_i|^p\right)^{1/p}=|\lambda|\|\boldsymbol{x}\|_p.\end{aligned}$$

即齐次性得证。最后由 Minkowski 不等式 (1.9) 可知 $\|\boldsymbol{x}\|_p$ 满足三角不等式。所以 $\|\boldsymbol{x}\|_p$ 是 $\boldsymbol{C}^n$ 上的向量范数。

1.6.2 矩阵范数及其性质

$\boldsymbol{C}^n$ 上的一个向量也是 $\boldsymbol{C}^{n\times 1}$ 上的一个矩阵，而 $\boldsymbol{C}^{n\times n}$ 上的一个矩阵也可以看作 $\boldsymbol{C}^{n^2}$ 上的一个向量。所以，$\boldsymbol{C}^{n\times n}$ 上的矩阵范数与向量范数具有一定的联系，此外矩阵有乘法运算，因而应对矩阵乘积的范数作出相应的规范。

定义 1.20 如果 $\boldsymbol{C}^{n\times n}$ 上的一个实函数 $\|\cdot\|:\boldsymbol{C}^{n\times n}\to\boldsymbol{R}$ 满足以下条件：

(1) （非负性）$\|A\|\ge 0$ 对所有的 $A\in\boldsymbol{C}^{n\times n}$ 成立，并且 $\|A\|=0$ 的充要条件是 $A=0$；

(2) （齐次性）$\|\lambda A\|=|\lambda|\|A\|$ 对所有的 $\lambda\in\boldsymbol{C}$ 与 $A\in\boldsymbol{C}^{n\times n}$ 成立；

(3) （三角不等式）$\|A+B\| \le \|A\| + \|B\|$ 对所有的 $A, B \in C^{n\times n}$ 成立；

(4) （相容性）$\|AB\| \le \|A\|\|B\|$ 对所有的 $A, B \in C^{n\times n}$ 成立，

则 $\|\cdot\|$ 称为是一个矩阵范数，而 $\|A\|$ 称为是 $C^{n\times n}$ 上矩阵 A 的范数。

例 1.25 $C^{n\times n}$ 上的 m_∞ 范数定义为

$$\|A\|_{m_\infty} = n \max_{1\le i,j\le n} |a_{ij}|, \quad A = (a_{ij}) \in C^{n\times n}.$$

证明：易验证 $\|A\|_{m_\infty}$ 满足非负性、齐次性和三角不等式，以下验证相容性。事实上，设 $A = (a_{ij})$，$B = (b_{ij}) \in C^{n\times n}$，有

$$\begin{aligned}
\|AB\|_{m_\infty} &= \left\| \left(\sum_{k=1}^{n} a_{ik} b_{kj} \right) \right\|_{m_\infty} = n \max_{1\le i,j\le n} \left| \sum_{k=1}^{n} a_{ik} b_{kj} \right| \\
&\le n \max_{1\le i,j\le n} \sum_{k=1}^{n} |a_{ik}||b_{kj}| \le n \sum_{k=1}^{n} \max_{1\le i,k\le n} |a_{ik}| \max_{1\le k,j\le n} |b_{kj}| \\
&= n \max_{1\le i,k\le n} |a_{ik}| \cdot n \max_{1\le k,j\le n} |b_{kj}| \\
&= \|A\|_{m_\infty} \|B\|_{m_\infty}.
\end{aligned}$$

例 1.26 $C^{n\times n}$ 上的 F 范数定义为

$$\|A\|_F = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2}, \quad A = (a_{ij}) \in C^{n\times n}.$$

证明：易验证 $\|A\|_{m_\infty}$ 满足非负性、齐次性和三角不等式，以下验证相容性。事实上，设 $A = (a_{ij})$，$B = (b_{ij}) \in C^{n\times n}$，由 Cauchy-Schwarz 不等式有

$$\begin{aligned}
\|AB\|_F^2 &= \left\| \left(\sum_{k=1}^{n} a_{ik} b_{kj} \right) \right\|_F^2 = \sum_{i,j=1}^{n} \left| \sum_{k=1}^{n} a_{ik} b_{kj} \right|^2 \\
&\le \sum_{i,j=1}^{n} \left(\sum_{k=1}^{n} |a_{ik}||b_{kj}| \right)^2 \le \sum_{i,j=1}^{n} \sum_{k=1}^{n} |a_{ik}|^2 \sum_{k=1}^{n} |b_{kj}|^2 \\
&= \sum_{i,k=1}^{n} |a_{ik}|^2 \sum_{k,j=1}^{n} |b_{kj}|^2 \\
&= \|A\|_F^2 \|B\|_F^2.
\end{aligned}$$

注 1.3 根据矩阵 F 范数的定义，事实上我们有 $\|A\|_F = \sqrt{\mathrm{tr}(A^H A)} = \sqrt{\mathrm{tr}(A A^H)}$.

设 $A=(a_{ij})\in \boldsymbol{C}^{m\times n}$，用$\bar{A}=(\bar{a}_{ij})$ 表示以 A 的元素的共轭复数为元素组成的矩阵，而$A^H=\left(\bar{A}\right)^T$ 为 A 的共轭转置矩阵。容易验证矩阵的共轭转置运算具有下列性质：

（1）$A^H=\overline{(A^T)}$；（2）$(A+B)^H=A^H+B^H$；（3）$(kA)^H=\bar{k}A^H$；

（4）$(AB)^H=B^HA^H$；（5）$\left(A^H\right)^H=A$；（6）如果A可逆，则$\left(A^H\right)^{-1}=(A^{-1})^H$.

定义 1.21 设$A\in \boldsymbol{C}^{n\times n}$，若$A$满足$A^HA=I$，则称$A$为酉矩阵。

由定义可知$A^H=A^{-1}$，从而有$AA^H=I$。当A为实矩阵时，酉矩阵A就是正交矩阵。酉矩阵具有如下性质。

定理 1.19 设$A,B\in \boldsymbol{C}^{n\times n}$。

（1）若A是酉矩阵，则$A^{-1},A^H,A^T,\bar{A},A^k\,(k=1,2,\cdots)$均为酉矩阵；

（2）若A,B是酉矩阵，则AB也是酉矩阵；

（3）若A是酉矩阵，则$|\det A|=1$；

（4）A是酉矩阵的充要条件是A的n个列向量是标准正交向量组；

（5）A是酉矩阵的充要条件是A的行向量是标准正交向量组；

（6）A是酉矩阵的充要条件是对任意$\alpha,\beta\in C^n$，有$(A\alpha,A\beta)=(\alpha,\beta)$；

（7）若A是酉矩阵，λ为A的特征值，则$|\lambda|=1$.

定理 1.20 (矩阵F 范数的酉不变性) 设 $U,V\in \boldsymbol{C}^{n\times n}$ 是酉矩阵，则对任意 $A\in \boldsymbol{C}^{n\times n}$ 有

$$\|UA\|_F=\|AV\|_F=\|UAV\|_F=\|A\|_F.$$

证明：根据矩阵 F 范数的定义，有

$$\|UA\|_F=\sqrt{\mathrm{tr}((UA)^HUA)}=\sqrt{\mathrm{tr}(A^HU^HUA)}=\sqrt{\mathrm{tr}(A^HA)}=\|A\|_F,$$

$$\|AV\|_F=\sqrt{\mathrm{tr}(AV(AV)^H)}=\sqrt{\mathrm{tr}(AVV^HA)}=\sqrt{\mathrm{tr}(A^HA)}=\|A\|_F.$$

从而亦有 $\|UAV\|_F=\|AV\|_F=\|A\|_F$。 证毕。

下面定理说明，经过相似变换，可以用一个矩阵范数定义另一个矩阵范数。

定理 1.21 设 $\|\cdot\|$ 是 $\boldsymbol{C}^{n\times n}$ 上一个矩阵范数，$G\in \boldsymbol{C}_n^{n\times n}$ 是一个可逆矩阵，则

$$\|A\|_G=\|G^{-1}AG\|,\quad A\in \boldsymbol{C}^{n\times n}$$

是矩阵范数。

证明：可直接验证 $\|A\|_G$ 满足矩阵范数定义的条件。 证毕。

设 $\|\cdot\|$ 是 $\boldsymbol{C}^n$ 上的向量范数，定义 $\boldsymbol{C}^{n\times n}$ 上的函数为

$$\|A\|_m=\max_{\|\boldsymbol{x}\|=1}\|A\boldsymbol{x}\|,\quad A\in \boldsymbol{C}^{n\times n}. \tag{1.10}$$

定理 1.22 式 (1.10) 中定义的函数 $\|\cdot\|_m$ 是 $C^{n\times n}$ 上的范数，并有 $\|Ax\| \le \|A\|_m\|x\|$ 对所有 $A \in C^{n\times n}$ 和所有 $x \in C^n$ 成立，及对单位矩阵 I, $\|I\|_m = 1$.

证明：(1) 显然 $\|A\|_m \ge 0$。如果 $A = 0$，则 $Ax = 0$ 对所有 $x \in C^n$ 成立，那么由(1.10)， $\|A\|_m = 0$；反之，若 $\|A\|_m = 0$，则 $\|Ax\| = 0$ 即 $Ax = 0$ 对所有 $x \in B = \{x : \|x\| = 1\}$ 成立。那么对所有 $x \in C^n$, $x \ne 0$，有 $\frac{x}{\|x\|} \in B$，则

$$0 = A\left(\frac{x}{\|x\|}\right) = \frac{Ax}{\|x\|}.$$

故 $Ax = 0$，而 $x = 0$ 时，$Ax = 0$ 自然成立。从而 $A = 0$，非负性成立。

(2) 对任意 $\lambda \in C$，齐次性成立：

$$\|\lambda A\|_m = \max_{\|x\|=1} \|\lambda Ax\| = \max_{\|x\|=1} |\lambda|\|Ax\| = |\lambda| \max_{\|x\|=1} \|Ax\| = |\lambda|\|A\|_m.$$

(3) 对任意 $A, B \in C^{n\times n}$， 三角不等式成立：

$$\begin{aligned}\|A+B\|_m &= \max_{\|x\|=1} \|(A+B)x\| = \max_{\|x\|=1} \|Ax + Bx\| \\ &\le \max_{\|x\|=1} (\|Ax\| + \|Bx\|) \le \max_{\|x\|=1} \|Ax\| + \max_{\|x\|=1} \|Bx\| \\ &= \|A\|_m + \|B\|_m.\end{aligned}$$

(4) 对任意 $A, B \in C^{n\times n}$， 相容性成立：

$$\begin{aligned}\|AB\|_m &= \max_{\|x\|=1} \|ABx\| = \max_{\|x\|=1} \frac{\|ABx\|}{\|Bx\|}\|Bx\| \\ &\le \max_{\|x\|=1} \left\|A\left(\frac{Bx}{\|Bx\|}\right)\right\| \max_{\|x\|=1} \|Bx\| \le \max_{\|y\|=1} \|Ay\| \max_{\|x\|=1} \|Bx\| \\ &= \|A\|_m\|B\|_m,\end{aligned}$$

其中，最大值不妨假定只对使 $Bx \ne 0$ 的向量 x 取的。

如果 $x \ne 0$，则有

$$\frac{\|Ax\|}{\|x\|} = \left\|A\left(\frac{x}{\|x\|}\right)\right\| \le \max_{\|x\|=1} \|Ax\| = \|A\|_m.$$

所以 $\|Ax\| \le \|A\|_m\|x\|$，当 $x = 0$ 时，不等式也成立。最后

$$\|I\|_m = \max_{\|x\|=1} \|Ix\| = \max_{\|x\|=1} \|x\| = 1.$$

证毕。

定义 1.22 称由式 (1.10) 中定义的矩阵范数为算子范数，或由向量范数 $\|\cdot\|$ 诱导的矩阵范数。

从定理 1.22 可知，一个矩阵范数 $\|\cdot\|$ 是算子范数的必要条件为：$\|I\|=1$。因为

$$\|I\|_{m_1}=n,\ \|I\|_{m_\infty}=n,\ \|I\|_F=\sqrt{n},$$

所以一般地矩阵 m_1，m_∞，F 范数均不是算子范数，或者说，它们均不是由某向量范数诱导的矩阵范数。

常用的算子范数有矩阵 1、2、∞ 范数，下面分别说明之。

例 1.27 （极大列和范数）矩阵 1 范数定义为：

$$\|A\|_1=\max_{1\le j\le n}\sum_{i=1}^{n}|a_{ij}|,\quad A=(a_{ij})\in \boldsymbol{C}^{n\times n}.$$

以下将证明 $\|A\|_1$ 是由向量 1-范数诱导的，因而它一定是矩阵范数。

事实上，记 A 的第 j 列为 a_j，即 $A=(\alpha_1,\alpha_2,\cdots,\alpha_n)$。于是 $\|A\|_1=\max\limits_{1\le j\le n}\|\alpha_j\|_1$。如果 $\boldsymbol{x}=(\xi_1,\xi_2,\cdots,\xi_n)^T\in \boldsymbol{C}^n$，则

$$\|A\boldsymbol{x}\|_1=\left\|\sum_{j=1}^{n}\xi_j\alpha_j\right\|_1\le\sum_{j=1}^{n}\|\xi_j\alpha_j\|_1=\sum_{j=1}^{n}|\xi_j|\ \|\alpha_j\|_1$$

$$\le\sum_{j=1}^{n}|\xi_j|\left(\max_{1\le j\le n}\|\alpha_j\|_1\right)=\sum_{j=1}^{n}|\xi_j|\ \|A\|_1=\|\boldsymbol{x}\|_1\|A\|_1.$$

因而，$\max\limits_{\|\boldsymbol{x}\|_1=1}\|A\boldsymbol{x}\|_1\le\|A\|_1$. 设 e_k是 $\boldsymbol{C}^n$ 中标准基的第 k 个向量，则对任意 $k=1,2,\cdots,n$，有

$$\max_{\|\boldsymbol{x}\|_1=1}\|A\boldsymbol{x}\|_1\ge\|Ae_k\|_1=\|\alpha_k\|_1,$$

所以

$$\max_{\|\boldsymbol{x}\|_1=1}\|A\boldsymbol{x}\|_1\ge\max_{1\le k\le n}\|\alpha_k\|_1=\|A\|_1.$$

故 $\|A\|_1=\max\limits_{\|\boldsymbol{x}\|_1=1}\|A\boldsymbol{x}\|_1$.

例 1.28 （极大行和范数）矩阵 ∞ 范数定义为：

$$\|A\|_\infty=\max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|,\quad A=(a_{ij})\in \boldsymbol{C}^{n\times n}.$$

以下将证明 $\|A\|_\infty$ 是由向量 ∞-范数诱导的，因而它一定是矩阵范数。

如果 $\boldsymbol{x}=(\xi_1,\xi_2,\cdots,\xi_n)^T\in C^n$，则

$$\|A\boldsymbol{x}\|_\infty=\max_{1\le i\le n}\left|\sum_{j=1}^n a_{ij}\xi_j\right|\le\max_{1\le i\le n}\sum_{j=1}^n|a_{ij}\xi_j|$$

$$\le\max_{1\le i\le n}\sum_{j=1}^n|a_{ij}|\,\|\boldsymbol{x}\|_\infty=\|A\|_\infty\|\boldsymbol{x}\|_\infty.$$

因而，$\max\limits_{\|\boldsymbol{x}\|_\infty=1}\|A\boldsymbol{x}\|_\infty\le\|A\|_\infty$. 现设

$$\max_{1\le i\le n}\sum_{j=1}^n|a_{ij}|=\sum_{j=1}^n|a_{kj}|.$$

定义向量 $\boldsymbol{z}=(\tau_1,\tau_2,\cdots,\tau_n)^T\in C^n$ 为：对 $j=1,2,\cdots,n$

$$\tau_j=\begin{cases}\dfrac{|a_{kj}|}{a_{kj}}, & \text{如 } a_{kj}\neq 0\\ 1, & \text{如 } a_{kj}=0.\end{cases}$$

于是 $\|\boldsymbol{z}\|_\infty=1$，且对所有 $j=1,2,\cdots,n$，有 $a_{kj}\tau_j=|a_{kj}|$，并且

$$\max_{\|\boldsymbol{x}\|_\infty=1}\|A\boldsymbol{x}\|_\infty\ge\|A\boldsymbol{z}\|_\infty=\max_{1\le i\le n}\left|\sum_{j=1}^n a_{ij}\tau_j\right|\ge\left|\sum_{j=1}^n a_{kj}\tau_j\right|$$

$$=\sum_{j=1}^n|a_{kj}|=\max_{1\le i\le n}\sum_{j=1}^n|a_{ij}|=\|A\|_\infty.$$

故 $\|A\|_\infty=\max\limits_{\|\boldsymbol{x}\|_\infty=1}\|A\boldsymbol{x}\|_\infty$.

例 1.29 （谱范数）矩阵 2 范数定义为

$$\|A\|_2=\max\{\sqrt{\lambda}:\lambda\text{是 }A^HA\text{ 的特征值}\}.$$

以下将证明 $\|A\|_2$ 是由向量 2-范数诱导的，因而它一定是矩阵范数。

定义 1.23 设$A\in C^{n\times n}$，如果$A^H=A$，则称A为Hermite矩阵；如果$A^H=-A$，则称A为反Hermite矩阵。

显然，Hermite矩阵主对角线元素全为实数，反Hermite矩阵的主对角元全为零或纯虚数。实对称矩阵是Hermite矩阵的特例。

定义 1.24 设$A\in C^{n\times n}$，且$A^HA=AA^H$，则称A为正规矩阵。

容易验证，对角矩阵、实对称矩阵、实反对称矩阵$(A^T=-A)$、Hermite矩阵、反Hermite矩阵、正交矩阵以及酉矩阵等，都是正规矩阵。当然，正规矩阵并非只是

这些常用的矩阵，例如

$$A=\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}$$

是一个正规矩阵，但它不属于上述矩阵的任何一种。

因为 A^HA 是 Hermite 矩阵，所以它的特征值 $\lambda_1,\lambda_2,\cdots,\lambda_n$ 均非负，根据第2.3节的定理2.8，存在酉矩阵 U 使得

$$U^HA^HAU=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_n).$$

如果 $\boldsymbol{x}=(\xi_1,\xi_2,\cdots,\xi_n)^T\in\boldsymbol{C}^n$，记 $U^H\boldsymbol{x}=\boldsymbol{y}=(\eta_1,\eta_2,\cdots,\eta_n)^T\in\boldsymbol{C}^n$。则

$$\begin{aligned}\|A\boldsymbol{x}\|_2^2&=(A\boldsymbol{x})^H(A\boldsymbol{x})=\boldsymbol{x}^HA^HA\boldsymbol{x}=\boldsymbol{y}^HU^HA^HAU\boldsymbol{y}\\&=\boldsymbol{y}^H\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_n)\boldsymbol{y}=\sum_{i=1}^{n}\lambda_i|\eta_i|^2\\&\le\|A\|_2^2\sum_{i=1}^{n}|\eta_i|^2=\|A\|_2^2\|\boldsymbol{y}\|_2^2=\|A\|_2^2\|\boldsymbol{x}\|_2^2.\end{aligned}$$

因而 $\max\limits_{\|\boldsymbol{x}\|_2=1}\|A\boldsymbol{x}\|_2\le\|A\|_2$. 设 $\max\{\lambda_1,\lambda_2,\cdots,\lambda_n\}=\lambda_k$. 定义向量 $\mathbf{z}=Ue_k$，其中 e_k 为 $\boldsymbol{C}^n$ 中标准基的第 k 个向量. 则 $\|\mathbf{z}\|_2=\|e_k\|_2=1$，并且

$$\begin{aligned}\max_{\|\boldsymbol{x}\|_2=1}\|A\boldsymbol{x}\|_2&\ge\|A\boldsymbol{z}\|_2=\sqrt{(A\boldsymbol{z})^H(A\boldsymbol{z})}\\&=\sqrt{e_k^H\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_n)e_k}=\sqrt{\lambda_k}=\|A\|_2.\end{aligned}$$

故 $\|A\|_2=\max\limits_{\|\boldsymbol{x}\|_2=1}\|A\boldsymbol{x}\|_2$.

定理 1.23 （1）（谱范数的酉不变性）设 U,V 是酉矩阵，则

$$\|A\|_2=\|UA\|_2=\|AV\|_2=\|UAV\|_2.$$

（2）如果 A 是正规矩阵，则 $\|A\|_2=\max\{|\lambda|:\lambda是\ A\ 的特征值\}$.

证明：（1）因为 $(UA)^HUA=A^HU^HUA=A^HA$，所以 $\|A\|_2=\|UA\|_2$. 又 $(AV)^HAV$ 与 $AV(AV)^H$ 的特征值相同，而 $AV(AV)^H=AA^H$，且 AA^H 与 A^HA 的特征值相同，因此 $\|A\|_2=\|AV\|_2$. 从而亦有 $\|A\|_2=\|UAV\|_2$

（2）如果 A 是正规矩阵，则根据第2.3节的定理2.8，存在酉矩阵 U 使得

$$U^HAU=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_n),$$

其中 $\lambda_1,\lambda_2,\cdots,\lambda_n$ 是 A 的特征值。那么

$$\begin{aligned}A^HA&=U\mathrm{diag}(\bar{\lambda}_1,\bar{\lambda}_2,\cdots,\bar{\lambda}_n)U^HU\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_n)U^H\\&=U\mathrm{diag}(|\lambda_1|^2,|\lambda_2|^2,\cdots,|\lambda_n|^2)U^H.\end{aligned}$$

因此 A^HA 的特征值为 $|\lambda_1|^2,|\lambda_2|^2,\cdots,|\lambda_n|^2$，所以 $\|A\|_2=\max\limits_{1\le i\le n}|\lambda_i|$。 证毕。

1.6.3 向量范数、矩阵范数的相容性

定义 1.25 设 $\|\cdot\|_\alpha$ 是 C^n 上的一个向量范数，$\|\cdot\|_m$ 是 $C^{n\times n}$ 上的一个矩阵范数。如果对任意 $\boldsymbol{x}\in C^n$ 及 $A\in C^{n\times n}$ 有

$$\|A\boldsymbol{x}\|_\alpha \le \|A\|_m\|\boldsymbol{x}\|_\alpha,$$

则称向量范数$\|\cdot\|_\alpha$ 与矩阵范数 $\|\cdot\|_m$ 是相容的。

例 1.30 向量 1-范数与矩阵 m_1 范数相容。

证明：设 $\boldsymbol{x}=(\xi_1,\xi_2,\cdots,\xi_n)^T\in C^n$，$A=(a_{ij})\in C^{n\times n}$，则有

$$\begin{aligned}\|A\boldsymbol{x}\|_1 &= \sum_{i=1}^n\left|\sum_{j=1}^n a_{ij}\xi_j\right| \le \sum_{i=1}^n\sum_{j=1}^n(|a_{ij}||\xi_j|)\\ &\le \sum_{i=1}^n\sum_{j=1}^n\left(|a_{ij}|\sum_{j=1}^n|\xi_j|\right)=\left(\sum_{i=1}^n\sum_{j=1}^n|a_{ij}|\right)\left(\sum_{j=1}^n|\xi_j|\right)\\ &=\|A\|_{m_1}\|\boldsymbol{x}\|_1.\end{aligned}$$

例 1.31 向量 2-范数与矩阵 F 范数相容。

证明：设 $\boldsymbol{x}=(\xi_1,\xi_2,\cdots,\xi_n)^T\in C^n$，$A=(a_{ij})\in C^{n\times n}$，由 Cauchy-Schwarz 不等式有

$$\begin{aligned}\|A\boldsymbol{x}\|_2 &= \left(\sum_{i=1}^n\left|\sum_{j=1}^n a_{ij}\xi_j\right|^2\right)^{1/2} \le \left(\sum_{i=1}^n\left(\sum_{j=1}^n(|a_{ij}||\xi_j|)\right)^2\right)^{1/2}\\ &\le \left(\sum_{i=1}^n\sum_{j=1}^n|a_{ij}|^2\sum_{j=1}^n|\xi_j|^2\right)^{1/2}\\ &\le \left(\sum_{i=1}^n\sum_{j=1}^n|a_{ij}|^2\right)^{1/2}\left(\sum_{j=1}^n|\xi_j|^2\right)^{1/2}\\ &=\|A\|_F\|\boldsymbol{x}\|_2.\end{aligned}$$

例 1.32 向量 1-范数、2-范数、∞-范数均与矩阵 m_∞ 范数相容。

证明：设 $\boldsymbol{x}=(\xi_1,\xi_2,\cdots,\xi_n)^T\in C^n$，$A=(a_{ij})\in C^{n\times n}$，则有

$$\|A\boldsymbol{x}\|_1 = \sum_{i=1}^n\left|\sum_{j=1}^n a_{ij}\xi_j\right| \le \sum_{i=1}^n\sum_{j=1}^n|a_{ij}||\xi_j|$$

$$\leq \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\max_{1\leq i,j\leq n}|a_{ij}|\cdot|\xi_j|\right)=n\max_{1\leq i,j\leq n}|a_{ij}|\sum_{j=1}^{n}|\xi_j|$$

$$=\|A\|_{m_\infty}\|\boldsymbol{x}\|_1.$$

而

$$\|A\boldsymbol{x}\|_2=\left(\sum_{i=1}^{n}\left|\sum_{j=1}^{n}a_{ij}\xi_j\right|^2\right)^{1/2}\leq\left(\sum_{i=1}^{n}\left(\sum_{j=1}^{n}|a_{ij}||\xi_j|\right)^2\right)^{1/2}$$

$$\leq\left(\sum_{i=1}^{n}\left(\sum_{j=1}^{n}|a_{ij}|^2\sum_{j=1}^{n}|\xi_j|^2\right)\right)^{1/2}$$

$$=\left(\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2\right)^{1/2}\left(\sum_{j=1}^{n}|\xi_j|^2\right)^{1/2}\leq n\max_{1\leq i,j\leq n}|a_{ij}|\left(\sum_{j=1}^{n}|\xi_j|^2\right)^{1/2}$$

$$=\|A\|_{m_\infty}\|\boldsymbol{x}\|_2.$$

以及

$$\|A\boldsymbol{x}\|_\infty=\max_{1\leq i\leq n}\left|\sum_{j=1}^{n}a_{ij}\xi_j\right|\leq\max_{1\leq i\leq n}\sum_{j=1}^{n}|a_{ij}||\xi_j|$$

$$\leq n\max_{1\leq i,j\leq n}|a_{ij}|\max_{1\leq j\leq n}|\xi_j|$$

$$=\|A\|_{m_\infty}\|\boldsymbol{x}\|_\infty.$$

由定理 1.22 可知，任何向量范数都存在与之相容的矩阵范数（算子范数），反之我们有以下结论。

定理 1.24 设 $\|\cdot\|_m$ 是 $\boldsymbol{C}^{n\times n}$ 上的一个矩阵范数，则在 $\boldsymbol{C}^n$ 上存在与之相容的向量范数。

证明：对 $\boldsymbol{x}\in\boldsymbol{C}^n$，记 $\boldsymbol{X}$ 是由 n 个 $\boldsymbol{x}$ 为列向量组成的矩阵，即 $\boldsymbol{X}=(\boldsymbol{x},\boldsymbol{x},\cdots,\boldsymbol{x})^T\in\boldsymbol{C}^{n\times n}$。定义 $\boldsymbol{C}^n$ 上函数 $\|\boldsymbol{x}\|=\|\boldsymbol{X}\|_m$。易证 $\|\boldsymbol{x}\|$ 是 $\boldsymbol{C}^n$ 上的范数，并且对任意 $A\in\boldsymbol{C}^{n\times n}$，有

$$\|A\boldsymbol{x}\|=\|(A\boldsymbol{x},A\boldsymbol{x},\cdots,A\boldsymbol{x})^T\|_m=\|A\boldsymbol{X}\|_m\leq\|A\|_m\|\boldsymbol{X}\|_m=\|A\|_m\|\boldsymbol{x}\|,$$

即向量范数 $\|\cdot\|$ 与给定矩阵范数 $\|\cdot\|_m$ 相容。 证毕。

1.7 矩阵的谱半径，条件数

作为矩阵范数的一个重要应用，它可以给出矩阵特征值（谱）的范围。本节讨论矩阵的谱半径及其在研究矩阵级数收敛性中的应用。

1.7.1 矩阵的谱半径

定义 1.26 矩阵 $A\in C^{n\times n}$ 的谱半径 $\rho(A)$ 定义为

$$\rho(A)=\max\{|\lambda|:\lambda \text{ 是 } A \text{ 的特征值}\}.$$

定理 1.25 设 $\|\cdot\|$ 是 $C^{n\times n}$ 上一个矩阵范数。则对任意 $A\in C^{n\times n}$，有 $\rho(A)\le\|A\|$.

证明：根据谱半径的定义，至少有 A 的一个特征值 λ_* 可使得 $|\lambda_*|=\rho(A)$。设 $A\boldsymbol{x}=\lambda_*\boldsymbol{x}$，$\boldsymbol{x}\neq 0$。记 $\boldsymbol{X}=(\boldsymbol{x},\boldsymbol{x},\cdots,\boldsymbol{x})\in C^{n\times n}$，那么 $A\boldsymbol{X}=(A\boldsymbol{x},A\boldsymbol{x},\cdots,A\boldsymbol{x})=(\lambda_*\boldsymbol{x},\lambda_*\boldsymbol{x},\cdots,\lambda_*\boldsymbol{x})=\lambda_*\boldsymbol{X}$。则

$$|\lambda_*|\,\|\boldsymbol{X}\|=\|\lambda_*\boldsymbol{X}\|=\|A\boldsymbol{X}\|\le\|A\|\|\boldsymbol{X}\|,$$

因此 $\rho(A)=|\lambda_*|\le\|A\|$。 证毕。

矩阵的谱半径是 $C^{n\times n}$ 上的一个函数，但它不是矩阵范数。如设

$$A=\begin{pmatrix}0&1\\0&0\end{pmatrix},\ B=\begin{pmatrix}0&0\\1&0\end{pmatrix},\ E=\begin{pmatrix}0&1\\1&0\end{pmatrix},\ F=\begin{pmatrix}1&1\\0&1\end{pmatrix},$$

则 $A\neq 0$，但 $\rho(A)=0$；$\rho(A+B)=1>0=\rho(A)+\rho(B)$；$\rho(EF)=(1+\sqrt{5})/2>1=\rho(E)\rho(F)$。

例 1.33 设 $A\in C^{n\times n}$，证明：（1）$\|A\|_2\le\max\{\|A\|_\infty,\|A\|_1\}$；（2）当 A 为正规矩阵时，$\|A\|_2=\rho(A)$。

证明：（1）我们有

$$\begin{aligned}\|A\|_2&=\sqrt{\rho(A^HA)}\le\sqrt{\|A^HA\|_1}\le\sqrt{\|A^H\|_1\|A\|_1}\\&=\sqrt{\|A\|_\infty\|A\|_1}\le\max\{\|A\|_\infty,\|A\|_1\}\end{aligned}$$

（2）根据定理 1.23及谱半径的定义，对正规矩阵 A，有 $\|A\|_2=\rho(A)$。

虽然谱半径本身不是矩阵范数，但是，对每个固定的矩阵 $A\in C^{n\times n}$，它是 A 的所有矩阵范数值的最大下界，以下定理可说明该结论。

定理 1.26 设 $A\in C^{n\times n}$，则对给定的 $\varepsilon>0$，存在一个矩阵范数 $\|\cdot\|$ 使得 $\|A\|\le\rho(A)+\varepsilon$.

证明：根据第 2.2 节定理 2.6， A 相似于 Jordan 标准形矩阵J，即存在可逆矩阵 $P \in C_n^{n\times n}$ 使得

$$P^{-1}AP = J = \begin{pmatrix} \lambda_1 & \delta_1 & & \\ & \lambda_2 & \ddots & \\ & & \ddots & \delta_{n-1} \\ & & & \lambda_n \end{pmatrix}, \quad \delta_i = 0 \text{ 或 } 1$$

其中 $\lambda_1, \lambda_2, \cdots, \lambda_n$ 是 A 的特征值。 现令 $D = \mathrm{diag}(1, \varepsilon, \varepsilon^2, \cdots, \varepsilon^{n-1})$, 则有

$$D^{-1}P^{-1}APD = D^{-1}JD = \begin{pmatrix} \lambda_1 & \varepsilon\delta_1 & & \\ & \lambda_2 & \ddots & \\ & & \ddots & \varepsilon\delta_{n-1} \\ & & & \lambda_n \end{pmatrix}$$

于是

$$\|D^{-1}P^{-1}APD\|_1 \le \max_j(|\lambda_j| + \varepsilon) = \rho(A) + \varepsilon$$

定义

$$\|B\| = \|D^{-1}P^{-1}APD\|_1, \quad B \in \boldsymbol{C}^{n\times n}$$

根据定理 1.21， $\|\cdot\|$ 是 $\boldsymbol{C}^{n\times n}$ 上的矩阵范数，且有

$$\|A\| = \|D^{-1}P^{-1}APD\|_1 \le \rho(A) + \varepsilon.$$

证毕。

1.7.2 矩阵序列及级数中的应用

可以类似于向量序列，用矩阵范数来讨论矩阵序列的收敛性以及矩阵级数的收敛性。 由于 $\boldsymbol{C}^{n\times n}$ 上的矩阵范数也是 $\boldsymbol{C}^{n^2}$ 上的向量范数， 所以关于向量范数的一些性质对矩阵范数也是适用的，不再赘述。

定义 1.27 称满足 $\lim\limits_{k\to\infty} A^k = 0$ 的矩阵 $A \in \boldsymbol{C}^{n\times n}$ 为收敛矩阵。

注 1.4 条件 $\lim\limits_{k\to\infty} A^k = 0$ 等价于 $\lim\limits_{k\to\infty} \|A^k\| = 0$ 对 $\boldsymbol{C}^{n\times n}$ 上任意矩阵范数 $\|\cdot\|$ 成立。

定理 1.27 设 $A \in \boldsymbol{C}^{n\times n}$， 则 A 为收敛矩阵的充要条件为 $\rho(A) < 1$.

证明：如果 A 为收敛矩阵，则对 $\boldsymbol{C}^{n\times n}$ 上任一矩阵范数 $\|\cdot\|$ 有 $\lim\limits_{k\to\infty} \|A^k\| = 0$。因为

$$(\rho(A))^k = \rho(A^k) \le \|A^k\| \to 0, \quad k \to \infty,$$

所以必有 $\rho(A)<1$.

反之，如果 $\rho(A)<1$，则有 $\varepsilon>0$ 使得 $\rho(A)+\varepsilon<1$。根据定理 1.26，存在某个矩阵范数 $\|\cdot\|$ 使得 $\|A\|\le\rho(A)+\varepsilon<1$。那么

$$\|A^k\|\le\|A\|^k\le(\rho(A)+\varepsilon)^k\to 0.$$

所以 A 是收敛矩阵。 证毕。

定理 1.28 设 $A\in C^{n\times n}$，若有 $C^{n\times n}$ 上的矩阵范数 $\|\cdot\|$ 使得 $\|A\|<1$，则 A 为收敛矩阵。

定理 1.29 设 $\|\cdot\|$ 是 $C^{n\times n}$ 上的矩阵范数，则对所有 $A\in C^{n\times n}$，有

$$\rho(A)=\lim_{k\to\infty}\|A^k\|^{1/k}.$$

证明：因为 $(\rho(A))^k=\rho(A^k)\le\|A^k\|$，所以对所有 $k=1,2,\cdot$，有 $\rho(A)\le\|A^k\|^{1/k}$. 任意给定 $\varepsilon>0$，则矩阵 $\widetilde{A}=[\rho(A)+\varepsilon]^{-1}A$ 的谱半径严格小于 1，因而 $\widetilde{A}$ 是收敛矩阵，即 $k\to\infty$ 时，$\|\widetilde{A}^k\|\to 0$。那么存在某个正整数 N 使得 $k\ge N$ 时，$\|\widetilde{A}^k\|(<1$，即 $\|A^k\|\le[\rho(A)+\varepsilon]^k$。因此

$$\rho(A)\le\|A^k\|^{1/k}\le\rho(A)+\varepsilon.$$

由 $\varepsilon>0$ 的任意性，得出 $\lim_{k\to\infty}\|A^k\|^{1/k}$ 存在且等于 $\rho(A)$。 证毕。

例 1.34 判断下列矩阵是否为收敛矩阵：

$$A_1=\frac{1}{5}\begin{pmatrix}2&1\\3&2\end{pmatrix};\quad A_2=\frac{1}{3}\begin{pmatrix}1&-1\\-1&3\end{pmatrix};\quad A_3=\begin{pmatrix}0.1&0.5&-0.3\\-0.4&0.4&0.1\\0.3&-0.2&0.3\end{pmatrix}.$$

解：(1) 由于 A_1 的特征值为 $\lambda_1=0.8$，$\lambda_2=0$，所以 $\rho(A_1)=0.8<1$. 从而 A_1 是收敛矩阵。

(2) 由于 A_2 的特征值为 $\lambda_1=(2+\sqrt{2})/3$，$\lambda_2=(2-\sqrt{2})/3$，所以 $\rho(A_2)=(2+\sqrt{2})/3>1$. 从而 A_2 不是收敛矩阵。

(3) 由于 $\|A_3\|_\infty=0.9<1$，所以 A_3 是收敛矩阵。

定义 1.28 设 $\{A^{(k)}\}_{k=1}^{\infty}$ 是 $C^{n\times n}$ 上的矩阵序列，它们的无穷和式

$$A^{(1)}+A^{(2)}+\cdots+A^{(k)}+\cdots$$

称为矩阵级数，记为 $\sum_{k=1}^{\infty}A^{(k)}$. 对任一正整数 N，称 $S^{(N)}=\sum_{k=1}^{N}A^{(k)}$ 为矩阵级数的部分和。如果部分和序列 $\{S^{(N)}\}$ 收敛于矩阵 S，则称矩阵级数收敛，其和为 S，记为 $\sum_{k=1}^{\infty}A^{(k)}=S$。不收敛的级数称为发散级数。

从定义可知：若 $A^{(k)}=(a_{ij}^{(k)}),\ S=(s_{ij})\in C^{n\times n}$，则 $\sum\limits_{k=1}^{\infty}A^{(k)}=S$ 等价于

$$\sum_{k=1}^{\infty}a_{ij}^{(k)}=s_{ij},\quad i,j=1,2,\cdots,n.$$

定义 1.29 设 $A^{(k)}=(a_{ij}^{(k)})\in C^{n\times n}$，$(k=1,2,\cdots)$。若 n^2 个数项级数

$$\sum_{k=1}^{\infty}a_{ij}^{(k)},\quad i,j=1,2,\cdots,n.$$

均绝对收敛，则称矩阵级数 $\sum\limits_{k=1}^{\infty}A^{(k)}$ 绝对收敛。

从定义可知绝对收敛的矩阵级数一定是收敛的矩阵级数。矩阵级数的绝对收敛性也可以通过矩阵范数将之化为正项级数的收敛性。

定理 1.30 矩阵级数 $\sum\limits_{k=1}^{\infty}A^{(k)}$ 绝对收敛的充要条件是正项级数 $\sum\limits_{k=1}^{\infty}\|A^{(k)}\|$ 收敛，其中 $\|\cdot\|$ 是 $C^{n\times n}$ 上的任一矩阵范数。

证明：由于 $C^{n\times n}$ 上的所有矩阵范数是等价的，所以只要证明 $\sum\limits_{k=1}^{\infty}A^{(k)}$ 绝对收敛的充要条件是 $\sum\limits_{k=1}^{\infty}\|A^{(k)}\|_{m_1}$ 收敛即可。

设 $A^{(k)}=(a_{ij}^{(k)})\in C^{n\times n}$，$(k=1,2,\cdots)$。若 $\sum\limits_{k=1}^{\infty}A^{(k)}$ 绝对收敛，即 $\sum\limits_{k=1}^{\infty}|a_{ij}^{(k)}|$ 收敛 $(i,j=1,2,\cdots,n)$，从而其部分和有界，设

$$\sum_{k=1}^{N}|a_{ij}^{(k)}|\le M,\quad i,j=1,2,\cdots,n.$$

因此

$$\begin{aligned}\sum_{k=1}^{N}\|A^{(k)}\|_{m_1}&=\sum_{k=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}^{(k)}|\\&=\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{N}|a_{ij}^{(k)}|\le n^2M.\end{aligned}$$

所以 $\sum\limits_{k=1}^{\infty}\|A^{(k)}\|_{m_1}$ 收敛。

反之，若 $\sum\limits_{k=1}^{\infty}\|A^{(k)}\|_{m_1}$ 收敛，因为

$$|a_{ij}^{(k)}|\le\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}^{(k)}|=\|A^{(k)}\|_{m_1},\quad i,j=1,2,\cdots,n,$$

所以 $\sum\limits_{k=1}^{\infty}|a_{ij}^{(k)}|$ 收敛，即 $\sum\limits_{k=1}^{\infty}A^{(k)}$ 绝对收敛。 证毕。

下面讨论一类特殊的矩阵级数——矩阵幂级数，它在后面矩阵函数的章节中有重要的作用。

定义 1.30 设 $A\in\boldsymbol{C}^{n\times n}$，$a_k\in\boldsymbol{C}$ $(k=0,1,2,\cdots)$，称矩阵级数 $\sum\limits_{k=0}^{\infty}a_kA^k$ 为矩阵 A 的幂级数。

定理 1.31 设复变量幂级数 $\sum\limits_{k=0}^{\infty}a_kz^k$ 的收敛半径为 R，$A\in\boldsymbol{C}^{n\times n}$，则

(1) 当 $\rho(A)<R$ 时，矩阵幂级数 $\sum\limits_{k=0}^{\infty}a_kA^k$ 绝对收敛；

(2) 当 $\rho(A)>R$ 时，矩阵幂级数 $\sum\limits_{k=0}^{\infty}a_kA^k$ 发散。

证明：(1) 当 $\rho(A)<R$ 时，存在 $\varepsilon>0$ 使得 $\rho(A)+\varepsilon<R$，从而由定理 1.26，存在 $\boldsymbol{C}^{n\times n}$ 上的矩阵范数 $\|\cdot\|$ 使得 $\|A\|\leq\rho(A)+\varepsilon<R$。那么有

$$\|a_kA^k\|\leq|a_k|\|A\|^k\leq|a_k|(\rho(A)+\varepsilon)^k.$$

由于级数 $\sum\limits_{k=0}^{\infty}|a_k|(\rho(A)+\varepsilon)^k$ 收敛，所以 $\sum\limits_{k=0}^{\infty}\|a_kA^k\|$ 收敛，故由定理 1.30，矩阵幂级数 $\sum\limits_{k=0}^{\infty}a_kA^k$ 绝对收敛。

(2) 当 $\rho(A)<R$ 时，设 A 的特征值为 $\lambda_1,\lambda_2,\cdots,\lambda_n$，则一定有某个特征值的模大于 R，不妨设 $|\lambda_1|>R$。设 A 的 Jordan 矩阵为 J，由 Jordan 定理，存在可逆矩阵 P 使得

$$P^{-1}AP=J=\begin{pmatrix}\lambda_1 & \delta_1 & & \\ & \lambda_2 & \ddots & \\ & & \ddots & \delta_{n-1}\\ & & & \lambda_n\end{pmatrix},\quad \delta_i=0\ \text{或}\ 1.$$

由于矩阵级数 $\sum\limits_{k=0}^{\infty}a_kJ^k$ 对角线元素为 $\sum\limits_{k=0}^{\infty}a_k\lambda_j^k$ $(j=1,2,\cdots,n)$，其中有一项 $\sum\limits_{k=0}^{\infty}a_k\lambda_1^k$ 是发散的，所以矩阵级数 $\sum\limits_{k=0}^{\infty}a_kJ^k$ 发散。而 $\sum\limits_{k=0}^{\infty}a_kJ^k=P^{-1}\left(\sum\limits_{k=0}^{\infty}a_kA^k\right)P$，故矩阵级数 $\sum\limits_{k=0}^{\infty}a_kA^k$ 发散。 证毕。

定理 1.32 设复变量幂级数 $\sum\limits_{k=0}^{\infty}a_kz^k$ 的收敛半径为 R，$A\in\boldsymbol{C}^{n\times n}$。若存在 $\boldsymbol{C}^{n\times n}$ 上的矩阵范数 $\|\cdot\|$ 使得 $\|A\|<R$，则矩阵幂级数 $\sum\limits_{k=0}^{\infty}a_kA^k$ 绝对收敛。

例 1.35 判断以下矩阵幂级数的敛散性：

$$(1)\sum_{k=0}^{\infty}\frac{k}{5^k}\begin{pmatrix}2&1\\3&2\end{pmatrix}^k;\quad (2)\sum_{k=0}^{\infty}\frac{k+1}{3^k}\begin{pmatrix}1&-1\\-1&3\end{pmatrix}^k;$$

$$(3)\sum_{k=0}^{\infty}k(k+1)\begin{pmatrix}0.1&0.5&-0.3\\-0.4&0.4&0.1\\0.3&-0.2&0.3\end{pmatrix}^k.$$

解：设

$$A=\frac{1}{5}\begin{pmatrix}2&1\\3&2\end{pmatrix};\quad B=\frac{1}{3}\begin{pmatrix}1&-1\\-1&3\end{pmatrix};\quad D=\begin{pmatrix}0.1&0.5&-0.3\\-0.4&0.4&0.1\\0.3&-0.2&0.3\end{pmatrix}.$$

(1) 因为 $\rho(A)=0.8<1$，而幂级数 $\sum\limits_{k=0}^{\infty}kz^k$ 的收敛半径为 1，所以矩阵幂级数 $\sum\limits_{k=0}^{\infty}kA^k$ 绝对收敛。

(2) 因为 $\rho(B)=(2+\sqrt{2})/3>1$，而幂级数 $\sum\limits_{k=0}^{\infty}(k+1)z^k$ 的收敛半径为 1，所以矩阵幂级数 $\sum\limits_{k=0}^{\infty}(k+1)B^k$ 发散。

(3) 因为 $\|D\|_\infty=0.9<1$，而幂级数 $\sum\limits_{k=0}^{\infty}k(k+1)z^k$ 的收敛半径为 1，所以矩阵幂级数 $\sum\limits_{k=0}^{\infty}k(k+1)D^k$ 绝对收敛。

定理 1.33 设 $A\in\boldsymbol{C}^{n\times n}$，则矩阵幂级数 $\sum\limits_{k=0}^{\infty}A^k$ 收敛的充要条件是 $\rho(A)<1$，且在收敛时，其和为 $(I-A)^{-1}$。

证明：如果 $\rho(A)<1$，则 A 是收敛矩阵。因为幂级数 $\sum\limits_{k=0}^{\infty}z^k$ 的收敛半径为 1，所以矩阵幂级数 $\sum\limits_{k=0}^{\infty}A^k$ 收敛于某个矩阵 S。而当 $N\to\infty$ 时，由于

$$(I-A)\sum_{k=0}^{N}A^k=\sum_{k=0}^{N}A^k-\sum_{k=0}^{N}A^{k+1}=I-A^{N+1}\to I,$$

所以得出 $S=(I-A)^{-1}$。

反之，如果 $\sum\limits_{k=0}^{\infty}A^k$ 收敛，设其和为 S，记其部分和为 $S^{(N)}=\sum\limits_{k=0}^{N}A^k$。由于

$$\lim_{N\to\infty}A^N=\lim_{N\to\infty}(S^{(N)}-S^{(N-1)})=\lim_{N\to\infty}S^{(N)}-\lim_{N\to\infty}S^{(N-1)}=S-S=0,$$

所以 A 是收敛矩阵，故由定理 1.27，$\rho(A)<1$。 证毕。

定理 1.34 设 $A \in C^{n\times n}$，如果存在矩阵范数 $\|\cdot\|$ 使得 $\|A\| < 1$，则 $I - A$ 是可逆矩阵，且 $(I-A)^{-1} = \sum\limits_{k=0}^{\infty} A^k$，从而 $\|(I-A)^{-1}\| \leq (1-\|A\|)^{-1}$。

1.7.3 矩阵的条件数

如果给定了可逆矩阵 $A \in C^{n\times n}$，可以计算出其逆矩阵 A^{-1}。但是如果计算是在数字计算机中进行的，就难免要出现舍入误差和舍位误差。此外，矩阵 A 的某些元素可能是通过实验或观测获得的结果，也难免会出现误差。也就是说，若要计算 A^{-1} 时，实际计算的可能是 $(A+\delta A)^{-1}$，其中 δA 是微小的误差矩阵。这些数据误差对计算 A^{-1} 时出现的误差影响多大呢？在估计这个影响的大小中起重要作用的一个量就是矩阵 A 的条件数。

定义 1.31 设 $A \in C_n^{n\times n}$，$\|\cdot\|$ 是 $C^{n\times n}$ 上的一个矩阵范数。矩阵 A 的条件数定义为

$$\mathrm{cond} = \|A\|\,\|A^{-1}\|.$$

例 1.36 设 $A \in C_n^{n\times n}$ 是一个正规矩阵，则

$$\mathrm{cond}_2(A) = \|A\|_2\|A^{-1}\|_2 = \frac{\max\{|\lambda| : \lambda \text{ 是 } A \text{ 的特征值}\}}{\min\{|\lambda| : \lambda \text{ 是 } A \text{ 的特征值}\}}$$

证明：设 A 的特征值为 $\lambda_1, \lambda_2, \cdots, \lambda_n$，不妨设

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n| > 0.$$

则 $\|A\|_2 = \rho(A) = |\lambda_1|$。而 A^{-1} 亦为正规矩阵，且其特征值为 $1/\lambda_j$ $(j = 1, 2, \cdots, n)$，满足

$$\frac{1}{|\lambda_n|} \geq \frac{1}{|\lambda_{n-1}|} \geq \cdots \geq \frac{1}{|\lambda_1|}.$$

所以 $\|A^{-1}\|_2 = \rho(A^{-1}) = \frac{1}{|\lambda_n|}$。从而

$$\mathrm{cond}_2(A) = \|A\|_2\|A^{-1}\|_2 = \frac{|\lambda_1|}{|\lambda_n|}.$$

例 1.37 设 $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 3 \\ 3 & 4 & 5 \end{pmatrix}$，求 A 的条件数 $\mathrm{cond}_\infty(A) = \|A\|_\infty\|A^{-1}\|_\infty$。

解：因为 $A^{-1} = \frac{1}{2}\begin{pmatrix} -3 & -2 & 3 \\ 1 & 4 & -3 \\ 1 & -2 & 1 \end{pmatrix}$，所以 A 的条件数为

$$\mathrm{cond}_\infty(A) = \|A\|_\infty\|A^{-1}\|_\infty = 12 \times 4 = 48.$$

1.7.4 在误差估计中的应用

首先我们讨论矩阵 A 的数据误差对在求矩阵 A 的逆中产生的相对误差的影响。

定理 1.35 设 $A \in C_n^{n\times n}$ 是一个可逆矩阵，$\delta A \in C^{n\times n}$ 是一个矩阵，$\|\cdot\|$ 是 $C^{n\times n}$ 上的一个矩阵范数。如果 $\|A^{-1}\delta A\| < 1$，则 $A+\delta A$ 可逆，且有

$$\frac{\|A^{-1}-(A+\delta A)^{-1}\|}{\|A^{-1}\|} \leq \frac{\|A^{-1}\delta A\|}{1-\|A^{-1}\delta A\|}. \tag{1.11}$$

证明：如果 $\|A^{-1}\delta A\| < 1$，由推论 1.34，矩阵 $A+\delta A = A(I+A^{-1}\delta A)$ 可逆，且

$$(A+\delta A)^{-1} = (I+A^{-1}\delta A)^{-1}A^{-1} = \sum_{k=0}^{\infty}(-1)^k(A^{-1}\delta A)^k A^{-1}.$$

所以

$$\|A^{-1}-(A+\delta A)^{-1}\| = \left\|\sum_{k=1}^{\infty}(-1)^{k+1}(A^{-1}\delta A)^k A^{-1}\right\|$$

$$\leq \sum_{k=1}^{\infty}\|A^{-1}\delta A\|^k\|A^{-1}\| = \frac{\|A^{-1}\delta A\|}{1-\|A^{-1}\delta A\|}\|A^{-1}\|.$$

从而

$$\frac{\|A^{-1}-(A+\delta A)^{-1}\|}{\|A^{-1}\|} \leq \frac{\|A^{-1}\delta A\|}{1-\|A^{-1}\delta A\|}.$$

定理 1.36 设 $A \in C_n^{n\times n}$，$\delta A \in C^{n\times n}$。如果存在 $C^{n\times n}$ 上的一个矩阵范数 $\|\cdot\|$ 使得 $\|A^{-1}\|\,\|\delta A\| < 1$，则有

$$\frac{\|A^{-1}-(A+\delta A)^{-1}\|}{\|A^{-1}\|} \leq \frac{\mathrm{cond}(A)}{1-\mathrm{cond}(A)(\|\delta A\|/\|A\|)}\frac{\|\delta A\|}{\|A\|}. \tag{1.12}$$

证明：注意到 $\|A^{-1}\delta A\| \leq \|A^{-1}\|\|\delta A\| < 1$，且函数 $\dfrac{t}{1-t}$ 是增函数，由式(1.11) 可直接得出式 (1.12)。

不等式 (1.12) 用数据的相对误差表示矩阵逆的相对误差，只要条件数 $\mathrm{cond}(A)$ 不很大，矩阵逆的相对误差与数据的相对误差就为同一个数量级。

其次我们讨论在求解线性方程组 $A\boldsymbol{x} = \boldsymbol{b}$ 中系数矩阵 A 或向量 $\boldsymbol{b}$ 的数据误差对解的误差的影响。

定理 1.37 设 $A \in C_n^{n\times n}$，$\delta A \in C^{n\times n}$，$\boldsymbol{b}, \delta\boldsymbol{b} \in C^n$，而 C^n 上的向量范数 $\|\cdot\|_\alpha$ 与 $C^{n\times n}$ 上的矩阵范数 $\|\cdot\|$ 相容。设 $\boldsymbol{x}$ 是线性方程组 $A\boldsymbol{x} = \boldsymbol{b}$ 的解，$\hat{\boldsymbol{x}}$ 是线性方程组 $(A+\delta A)\hat{\boldsymbol{x}} = \boldsymbol{b}+\delta\boldsymbol{b}$ 的解。如果 $\|A^{-1}\|\|\delta A\| < 1$，则

$$\frac{\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_\alpha}{\|\boldsymbol{x}\|_\alpha} \leq \frac{\mathrm{cond}(A)}{1-\mathrm{cond}(A)(\|\delta A\|/\|A\|)}\frac{\|\delta A\|}{\|A\|}$$

$$+\frac{\mathrm{cond}(A)}{1-\mathrm{cond}(A)(\|\delta A\|/\|A\|)}\frac{\|\delta \boldsymbol{b}\|_\alpha}{\|\boldsymbol{b}\|_\alpha}. \tag{1.13}$$

证明：如果 $\|A^{-1}\|\|\delta A\| < 1$，由推论 1.34，矩阵 $A+\delta A = A(I+A^{-1}\delta A)$ 可逆，且

$$\begin{aligned}\boldsymbol{x}-\hat{\boldsymbol{x}} &= A^{-1}\boldsymbol{b}-(A+\delta A)^{-1}(\boldsymbol{b}+\delta \boldsymbol{b})\\ &= [A^{-1}-(A+\delta A)^{-1}]\boldsymbol{b}-(A+\delta A)^{-1}\delta \boldsymbol{b}\\ &= \sum_{k=1}^{\infty}(-1)^{k+1}(A^{-1}\delta A)^k A^{-1}\boldsymbol{b}+\sum_{k=0}^{\infty}(-1)^{k+1}(A^{-1}\delta A)^k A^{-1}\delta \boldsymbol{b}.\end{aligned}$$

所以，利用 $\|\boldsymbol{b}\|_\alpha = \|A\boldsymbol{x}\|_\alpha \le \|A\|\|\boldsymbol{x}\|_\alpha$，可得

$$\begin{aligned}\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_\alpha &\le \sum_{k=1}^{\infty}(\|A^{-1}\|\|\delta A\|)^k\|\boldsymbol{x}\|_\alpha+\sum_{k=0}^{\infty}(\|A^{-1}\|\|\delta A\|)^k\|A^{-1}\|\|\delta \boldsymbol{b}\|_\alpha\\ &= \frac{\|A^{-1}\|\|\delta A\|}{1-\|A^{-1}\|\|\delta A\|}\|\boldsymbol{x}\|_\alpha+\frac{1}{1-\|A^{-1}\|\|\delta A\|}\|A^{-1}\|\|\delta \boldsymbol{b}\|_\alpha\\ &\le \frac{\mathrm{cond}(A)}{1-\mathrm{cond}(A)(\|\delta A\|/\|A\|)}\frac{\|\delta A\|}{\|A\|}\|\boldsymbol{x}\|_\alpha\\ &\quad+\frac{\mathrm{cond}(A)}{1-\mathrm{cond}(A)(\|\delta A\|/\|A\|)}\frac{\|\delta \boldsymbol{b}\|_\alpha}{\|\boldsymbol{b}\|_\alpha}\|\boldsymbol{x}\|_\alpha\end{aligned}$$

从而可得出 (1.13)。

估计式 (1.13) 给出的相对误差的界有两项，一项是关于系数矩阵 A 的相对误差界，一项是关于右边向量 $\boldsymbol{b}$ 的相对误差界。然而这个估计式给出的相对误差的界不涉及所计算出的解或者由它导出的任何量。但是，假定我们通过某种方法计算出方程组 $A\boldsymbol{x}=\boldsymbol{b}$ 的某个“解”$\boldsymbol{x}$，比如，用第九章介绍的某方法求出方程组 $A\boldsymbol{x}=\boldsymbol{b}$ 的某个近似解 $\boldsymbol{x}$。恰好有 $A\boldsymbol{x}=\boldsymbol{b}$ 的情形往往是不可能的，不过，可以从剩余向量 $\boldsymbol{r}=\boldsymbol{b}-A\hat{\boldsymbol{x}}$ 得到 $\hat{\boldsymbol{x}}$ 与真解 $\boldsymbol{x}$ 如何接近的估计式。

定理 1.38 设 $A\in C_n^{n\times n}$，$\boldsymbol{b},\boldsymbol{r}\in C^n$，而 C^n 上的向量范数 $\|\cdot\|_\alpha$ 与 $C^{n\times n}$ 上的矩阵范数 $\|\cdot\|$ 相容。如果向量 $\boldsymbol{x},\hat{\boldsymbol{x}}\in C^n$ 分别满足 $A\boldsymbol{x}=\boldsymbol{b}$，$A\hat{\boldsymbol{x}}=\boldsymbol{b}-\boldsymbol{r}$，则有

$$\frac{\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_\alpha}{\|\boldsymbol{x}\|_\alpha}\le \mathrm{cond}(A)\frac{\|\boldsymbol{r}\|_\alpha}{\|\boldsymbol{b}\|_\alpha}. \tag{1.14}$$

证明：因为 $\|\boldsymbol{b}\|_\alpha = \|A\boldsymbol{x}\|_\alpha \le \|A\|\|\boldsymbol{x}\|_\alpha$，并且

$$\boldsymbol{x}-\hat{\boldsymbol{x}} = A^{-1}\boldsymbol{b}-A^{-1}(\boldsymbol{b}-\boldsymbol{r}) = A^{-1}\boldsymbol{r},$$

所以

$$\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_\alpha \le \|A^{-1}\boldsymbol{r}\|_\alpha \le \frac{\|A\|\|\boldsymbol{x}\|_\alpha}{\|\boldsymbol{b}\|_\alpha} = \|A\|\|A^{-1}\|\frac{\|\boldsymbol{r}\|_\alpha}{\|\boldsymbol{b}\|_\alpha}\|\boldsymbol{x}\|.$$

从而得到定理结论。

我们看到，无论在估计式 (1.12)，还是在估计式 (1.13) 和 (1.14)中， 矩阵 A 的条件数都起着决定性的作用，在条件数不大的情况下，矩阵逆的误差或线性方程组解的误差与数据的误差或剩余向量的误差同为一个数量级。

如果条件数 $\mathrm{cond}(A)$ 较小（接近 1），就称 A 关于求矩阵逆或求解线性方程组为良态的或好条件的； 如果条件数 $\mathrm{cond}(A)$ 较大，就称 A 关于求矩阵逆或求解线性方程组为病态的或坏条件的。

例 1.38 设

$$A=\begin{pmatrix}1 & -1\\ -1 & 1+\varepsilon\end{pmatrix},\quad \varepsilon>0.$$

证明对任意范数，当 $\varepsilon\to 0$ 时有 $\mathrm{cond}(A)=O(\varepsilon^{-1})$。因而矩阵 A 是病态的。

证明：可以求得 A 的特征值为 $\lambda_1=(2+\varepsilon+\sqrt{4+\varepsilon^2})/2$，$\lambda_2=(2+\varepsilon-\sqrt{4+\varepsilon^2})/2$。由于 A 为对称矩阵，所以对谱范数有

$$\begin{aligned}\mathrm{cond}_2(A)&=\frac{\lambda_1}{\lambda_2}=\frac{2+\varepsilon+\sqrt{4+\varepsilon^2}}{2+\varepsilon-\sqrt{4+\varepsilon^2}}\\ &=\frac{1}{\varepsilon}\left(2+\sqrt{4+\varepsilon^2}+\frac{\varepsilon}{2}\sqrt{4+\varepsilon^2}+\frac{\varepsilon^2}{2}+\varepsilon\right)\\ &=O(\varepsilon^{-1}),\quad (\text{当 }\varepsilon\to 0\text{ 时}).\end{aligned}$$

因为 $\boldsymbol{C}^{n\times n}$ 上所有矩阵范数是等价的，所以对任意范数，当 $\varepsilon\to 0$ 时有 $\mathrm{cond}(A)=O(\varepsilon^{-1})$。因而矩阵 A 是病态的。

例 1.39 设 $A=\begin{pmatrix}-1 & i & 0\\ -i & 0 & -i\\ 0 & i & -1\end{pmatrix}$, $\delta(A)\in\boldsymbol{C}^{3\times 3}$, $0\neq b\in\boldsymbol{C}^3$.

为使线性方程组 $A\boldsymbol{x}=b$ 的解 $\boldsymbol{x}$ 与 $(A+\delta(A))\boldsymbol{x}=b$ 的解 $\hat{\boldsymbol{x}}$ 的相对误差 $\dfrac{\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}\leq 10^{-4}$, 问 $\dfrac{\|\delta(A)\|_2}{\|A\|_2}$ 应不超过何值？

解：因为 $|\lambda\mathrm{I}-A|=(\lambda+1)(\lambda-1)(\lambda+2)$，所以 $\mathrm{cond}_2(A)=\|A\|_2\|A^{-1}\|_2=2$。从

$$\frac{\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}\leq\frac{\mathrm{cond}_2(A)}{1-\mathrm{cond}_2(A)\frac{\|\delta A\|_2}{\|A\|_2}}\frac{\|\delta A\|_2}{\|A\|_2}\leq 10^{-4},$$

即

$$\frac{2}{1-2\frac{\|\delta A\|_2}{\|A\|_2}}\frac{\|\delta A\|_2}{\|A\|_2}\leq 10^{-4},$$

得

$$\frac{\|\delta A\|_2}{\|A\|_2} \leq \frac{1}{2.0002} \times 10^{-4} \leq \frac{1}{2} \times 10^{-4} = 5 \times 10^{-5}.$$

第一章习题

1. 分别证明例1.1-例1.4定义的距离都满足距离的三个条件。

2. 证明极限的性质1.1。

3. 证明定理1.1。

4. 证明例1.16。

5. 证明定理1.7。

6. 证明稠密性具有传递性，即A在B中稠密，B在C中稠密，则A在C中稠密。

7. 设A是列紧的且是闭的，证明A是紧的。

8. 设X是度量空间，$x \in X$，A是X中的紧集，记

$$d(x, A) = \inf\{d(x, y) : y \in A\}$$

证明：当$d(x, A) = 0$时，那么$x \in A$； 若将A换成列紧的，结论是否成立？

9. 记$C[a, b]$为区间$[a, b]$上连续函数空间，对任何$f, g \in C[a, b]$，定义

$$d(f, g) = \max_{a \le x \le b} |f(x) - g(x)|.$$

证明d是$C[a, b]$上的距离，并且$C[a, b]$是完备空间。

10. 用压缩映射原理证明方程$x = a \sin x$只有唯一解，其中$a \in (0, 1)$。

11. 证明下述线性方程组

$$A = \begin{pmatrix} \frac{3}{4} & -\frac{1}{4} & \frac{1}{4} \\ -\frac{1}{7} & \frac{5}{14} & -\frac{1}{5} \\ -\frac{1}{100} & -\frac{1}{120} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{6} \\ \frac{1}{7} \\ \frac{1}{8} \end{pmatrix}$$

有唯一解，并写出方程近似解的迭代序列。

12. 用压缩映射原理构造迭代序列求解下述微分方程

$$\begin{cases} \dfrac{dy}{dx} = 1 + x^2, \\ x(0) = 0 \end{cases}$$

的解。

13. 赋范线性空间X是Banach空间的充要条件是对任意序列 $\{x_n\} \subset X$，若 $\sum\limits_{i=1}^{\infty}\|x_i\|$ 收敛，则 $\sum\limits_{i=1}^{\infty}x_i$ 在 X 中收敛。

14. 设X是无限维赋范线性空间，Y是X的有限维子空间，证明：必存在$x_0 \in X$且$\|x_0\|=1$，使得$d(x_0,Y)\geq 1$。

15. 设$\{e_n\}$是内积空间X中一个标准正交系，求证对任何$x,y\in X$，有

$$\sum_{n=1}^{\infty}|<x,e_n>\cdot<y,e_n>|\leq\|x\|\cdot\|y\|.$$

16. 设$\{e_n\}$是Hilbert空间X的一个标准正交系，对任何$x\in X$，求证

$$y=\sum_{n=1}^{\infty}<x,e_n>e_n$$

在X中更不能收敛，并且$x-y$与每个 e_n正交。

17. 设 $A\in C^{n\times n}$，证明：$\frac{1}{\sqrt{n}}\|A\|_F\leq\|A\|_2\leq\|A\|_F$

18. 设 $A=\begin{pmatrix}1/2 & 1\\ 0 & 1/2\end{pmatrix}$，对 $k=2,3,\cdots$，直接计算 A^k 及 $\rho(A^k)$，$\|A^k\|_1$，$\|A^k\|_\infty$，$\|A\|_2$。

19. 如果 $\|\cdot\|$ 是 $C^{n\times n}$ 上的矩阵范数，证明：对所有 $c\geq 1$，$c\|\cdot\|$ 是矩阵范数，但是，对任意 $c<1$，$c\|\cdot\|_1$ 不是矩阵范数。

20. 设 $A\in C^{n\times n}$，证明 Hermite 矩阵

$$\hat{A}=\begin{pmatrix}0 & A\\ A^H & 0\end{pmatrix}\in C^{2n\times 2n}.$$

与 A 有相同的谱范数。

21. 设 $A,B\in C^{n\times n}$ 都是 Hermite 矩阵，证明：$\rho(A+B)\leq\rho(A)+\rho(B)$.

22. 证明，如果 $A\in C^{n\times n}$ 可逆，则 $\mathrm{cond}(A)=\mathrm{cond}(A^{-1})$。

23. 设 $A=\begin{pmatrix}2 & 1 & 3\\ 8 & 2 & 4\\ 4 & 4 & 10\end{pmatrix}$，求 A 的条件数 $\mathrm{cond}_\infty(A)$。

24. 证明：$\mathrm{cond}(AB)\leq\mathrm{cond}(A)\mathrm{cond}(B)$。那么 $\mathrm{cond}(\cdot)$ 是 $C^{n\times n}$ 上的矩阵范数吗？

第 二 章　矩阵的标准型与特征值计算

本章主要介绍矩阵在相似变换下的标准形，及矩阵特征值估计和计算的幂迭代法。它们不仅是矩阵理论和矩阵计算的基本问题，而且在许多工程技术研究领域也有着广泛的应用。

2.1　λ-矩阵及标准形、不变因子和初等因子

本节引进 λ-矩阵这个工具，为后面的矩阵化为Jordan标准形作准备。先介绍多项式的最大公因式及一些性质。

设 $f(x)=a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0$ 为复数域 $\boldsymbol{C}$ 上关于未定元x的多项式，n 为正整数，则称 $f(x)$ 是一个 n 次多项式，a_nx^n 称为 $f(x)$ 的最高次项或首项，a_n 称为首项系数，a_0 称为常数项。若 $f(x)=a$，则称 $f(x)$ 为常数多项式，当 $a\neq 0$ 时，称它为零次多项式，当 $a=0$ 时，称之为零多项式。称多项式 $f(x)=a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0$ 为首1多项式，如果 $f(x)$ 的首项系数为 1，即有形式：

$$f(x)=x^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0.$$

设 $f(x)$，$g(x)$ 是复数域 $\boldsymbol{C}$ 上的多项式，如果存在复数域 $\boldsymbol{C}$ 上的多项式 $h(x)$，使

$$f(x)=g(x)h(x),$$

则称 $g(x)$ 整除 $f(x)$，记作 $g(x)|f(x)$. 此时称 $g(x)$ 为 $f(x)$ 的因式（或因子），$f(x)$ 称为 $g(x)$ 的倍式。

例如 $f(x)=(x+1)^2, g(x)=2(x+1)$, 则 $g(x)|f(x)$，因 $f(x)=\frac{1}{2}(x+1)g(x)$.

由整除的定义知道，零多项式的因式可以是任一多项式，但零多项式不能是非零多项式的因式。下面不予证明的给出整除性的几个常用性质：

（1）若 $g(x)|f(x)$，则 $cg(x)|f(x)\quad(c\neq 0)$, 由此知非零常数是任一非零多项式的因子；

（2）若 $g(x)|f(x)$，$f(x)|g(x)$，且 $f(x), g(x)$ 都是非零多项式，则 $f(x)=cg(x)$，其中 c 为非零常数；

（3）若 $g(x)|f(x)$，$f(x)|h(x)$，则 $g(x)|h(x)$；

（4）若 $g(x)|f(x)$，$g(x)|h(x)$，则对任意的多项式 $u(x), v(x)$，有

$$g(x)|(f(x)u(x)+h(x)v(x)).$$

如果多项式 $\varphi(x)$ 满足：$\varphi(x)|f(x), \varphi(x)|g(x)$，则称 $\varphi(x)$ 是 $f(x)$ 与 $g(x)$ 的一个公因式（或公因子）。

设 $f(x), g(x)$ 是复数域 $\boldsymbol{C}$ 上多项式，如果存在复数域 $\boldsymbol{C}$ 上多项式 $d(x)$ 满足：（1）$d(x)|f(x), d(x)|g(x)$；（2）若 $d_1(x)$ 是 $f(x)$ 与 $g(x)$ 的公因式，则必有 $d_1(x)|d(x)$. 则称 $d(x)$ 是 $f(x), g(x)$ 的一个最大公因式。

若 $d_1(x)$，$d_2(x)$ 是 $f(x)$与$g(x)$ 的两个最大公因式，则由定义知二者必互相整除，即 $d_1(x)|d_2(x)$，$d_2(x)|d_1(x)$，这样就有 $d_1(x)=cd_2(x), c\neq 0$，这说明 $f(x)$ 与 $g(x)$ 的两个最大公因式至多相差一个非零常数。用 $(f(x), g(x))$ 表示首项系数为 1 的最大公因式。显然 $(f(x), g(x))$ 是唯一确定的。特别地，若 c 为非零常数，$f(x)$ 是首 1 多项式，则有：$(f(x), c)=1$，$(f(x), 0)=f(x)$. 多个多项式 $f_1(x), \cdots, f_s(x)$ 的最大公因式可同样定义，仍用 $(f_1(x), \cdots, f_s(x))$ 表示首项系数为 1 的最大公因式，如对 $s=3$，有

$$(f_1(x), f_2(x), f_3(x))=((f_1(x), f_2(x)), f_3(x)).$$

一般地，在复数域上求若干个多项式的最大公因式时，可先把每个多项式分解成一次因式的方幂的乘积形式，然后取含有公共一次因式的最低方幂的乘积，即得所求的最大公因式。例如 $f(x)=(x-1)^3(x+2)^2(x+5), g(x)=(x-1)^2(x+2)^3(x+3), h(x)=(x-1)^2(x+2)(x+3)^5$，则有 $(f(x), g(x), h(x))=(x-1)^2(x+2)$。

2.1.1 λ-矩阵的概念

定义 2.1 *设 $a_{ij}(\lambda)(i=1,2,\cdots,m; j=1,2,\cdots,n)$ 为复数域 $\boldsymbol{C}$ 上的多项式，则称以 $a_{ij}(\lambda)$ 为元素的 $m\times n$ 阶矩阵*

$$A(\lambda)=\begin{pmatrix} a_{11}(\lambda) & a_{12}(\lambda) & \cdots & a_{1n}(\lambda) \\ a_{21}(\lambda) & a_{22}(\lambda) & \cdots & a_{2n}(\lambda) \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}(\lambda) & a_{m2}(\lambda) & \cdots & a_{mn}(\lambda) \end{pmatrix}$$

为λ-矩阵或多项式矩阵。

为与 λ-矩阵相区别，我们把以复数域 $\boldsymbol{C}$ 中的数为元素的矩阵称为数字矩阵。显然，数字矩阵可看作是特殊的 λ-矩阵，因它的元素 a_{ij} 为零次多项式。方阵 A 的特征矩阵 $\lambda I-A$ 也是一个特殊的 λ-矩阵。

λ-矩阵可同数字矩阵一样定义相等、加法、数乘、乘法等运算，且与数字矩阵有相同的运算规律。对于$n\times n$ 阶的方 λ-矩阵可同样定义行列式，而且与数字矩阵的行列式有相同的性质。有了 λ-矩阵行列式的概念，也就有了 λ-矩阵的子式、余子式、伴随矩阵等概念，进而利用子式这个概念，就可给出 λ-矩阵秩的定义。

定义 2.2 如果 λ-矩阵 $A(\lambda)$ 中有一个 r 阶（$r \geq 1$）子式为非零多项式，而所有 $r+1$ 阶子式（如果有的话）全为零多项式，则称 $A(\lambda)$ 的秩为 r，记作 $rankA(\lambda)=r$ 或 $r_{A(\lambda)}=r$。零矩阵的秩规定为零。

如果 n 阶λ-矩阵 $A(\lambda)$ 秩为 n，即 $\det A(\lambda) \neq 0$，则称 $A(\lambda)$ 为满秩的或非奇异的。若 n 阶 λ-矩阵 $A(\lambda)$ 的秩小于 n，也即 $\det A(\lambda)=0$，就称 $A(\lambda)$ 是降秩的或奇异的。例如，数字矩阵 $A=(a_{ij})_{n\times n}$ 的特征矩阵 $A(\lambda)=\lambda I-A$ 就是重要的满秩矩阵，因为 $f_A(\lambda)=\det(\lambda I-A)$ 不是零多项式（至多有 n 个根）。对λ-矩阵也可以有初等变换。

定义 2.3 对 λ-矩阵 $A(\lambda)$ 施行的下列 3 种变换称为 λ-矩阵的初等行（列）变换：

（1）交换 $A(\lambda)$ 的任意两行（列），（交换 i,j 两行（列），记作 $r_i \leftrightarrow r_j\,(c_i \leftrightarrow c_j)$）；

（2）数 $k(k\neq 0)$ 乘 $A(\lambda)$ 的某一行（列），（第 i 行（列）乘以 k，记作 $kr_i\,(kc_i)$）；

（3）$A(\lambda)$ 的某一行（列）的 $\varphi(\lambda)$ 倍加到另一行（列）上去，其中 $\varphi(\lambda)$ 是 λ 的一个多项式。（第 i 行（列）的 $\varphi(\lambda)$ 倍加到第 j 行（列）上，记作 $\varphi(\lambda)r_i+r_j\ (\varphi(\lambda)c_i+c_j)$.

定义 2.4 如果 λ-矩阵 $A(\lambda)$ 经过有限次初等变换变为 λ-矩阵 $B(\lambda)$，则称 $A(\lambda)$ 与 $B(\lambda)$ 等价，记作 $A(\lambda)\cong B(\lambda)$。

容易验证，两 λ-矩阵等价具有反身性、对称性、传递性。

2.1.2 λ-矩阵的Smith标准形、不变因子和行列式因子

下面主要给出 λ-矩阵在初等变换下的一种重要的标准形–Smith标准形，进而给出 λ-矩阵的不变因子和行列式因子。

引理 2.1 若λ-矩阵$A(\lambda)=(a_{ij}(\lambda))_{m\times n}$的左上角元素$a_{11}(\lambda)\neq 0$，并且$A(\lambda)$中至少有一个元素不能被$a_{11}(\lambda)$所整除，则必可找到一个与$A(\lambda)$等价的矩阵$B(\lambda)$，其左上角元素$b_{11}(\lambda)$也不为零，且$b_{11}(\lambda)$的次数低于$a_{11}(\lambda)$的次数。

定理 2.1 任一非零的λ-矩阵$A(\lambda)=(a_{ij}(\lambda))_{m\times n}$都等价于如下形式的矩阵

$$A(\lambda)\cong S(\lambda)=\mathrm{diag}(d_1(\lambda),d_2(\lambda),\cdots,d_r(\lambda),0,\cdots,0)$$

其中$r\geq 1$是$A(\lambda)$的秩，$d_i(\lambda)\,(i=1,2,\cdots,r)$是首项系数为1的多项式，且 $d_i(\lambda)|d_{i+1}(\lambda)$, $i=1,\cdots,r-1$. λ-矩阵 $S(\lambda)$ 称为 $A(\lambda)$ 的Smith标准形，$d_i(\lambda)$ $(i=1,2,\cdots,r)$ 称为 $A(\lambda)$ 的不变因子（或不变因式）。

证明：因$A(\lambda)\neq 0$，知至少有一个非零元素，所以不妨设$a_{11}(\lambda)\neq 0$，否则，总可以经过行列调整，使得$A(\lambda)$的左上角元素不为零。如果$a_{11}(\lambda)$不能整除$A(\lambda)$的其余元素，则由引理 2.1知，可以找到与$A(\lambda)$等价的矩阵$B_1(\lambda)$，使$B_1(\lambda)$的左上角元素$b_{11}^{(1)}(\lambda)\neq 0$，并且$b_{11}^{(1)}(\lambda)$的次数低于$a_{11}(\lambda)$的次数。若$b_{11}^{(1)}(\lambda)$仍不能整除$B_1(\lambda)$的其余所有元素，则可用同样的方法，逐步降低左上角元素的次数，但这些次数都是非负整数，不能无限地降低，因此在有限步以后，将会终止于一个λ-矩阵$B_s(\lambda)$，其左上角元素$b_{11}^{(s)}(\lambda)\neq 0$，并且$b_{11}^{(s)}(\lambda)$可以整除$B_s(\lambda)$的其它所有元素$b_{ij}^{(s)}(\lambda)$.令$b_{ij}^{(s)}(\lambda)=b_{11}^{(s)}(\lambda)\,q_{ij}(\lambda)$，对$B_s(\lambda)$作初等变换：

$$B_s(\lambda)=\begin{pmatrix} b_{11}^{(s)}(\lambda) & \cdots & b_{1j}^{(s)}(\lambda) & \cdots \\ \vdots & & \vdots & \\ b_{i1}^{(s)}(\lambda) & \cdots & b_{ij}^{(s)}(\lambda) & \cdots \\ \vdots & & \vdots & \end{pmatrix}\begin{pmatrix} b_{11}^{(s)}(\lambda) & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & B_{s+1}(\lambda) & \\ 0 & & & \end{pmatrix}.$$

对λ-矩阵$B_{s+1}(\lambda)$，因其元素是$B_s(\lambda)$元素的组合，又 $B_s(\lambda)$ 的所有元素能被 $b_{11}^{(s)}(\lambda)$ 整除， 故$b_{11}^{(s)}(\lambda)$ 也可以整除 $B_{s+1}(\lambda)$的所有元素。而对上面后一个矩阵再进行一次初等变换，还可以把$b_{11}^{(s)}(\lambda)$化为首一多项式$d_1(\lambda)$，且不改变$B_{s+1}(\lambda)$的元素。如果$B_{s+1}(\lambda)\neq 0$，则对$B_{s+1}(\lambda)$可重复上述过程，于是又有$A(\lambda)$的等价矩阵

$$\begin{pmatrix} d_1(\lambda) & 0 & 0 & \cdots & 0 \\ 0 & d_2(\lambda) & 0 & \cdots & 0 \\ 0 & 0 & & & \\ \vdots & \vdots & & B_{S+2}(\lambda) & \\ 0 & 0 & & & \end{pmatrix},$$

其中$d_1(\lambda),d_2(\lambda)$均为首1多项式，且$d_1(\lambda)\mid d_2(\lambda)$，而$d_2(\lambda)$可整除$B_{s+2}(\lambda)$的各个元素。继续上述过程，最后可把$A(\lambda)$化成所要求的形式。 证毕。

例 2.1 *求λ-矩阵*

$$A(\lambda)=\begin{pmatrix} 1-\lambda & 2\lambda-1 & \lambda \\ \lambda & \lambda^2 & -\lambda \\ 1+\lambda^2 & \lambda^2+\lambda-1 & -\lambda^2 \end{pmatrix}$$

的Smith标准形和不变因子。

解：因$A(\lambda)$的左上角元素$1-\lambda$不能整除其它所有元素，所以由引理 2.1 知，可先降低它的次数，但$1-\lambda$是1 次的，降低次数只能是零次的，即是常数。故需把$1-\lambda$化为常数，为此只需第3列加到第1列，便得

$$A_2(\lambda)=\begin{pmatrix} 1 & 2\lambda-1 & \lambda \\ 0 & \lambda^2 & -\lambda \\ 1 & \lambda^2+\lambda-1 & -\lambda^2 \end{pmatrix},$$

其左上角元素1已能整除其余元素，因此可把与左上角元素1同一行同一列的其它非零元素都化为零，于是得

$$A_3(\lambda)=\begin{pmatrix}1 & 0 & 0\\ 0 & \lambda^2 & -\lambda\\ 0 & \lambda^2-\lambda & -\lambda^2-\lambda\end{pmatrix}.$$

在$A_3(\lambda)$中，由于λ^2不能整除剩余元素，如不能整除$\lambda^2-\lambda$，同样可降低λ^2的次数，显然λ^2应降为1次的，这只需把$A_3(\lambda)$的第2列与第3列交换即得

$$A_4(\lambda)=\begin{pmatrix}1 & 0 & 0\\ 0 & -\lambda & \lambda^2\\ 0 & -\lambda^2-\lambda & \lambda^2-\lambda\end{pmatrix},$$

这时可把$A_4(\lambda)$中$-\lambda$右边同一行、下方同一列的非零元都化为零，得

$$A_5(\lambda)=\begin{pmatrix}1 & 0 & 0\\ 0 & -\lambda & 0\\ 0 & 0 & -\lambda^3-\lambda\end{pmatrix},$$

$A_5(\lambda)$ 的第2行与第3行乘以-1便得所求的Smith标准形

$$S(\lambda)=\begin{pmatrix}1 & 0 & 0\\ 0 & \lambda & 0\\ 0 & 0 & \lambda^3+\lambda\end{pmatrix}.$$

不变因子为$d_1(\lambda)=1,\ d_2(\lambda)=\lambda, d_3(\lambda)=\lambda^3+\lambda$.

我们知道，数字矩阵在初等变换下的标准形是唯一的，同样可证λ-矩阵在初等变换下的Smith标准形是唯一的。为此引入

定义 2.5 *设λ-矩阵$A(\lambda)$的秩为$r\geq 1$，对于正整数$k\,(1\leq k\leq r)$，$A(\lambda)$中的所有非零的k阶子式的首一最大公因式$D_k(\lambda)$称为$A(\lambda)$的k阶行列式因子。*

显然，当$rankA(\lambda)=r\geq 1$时，$A(\lambda)$共有r个行列式因子$D_1(\lambda),\cdots,D_r(\lambda)$，它们是由$A(\lambda)$唯一确定的。因为$D_{k-1}(\lambda)$能整除每一个$k-1$阶子式，而$k$阶子式按一行展开后可表为一些$k-1$阶子式的组合，所以$D_{k-1}(\lambda)|\,D_k(\lambda)\ (k=2,\cdots,r)$。

定理 2.2 *等价的λ-矩阵具有相同的秩及相同的各阶行列式因子。即λ-矩阵的秩和各阶行列式因子在初等变换下是不变的。*

定理 2.3 *λ-矩阵$A(\lambda)$的Smith标准形*

$$S(\lambda)=\begin{pmatrix} d_1(\lambda) & & & & & \\ & d_2(\lambda) & & & & \\ & & \ddots & & & \\ & & & d_r(\lambda) & & \\ & & & & 0 & \\ & & & & & \ddots \\ & & & & & & 0 \end{pmatrix}$$

是唯一的，且$d_1(\lambda)=D_1(\lambda), d_k(\lambda)=\frac{D_k(\lambda)}{D_{k-1}(\lambda)}, k=2,\cdots,r.$

证明：因$A(\lambda)\simeq S(\lambda)$，所以$A(\lambda)$与$S(\lambda)$有相同的行列式因子。由 $d_1(\lambda),\cdots,d_r(\lambda)$ 为首1多项式，且 $d_k(\lambda)|d_{k+1}(\lambda)(k=1,2,\cdots,r)$，易知 $S(\lambda)$ 的 k阶行列式因子为 $d_1(\lambda)d_2(\lambda)\cdots d_k(\lambda)$，于是

$$D_k(\lambda)=d_1(\lambda)d_2(\lambda)\cdots d_k(\lambda),\quad k=1,2,\cdots,r. \tag{2.1}$$

因此，得

$$d_1(\lambda)=D_1(\lambda),\quad d_k(\lambda)=\frac{D_k(\lambda)}{D_{k-1}(\lambda)},\quad k=2,\cdots,r. \tag{2.2}$$

这说明$A(\lambda)$的不变因子由$A(\lambda)$的行列式因子唯一确定，而行列式因子在初等变换下不变，所以$A(\lambda)$的Simth标准形是唯一的。 证毕。

(2.1) 式和 (2.2) 给出了不变因式与行列式因子的关系，通过求行列式因子，也可求出$A(\lambda)$的不变因式，进而也可给出$A(\lambda)$的Smith标准形。

2.1.3 初等因子

定义 2.6 *把λ-矩阵$A(\lambda)$的每个次数≥ 1的不变因子$d_k(\lambda)$在复数域上分解为互不相同的一次因式的方幂的乘积，所有这些一次因式的方幂（相同的按出现的次数计算），称为$A(\lambda)$的初等因子，全体初等因子的集合称为初等因子组。*

设秩为r的λ-矩阵$A(\lambda)$的不变因子为$d_1(\lambda),d_2(\lambda),\cdots,d_r(\lambda)$. 将$d_k(\lambda)$ $(k=1,2,\cdots,r)$ 分解为互不相同的一次因式方幂的乘积：

$$\begin{aligned} d_1(\lambda)&=(\lambda-\lambda_1)^{k_{11}}(\lambda-\lambda_2)^{k_{12}}\cdots(\lambda-\lambda_s)^{k_{1s}},\\ d_2(\lambda)&=(\lambda-\lambda_1)^{k_{21}}(\lambda-\lambda_2)^{k_{22}}\cdots(\lambda-\lambda_s)^{k_{2s}},\\ &\cdots\cdots\cdots\\ d_r(\lambda)&=(\lambda-\lambda_1)^{k_{r1}}(\lambda-\lambda_2)^{k_{r2}}\cdots(\lambda-\lambda_s)^{k_{rs}}. \end{aligned}$$

这里 $\lambda_1,\lambda_2,\cdots,\lambda_s$ 为 $d_r(\lambda)$ 的全部互异零点，可知 $k_{r1},k_{r2},\cdots,k_{rs}$ 均不为零。又 $d_{k-1}(\lambda)\,|d_k(\lambda)\ (k=2,\cdots,r)$，所以

$$k_{1j}\le k_{2j}\le\cdots\le k_{rj}\quad (j=1,2,\cdots,s). \tag{2.3}$$

但 $k_{1j},k_{2j},\cdots,k_{r-1,j}\,(j=1,2,\cdots,s)$ 中可能有 $k_{ij}=0$，这时必有 $k_{1j}=\cdots=k_{i-1,j}=0$. 这样，下式

$$\begin{aligned}&(\lambda-\lambda_1)^{k_{11}},(\lambda-\lambda_2)^{k_{12}},\cdots,(\lambda-\lambda_s)^{k_{1s}},\\&(\lambda-\lambda_1)^{k_{21}},(\lambda-\lambda_2)^{k_{22}},\cdots,(\lambda-\lambda_s)^{k_{2s}},\\&\cdots\cdots\cdots\\&(\lambda-\lambda_1)^{k_{r1}},(\lambda-\lambda_2)^{k_{r2}},\cdots,(\lambda-\lambda_s)^{k_{rs}},\end{aligned}$$

中不是常数的因子就是$A(\lambda)$的全部初等因子。

如例 2.1 中$A(\lambda)$的不变因式分别为1,λ, $\lambda(\lambda^2+1)$，所以它的初等因子组是λ, λ, $\lambda+i$, $\lambda-i$。由此可见，初等因子组中可以有相同者。

推论 2.1 *设λ-矩阵$A(\lambda)$的秩为r，则$A(\lambda)$的不变因子与其初等因子互相决定。*

事实上，从式(2.3)可知，不同一次因式$(\lambda-\lambda_j)$的最高次幂相乘即得$d_r(\lambda)$，不同一次因式$(\lambda-\lambda_j)$的次高次幂相乘可得$d_{r-1}(\lambda)$，顺次可定出$d_{r-2}(\lambda),\cdots,d_1(\lambda)$.

例如，若已知5 × 6阶λ-矩阵$A(\lambda)$的秩为4，其初等因子为

$$\lambda,\lambda,\lambda^2,\lambda-1,(\lambda-1)^2,(\lambda-1)^3,(\lambda+1)^2,(\lambda+1)^3,(\lambda-2).$$

则可求得 $A(\lambda)$ 的不变因子为

$$d_4(\lambda)=\lambda^2(\lambda-1)^3(\lambda+1)^3(\lambda-2),\ d_3=\lambda(\lambda-1)^2(\lambda+1)^2,$$

$$d_2(\lambda)=\lambda(\lambda-1),\ d_1(\lambda)=1.$$

下面的结论对求初等因子是有用的。

定理 2.4 *设λ-矩阵*

$$A(\lambda)=\begin{pmatrix}B(\lambda)&0\\0&C(\lambda)\end{pmatrix},$$

其中$B(\lambda)$为$m\times n$阶λ-矩阵，$C(\lambda)$为$p\times l$阶λ-矩阵，则$B(\lambda)$，$C(\lambda)$的各个初等因子的全体是$A(\lambda)$的全部初等因子。

注意： 我们可从定理 2.4中的$B(\lambda)$和$C(\lambda)$的初等因子得到$A(\lambda)$的初等因子，但不能从$B(\lambda)$和$C(\lambda)$的不变因子求得$A(\lambda)$的不变因子。

2.2 Jordan标准形

Joradn标准形理论是矩阵理论的重要结论之一，它是说任一复矩阵均可相似于一特殊的上三角阵（下三角阵）。从而许多理论与实际的问题，用到它便可迎刃而解了。

2.2.1 矩阵相似的条件

设A是n阶数字矩阵，其特征矩阵$\lambda I - A$是λ-矩阵，利用这个特殊的λ-矩阵，可以给出两个n阶数字矩阵相似的判断准则。

定理 2.5 设$A, B \in C^{n\times n}$，则$A \sim B$的充分必要条件是$(\lambda I - A) \cong (\lambda I - B)$。

定义 2.7 设$A \in C^{n\times n}$，则其特征矩阵$\lambda I - A$的不变因子、行列式因子、初等因子分别称为A的不变因子、行列式因子、初等因子。

推论 2.2 设$A, B \in C^{n\times n}$，则下列命题等价：

(1) $A \sim B$；

(2) A, B有相同的各阶行列式因子；

(3) A, B有相同的不变因子；

(4) A, B有相同的初等因子。

2.2.2 矩阵的Jordan标准形

定义 2.8 形如

$$J_i = \begin{pmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{pmatrix}_{n_i \times n_i} \tag{2.4}$$

的方阵称为n_i阶 Jordan块，$\lambda_i \in C$.

例如:

$$\begin{pmatrix} 2 & 1 \\ & 2 \end{pmatrix}, \begin{pmatrix} -i & 1 & \\ & -i & 1 \\ & & -i \end{pmatrix}, \begin{pmatrix} 0 & 1 & & \\ & 0 & 1 & \\ & & 0 & 1 \\ & & & 0 \end{pmatrix}$$

都是Jordan块。特别地，一阶方阵是一阶Jordan块。

易知，式（2.4）中Jordan块J_i的初等因子为$(\lambda-\lambda_i)^{n_i}$。显然，因式$(\lambda-\lambda_i)^{n_i}$的幂指数为$J_i$的阶数，而$\lambda_i$即为$J_i$的主对角元，反之，若给定因式$(\lambda-\lambda_i)^{n_i}$，则必可写出一个$n_i$阶Jordan块$J_i$与之对应。因此，Jordan块被它的初等因子唯一决定，也称J_i为因式$(\lambda-\lambda_i)^{n_i}$对应的Jordan块。

定义 2.9 设$J_1, J_2, \cdots, J_s$为形如（2.4）的Jordan块，则称块对角矩阵

$$J=\begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_s \end{pmatrix} \tag{2.5}$$

为Jordan标准形。

由（2.5）知J的特征矩阵为

$$\lambda I-J=\operatorname{diag}(\lambda I_{n_1}-J_1,\cdots,\lambda I_{n_s}-J_s).$$

利用定理 2.4和Jordan块初等因子的讨论，知Jordan标准形J的初等因子为

$$(\lambda-\lambda_1)^{n_1},(\lambda-\lambda_2)^{n_2},\cdots,(\lambda-\lambda_s)^{n_s}.$$

可见，Jordan标准形的全部初等因子由它的全部Jordan块的初等因子组成，而Jordan块被它的初等因子唯一决定，因此，Jordan标准形除去其中Jordan块排列的次序外被它的初等因子唯一决定。

定理 2.6 设$A\in C^{n\times n}$，则A与一个Jordan标准形相似，并且这个Jordan标准形除了其中Jordan块的排列次序外被A唯一决定，常称其为A的Jordan标准形，并常记为J_A.

证明：设A的初等因子为

$$(\lambda-\lambda_1)^{n_1},\ (\lambda-\lambda_2)^{n_2},\ \cdots,\ (\lambda-\lambda_s)^{n_s}, \tag{2.6}$$

其中$\lambda_1,\cdots,\lambda_s$可能有相同的，$n_1,\cdots,n_s$也可能有相同，但总有$\sum_{i=1}^{s}n_i=n$. 每个初等因子$(\lambda-\lambda_i)^{n_i}$对应于一个Jordan块

$$J_i=\begin{pmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda_i \end{pmatrix}_{n_i\times n_i},\ i=1,\cdots,s.$$

并令

$$J=\begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_s \end{pmatrix}.$$

则J的初等因子也是（2.6）.这样，J与A有相同的初等因子，由推论 2.2 知$A\sim J$。

如果有另一个Jordan标准形$\tilde{J}$，使得$\tilde{J}\sim A$。则$\tilde{J}$与A也有相同的初等因子。因此，$\tilde{J}$与J除去其中Jordan块排列的次序外是相同的，这就证明了唯一性。证毕。

注意到$n_i>1$时，J_i是不可对角化的，而$n_i=1$时，$J_i=\lambda_i$是一阶Jordan块，其初等因子是一次的。故可得

推论 2.3 设$A\in C^{n\times n}$，则A可对角化当且仅当A的初等因子均是一次的。

现将方阵A的Jordan标准形的求解步骤归纳如下：

第一步，求出n阶方阵A的初等因子：

$$(\lambda-\lambda_1)^{n_1},(\lambda-\lambda_2)^{n_2},\cdots,(\lambda-\lambda_s)^{n_s},$$

其中$\lambda_1,\cdots,\lambda_s$可能有相同的，指数$n_1,n_2,\cdots,n_s$也可能有相同的，且$\sum\limits_{i=1}^{s}n_i=n$. 完成这一步，有三种方法：

（1）用初等变换化特征矩阵$\lambda I-A$为Smith标准形，求出A的不变因子 $d_1(\lambda)$, $d_2(\lambda)$, $\cdots$, $d_n(\lambda)$ （因$\operatorname{rank}(\lambda I-A)=n$，所以$A$有$n$个不变因子），再将次数大于0的不变因子分解成互不相同的一次因式方幂的乘积，即可得A的初等因子。

（2）用初等变换化$\lambda I-A$为对角形$\operatorname{diag}(f_1(\lambda),\cdots,f_n(\lambda))$，再将 $f_1(\lambda)$, $\cdots$, $f_n(\lambda)$ 分解成一次因式的方幂，即可得初等因子。

（3）直接求出$\lambda I-A$的行列式因子$D_1(\lambda),\cdots,D_n(\lambda)$，由此求出$A$的不变因子再分解，从而得到$A$的初等因子。

（1）、（2）称为初等变换法，（3）称为行列式因子法。

第二步，写出每个初等因子$(\lambda-\lambda_i)^{n_i}$ $(i=1,2,\cdots,s)$对应的Jordan块

$$J_i=\begin{pmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}_{n_i\times n_i},\quad i=1,2,\cdots,s.$$

第三步，写出以这些Jordan块构成的Jordan标准形

$$J=\begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_s \end{pmatrix}.$$

即为A的Jordan标准形。

例 2.2 求矩阵$A=\begin{pmatrix}0&1&0\\-4&4&0\\-2&1&2\end{pmatrix}$的Jordan标准形。

解：因

$$\lambda I-A=\begin{pmatrix}\lambda&-1&0\\4&\lambda-4&0\\2&-1&\lambda-2\end{pmatrix}\xrightarrow[c_1\leftrightarrow c_2]{}\begin{pmatrix}-1&\lambda&0\\\lambda-4&4&0\\-1&2&\lambda-2\end{pmatrix},$$

$$\to\begin{pmatrix}1&0&0\\0&(\lambda-2)^2&0\\0&-(\lambda-2)&\lambda-2\end{pmatrix}\to\begin{pmatrix}1&0&0\\0&\lambda-2&0\\0&0&(\lambda-2)^2\end{pmatrix}.$$

可见A的不变因子为$d_1(\lambda)=1,d_2(\lambda)=\lambda-2,d_3(\lambda)=(\lambda-2)^2$，$A$的初等因子为$\lambda-2,(\lambda-2)^2$. 故$A$的Jordan标准形为

$$J_A=\begin{pmatrix}2&0&0\\0&2&1\\0&0&2\end{pmatrix}.$$

例 2.3 求下列矩阵的Jordan标准形：

(1) $A=\begin{pmatrix}1&1&0\\6&2&0\\0&1&-1\end{pmatrix}$，(2) $B=\begin{pmatrix}3&1&0&0\\-4&-1&0&0\\7&1&2&1\\-7&-6&-1&0\end{pmatrix}$.

解：(1) 因

$$\lambda I-A=\begin{pmatrix}\lambda-1&-1&0\\-6&\lambda-2&0\\0&-1&\lambda+1\end{pmatrix}.$$

显然有$D_3(\lambda)=\det(\lambda I-A)=(\lambda+1)^2(\lambda-4)$，又$\lambda I-A$中有二阶子式

$$D_{13}=\begin{vmatrix}-6&\lambda-2\\0&-1\end{vmatrix}=6.$$

因为$D_2(\lambda)$整除每个二阶子式，且$D_2(\lambda)|D_3(\lambda)$，所以$D_2(\lambda)=1$，从而$D_1(\lambda)=1$. 于是$A$的不变因子为

$$d_1(\lambda)=d_2(\lambda)=1,\quad d_3(\lambda)=\frac{D_3(\lambda)}{D_2(\lambda)}=(\lambda+1)^2(\lambda-4).$$

得A的初等因子为$(\lambda+1)^2,(\lambda-4)$，故$A$的Jordan标准形为

$$J_A=\begin{pmatrix}-1&1&0\\0&-1&0\\0&0&4\end{pmatrix}.$$

（2）因

$$\lambda I-B=\begin{pmatrix}\lambda-3&-1&0&0\\4&\lambda+1&0&0\\-7&-1&\lambda-2&-1\\7&6&1&\lambda\end{pmatrix}.$$

易得$D_4(\lambda)=\det(\lambda I-B)=(\lambda-1)^4$，又可找到$\lambda I-B$中的两个三阶子式

$$D_{44}=\begin{vmatrix}\lambda-3&-1&0\\4&\lambda+1&0\\-7&-1&\lambda-2\end{vmatrix}=(\lambda-2)(\lambda-1)^2,$$

$$D_{23}=\begin{vmatrix}\lambda-3&-1&0\\-7&-1&-1\\7&6&\lambda\end{vmatrix}=-\lambda^2+2\lambda-11.$$

因$\left((\lambda-2)(\lambda-1)^2,-\lambda^2+2\lambda-11\right)=1$，所以$D_3(\lambda)=1$，进而有$D_2(\lambda)=D_1(\lambda)=1$，得$B$的不变因子为

$$d_1(\lambda)=d_2(\lambda)=d_3(\lambda)=1,\quad d_4(\lambda)=(\lambda-1)^4.$$

于是B的初等因子为$(\lambda-1)^4$，故B的Jordan标准形为

$$J_B=\begin{pmatrix}1&1&0&0\\0&1&1&0\\0&0&1&1\\0&0&0&1\end{pmatrix}.$$

由定理 2.6知，对任意的n阶矩阵A，存在n阶可逆矩阵P，使$P^{-1}AP=J_A$.上面给出了J_A的求法，但还未涉及相似变换阵P的求解。下面以三阶矩阵为例介绍P 的求法。

例 2.4 设$A=\begin{pmatrix}0&1&0\\-4&4&0\\-2&1&2\end{pmatrix}$，求$P$使$P^{-1}AP$为Jordan标准形。

解：由例 2.2知A的Jordan标准形为

$$J_A=\begin{pmatrix}2&0&0\\0&2&1\\0&0&2\end{pmatrix}.$$

设$P=(p_1,p_2,p_3)$,由$P^{-1}AP=J_A$，得$AP=PJ_A$，即

$$(Ap_1,Ap_2,Ap_3)=(2p_1,2p_2,p_2+2p_3).$$

从而得

$$\begin{cases} Ap_1=2p_1, \\ Ap_2=2p_2, \\ Ap_3=p_2+2p_3. \end{cases}$$

可见p_1,p_2是A的对应特征值2的两个线性无关的特征向量，而p_3可由求解非齐次线性方程组$(2I-A)x=-p_2$得到。

先解$(2I-A)x=0$，得特征值2的两个线性无关的特征向量$\xi_1=(1,2,0)^T,\xi_2=(0,0,1)^T$.可取$p_1=\xi_1$，也可试取$p_2=\xi_2$，求解$(2I-A)x=-p_2$，可发现无解，需重新选择$p_2$.由于$k_1\xi_1+k_2\xi_2$仍是$(2I-A)x=0$的解，因此可取$p_2=k_1\xi_1+k_2\xi_2$,其中$k_1,k_2$待定。因

$$(2I-A,-p_2)=\begin{pmatrix} 2 & -1 & 0 & -k_1 \\ 4 & -2 & 0 & -2k_1 \\ 2 & -1 & 0 & -k_2 \end{pmatrix}\to\begin{pmatrix} 2 & -1 & 0 & -k_1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & k_1-k_2 \end{pmatrix},$$

知$k_1=k_2$时，$(2I-A)x=-p_2$有解，且通解为$x_2=2x_1+k_1$，取$k_1=1$，可得$p_2=(1,2,1)^T$，再在$x_2=2x_1+1$中令$x_3=0,x_1=0$,得$p_3=(0,1,0)^T$. 于是$P=\begin{pmatrix} 1 & 1 & 0 \\ 2 & 2 & 1 \\ 0 & 1 & 0 \end{pmatrix}$,且$P^{-1}AP=J_A$.

2.2.3 Jordan标准形的应用

Jordan标准形在矩阵理论的很多方面都有应用。这里主要给出Jordan标准形在求矩阵多项式这个特殊矩阵函数上的应用。

设$A\in C^{n\times n},f(\lambda)=a_m\lambda^m+\cdots+a_1\lambda+a_0$为一多项式，如何计算$f(A)$呢？设$A$的Jordan标准形为

$$J_A=\begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_s \end{pmatrix}.$$

则$A=PJ_AP^{-1}$，因$A^m=PJ_A^mP^{-1}$,所以

$$f(A)=Pf(J_A)P^{-1}=P\begin{pmatrix} f(J_1) & & \\ & \ddots & \\ & & f(J_s) \end{pmatrix}P^{-1}.$$

P 和P^{-1}为已知，只要算出$f(J_i)$，即可求出$f(A)$. 记$g_k(\lambda)=\lambda^k$,则任一$f(\lambda)$可看作$g_o(\lambda),\cdots,g_m(\lambda)$的线性组合，故只要求出$g_k(J_i)$即可。

设

$$J_i=\begin{pmatrix}\lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1\\ & & & \lambda_i\end{pmatrix}_{n_i\times n_i},$$

则

$$g_k(J_i)=J_i^k=\begin{pmatrix}\lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1\\ & & & \lambda_i\end{pmatrix}^k=(\lambda_i I_{n_i}+H)^k,$$

其中

$$H=\begin{pmatrix}0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1\\ & & & 0\end{pmatrix}=\begin{pmatrix}0 & I_{n_i-1}\\ 0 & 0\end{pmatrix},$$

注意到$\lambda_i I_{n_i}$与H可交换，且

$$H^r=\begin{pmatrix}0 & I_{n_i-r}\\ 0 & 0\end{pmatrix}(1\le r\le n_i-1),H^s=0\,(s\ge n_i),$$

故

$$J_i^k=(\lambda_i I_{n_i}+H)^k=\lambda_i^k I_{n_i}+C_k^1\lambda_i^{k-1}H+\cdots+C_k^{n_i-1}\lambda_i^{k-n_i+1}H^{n_i-1},$$

其中 $j\ge k+1$ 时，λ_i^{k-j} 记为零。 又

$$C_k^r\lambda_i^{k-r}=\frac{k(k-1)\cdots(k-r+1)}{r!}\lambda_i^{k-r}=\frac{1}{r!}\left(\lambda_i^k\right)^{(r)}=\frac{1}{r!}g_k^{(r)}(\lambda_i),$$

因此

$$J_i^k=\begin{pmatrix}g_k(\lambda_i) & g'_k(\lambda_i) & \cdots & \frac{1}{(n_i-1)!}g_k^{(n_i-1)}(\lambda_i)\\ & \ddots & \ddots & \vdots\\ & & \ddots & g'_k(\lambda_i)\\ & & & g_k(\lambda_i)\end{pmatrix}.$$

进而可得

$$f(J_i)=\begin{pmatrix}f(\lambda_i) & f'(\lambda_i) & \cdots & \frac{1}{(n_i-1)!}f^{(n_i-1)}(\lambda_i)\\ & \ddots & \ddots & \vdots\\ & & \ddots & f'(\lambda_i)\\ & & & f(\lambda_i)\end{pmatrix}.$$

例 2.5 设$A=\begin{pmatrix}1&1&0\\6&2&0\\0&1&-1\end{pmatrix}$，求$A^m$.

解：设$f(\lambda)=\lambda^m$，则$f(A)=A^m$.由例 2.3 知

$$J_A=\begin{pmatrix}-1&1&0\\0&-1&0\\0&0&4\end{pmatrix}=\begin{pmatrix}J_1&\\&J_2\end{pmatrix}.$$

可求得

$$P=\begin{pmatrix}0&-\frac{1}{2}&5\\0&1&15\\1&0&3\end{pmatrix},$$

$$f(J_1)=\begin{pmatrix}f(-1)&f'(-1)\\0&f(-1)\end{pmatrix}=\begin{pmatrix}(-1)^m&m(-1)^{m-1}\\0&(-1)^m\end{pmatrix},\ f(J_2)=(4^m).$$

于是

$$\begin{aligned}A^m&=f(A)=Pf(J_A)P^{-1}\\&=\begin{pmatrix}0&-\frac{1}{2}&5\\0&1&15\\1&0&3\end{pmatrix}\begin{pmatrix}(-1)^m&m(-1)^{m-1}&0\\0&(-1)^m&0\\0&0&4^m\end{pmatrix}\cdot\frac{1}{25}\begin{pmatrix}-6&-3&25\\-30&10&0\\2&1&0\end{pmatrix}\\&=\frac{1}{25}\begin{pmatrix}15\cdot(-1)^m+10\cdot4^m&5\cdot\left((-1)^{m-1}+4^m\right)&0\\30\left((-1)^{m+1}+4^m\right)&10\cdot(-1)^m+15\cdot4^m&0\\(6-30m)(-1)^{m+1}+6\cdot4^m&3\left((-1)^{m+1}+4^m\right)+10m(-1)^{m-1}&25(-1)^m\end{pmatrix}.\end{aligned}$$

2.3 酉相似标准形

本节主要讨论一个方阵在酉相似下的标准形问题。

2.3.1 正规矩阵对角化

定理 2.7 （Schur定理）设$A\in C^{n\times n}$，则存在酉矩阵U和上三角矩阵R，使得

$$U^HAU=U^{-1}AU=R=\begin{pmatrix}\lambda_1&r_{12}&\cdots&r_{1n}\\&\lambda_2&\cdots&r_{2n}\\&&\ddots&\vdots\\&&&\lambda_n\end{pmatrix},$$

即A可酉相似于上三角矩阵，显然$\lambda_1, \lambda_2, \cdots, \lambda_n$为$A$的全部特征值。而且适当选取$U$，可使$A$的特征值按任意指定的顺序排列。

Schur定理是矩阵理论中的一个重要定理，它是很多重要结论的理论基础。我们知道一般方阵在复数域上总能相似于它的Jordan标准形，而Jordan标准形是一个特殊的上三角矩阵，但那里的相似变换矩阵是一般的可逆矩阵。

Schur定理说明任一方阵酉相似于上三角矩阵，那么怎样的矩阵能够酉相似于对角矩阵呢？借助于正规矩阵，就能解决这个问题。

引理 2.2 若A为正规矩阵，则与A酉相似的矩阵仍为正规矩阵。

引理 2.3 设A为上三角的正规矩阵，则A为对角矩阵。

证明：设$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ & a_{22} & \cdots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix}$，由$A^HA=AA^H$，得

$$\begin{pmatrix} \bar{a}_{11} & & & \\ \bar{a}_{12} & \bar{a}_{22} & & \\ \vdots & \vdots & \ddots & \\ \bar{a}_{1n} & \bar{a}_{2n} & \cdots & \bar{a}_{nn} \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ & a_{22} & \cdots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ & a_{22} & \cdots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix} \begin{pmatrix} \bar{a}_{11} & & & \\ \bar{a}_{12} & \bar{a}_{22} & & \\ \vdots & \vdots & \ddots & \\ \bar{a}_{1n} & \bar{a}_{2n} & \cdots & \bar{a}_{nn} \end{pmatrix}.$$

乘开并比较两端的主对角元，可得

$$|a_{11}|^2=\sum_{i=1}^{n}|a_{1i}|^2, |a_{12}|^2+|a_{22}|^2=\sum_{i=2}^{n}|a_{2i}|^2, \cdots, \sum_{i=1}^{n}|a_{in}|^2=|a_{nn}|^2.$$

从而得出$a_{ij}=0\,(i<j)$，即A为对角矩阵。证毕。

定理 2.8 设$A\in C^{n\times n}$，则A酉相似于对角矩阵的充分必要条件为A是正规矩阵。

证明：必要性. 若矩阵A酉相似于对角矩阵$\Lambda=\mathrm{diag}\,(\lambda_1, \lambda_2, \cdots, \lambda_n)$，即存在酉矩阵$U$，使$U^HAU=\Lambda$，则

$$A^HA=\left(U\Lambda U^H\right)^H\left(U\Lambda U^H\right)=U\Lambda^H\Lambda U^H=U\Lambda\Lambda^HU^H=\left(U\Lambda U^H\right)\left(U\Lambda^HU^H\right)=AA^H,$$

即A是正规矩阵。

充分性. 若A为正规矩阵。由Schur定理知，A酉相似于一个上三角矩阵R，而据引理 2.2，可得 R也是正规矩阵，再由引理 2.3，知R为对角矩阵。于是 A 酉相似于对角矩阵。证毕。

推论 2.4 *设$A \in C^{n\times n}$是正规矩阵，则A与A^H可有相同的酉矩阵U，使得 U^HAU，U^HA^HU 为对角矩阵。*

推论 2.5 *设$A \in C^{n\times n}$是正规矩阵，$\lambda \in C$，$x \in C^n$，则下列条件等价：*

(1) *x 是 A 的属于特征值λ的特征向量：$Ax = \lambda x$；*

(2) *x 是 A^H 的属于特征值 $\bar{\lambda}$的特征向量：$A^Hx = \bar{\lambda}x$.*

证明：不妨设x是单位特征向量， 即$x^Hx = 1$.

(1)⇒(2). 依定理 2.8 和推论 2.4，存在酉矩阵$U \in C^{n\times n}$，其第1列是x，使

$$U^HAU = \Lambda, U^HA^HU = \Lambda^H = \bar{\Lambda},$$

其中$\Lambda = \mathrm{diag}(\lambda, \lambda_2, \cdots, \lambda_n)$. 从而得 $A^HU = U\bar{\Lambda}$，此等式两端的第1列分别是 A^Hx 与 $\bar{\lambda}x$，因此 $A^Hx = \bar{\lambda}x$.

(2)⇒(1). 将(1)⇒(2)的过程应用于A^H，可直接得到 $Ax = \left(A^H\right)^Hx = \overline{(\bar{\lambda})}x = \lambda x$. 证毕。

推论 2.6 *正规矩阵属于不同特征值的特征向量是正交的。*

证明：设λ, μ 是正规矩阵A的两个不同特征值， x， y 是对应的特征向量，即 $Ax = \lambda x$， $Ay = \mu y$, 由推论 2.5 知也有$A^Hy = \bar{\mu}y$. 于是

$$\lambda(x, y) = (\lambda x, y) = y^HAx = \left(A^Hy\right)^Hx = (\bar{\mu}y)^Hx = \mu y^Hx = \mu(x, y),$$

即$(\lambda - \mu)(x, y) = 0$，因$\lambda \neq \mu$，得$(x, y) = 0$. 所以 x 与 y 正交。 证毕。

线性代数中介绍的实对称矩阵可正交相似于一个对角矩阵的结论，正是定理 2.8 充分性的特例。而从推论 2.6 可知，求n阶正规矩阵A酉相似于对角矩阵的计算步骤也完全类似于实对称矩阵对角化的情形。介绍如下：

(1) 求出A的互异特征值$\lambda_1, \lambda_2, \cdots, \lambda_s$，其重数分别为$n_1, n_2, \cdots, n_s$，且$n_1 + n_2 + \cdots + n_s = n$；

(2) 对每个互异特征值λ_i，求出其对应的n_i个线性无关的特征向量 $\xi_{i1}, \cdots, \xi_{in_i}$, $i = 1, 2, \cdots, s$；

(3) 分别将$\xi_{i1}, \cdots, \xi_{in_i}$标准正交化得$\eta_{i1}, \cdots, \eta_{in_i}, i = 1, \cdots, s$；

(4) 令$U = (\eta_{11}, \cdots, \eta_{1n_1}, \cdots, \eta_{s1}, \cdots, \eta_{sn_s})$.

则U为酉矩阵，且

$$U^HAU = U^{-1}AU = \mathrm{diag}(\lambda_1, \cdots, \lambda_1, \cdots, \lambda_s, \cdots, \lambda_s) = \mathrm{diag}(\lambda_1 I_{n_1}, \cdots, \lambda_s I_{n_s}).$$

例 2.6 已知 $A=\begin{pmatrix} 7 & 5i & -5-5i \\ -5i & 7 & 5-5i \\ -5+5i & 5+5i & 2 \end{pmatrix}$，试问是否存在酉矩阵$U$，使得$U^HAU$为对角矩阵。若存在，求这样的酉矩阵$U$和$U^{-1}AU$.

解：显然有 $A^H=A$，即A是Hermite矩阵，从而是正规矩阵，所以这样的酉矩阵是存在的。下求酉矩阵 U.

(1) 求A的特征值. 从

$$\det(\lambda I-A)=\lambda^3-16\lambda^2-48\lambda+1152=(\lambda-12)^2(\lambda+8)$$

得A的特征值$\lambda_1=\lambda_2=12,\lambda_3=-8$.

(2) 求特征向量. 对$\lambda_1=\lambda_2=12$，求解方程组$(12I-A)\boldsymbol{x}=0$，得特征向量为

$$\xi_1=(i,1,0)^T,\quad \xi_2=(1+i,0,-1)^T.$$

对$\lambda_3=-8$，解方程组$(-8I-A)\boldsymbol{x}=0$，得特征向量

$$\xi_3=(1,i,1-i)^T.$$

(3) 正交化. 由Schmidt正交化公式得

$$\beta_1=\xi_1=\begin{pmatrix} i \\ 1 \\ 0 \end{pmatrix},\quad \beta_2=\xi_2-\frac{(\xi_2,\beta_1)}{(\beta_1,\beta_1)}\beta_1=\begin{pmatrix} \frac{1+i}{2} \\ \frac{-1+i}{2} \\ -1 \end{pmatrix}.$$

(4) 单位化. 将 β_1,β_2,ξ_3 单位化得

$$\eta_1=\frac{\beta_1}{\|\beta_1\|}=\frac{1}{\sqrt{2}}\begin{pmatrix} i \\ 1 \\ 0 \end{pmatrix},\ \eta_2=\frac{\beta_2}{\|\beta_2\|}=\frac{1}{\sqrt{2}}\begin{pmatrix} \frac{1+i}{2} \\ \frac{-1+i}{2} \\ -1 \end{pmatrix},\ \eta_3=\frac{\xi_3}{\|\xi_3\|}=\frac{1}{2}\begin{pmatrix} 1 \\ i \\ 1-i \end{pmatrix}.$$

令 $U=(\eta_1,\eta_2,\eta_3)=\begin{pmatrix} \frac{i}{\sqrt{2}} & \frac{1+i}{2\sqrt{2}} & \frac{1}{2} \\ \frac{1}{\sqrt{2}} & \frac{-1+i}{2\sqrt{2}} & \frac{i}{2} \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1-i}{2} \end{pmatrix}$，则$U$为所求的酉矩阵，且

$$U^{-1}AU=\begin{pmatrix} 12 & & \\ & 12 & \\ & & -8 \end{pmatrix}.$$

例 2.6 中Hermite矩阵A的特征值均为实数，这是不是巧合呢？下面的推论给出了回答。

推论 2.7 设$A \in C^{n\times n}$，其特征值为$\lambda_1, \lambda_2, \cdots, \lambda_n$.

(1) 若A=A^H，即A为Hermite阵，则A的特征值均为实数；

(2) 若$A = -A^H$，即A为反Hermite阵，则A的特征值为零或纯虚数；

(3) 若A为酉矩阵，则A的特征值的模为1。

证明：(1)$A = A^H$，则A为正规矩阵，于是存在n阶酉矩阵U，使得

$$U^H AU = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} = \Lambda. \tag{2.7}$$

且

$$\Lambda^H = \left(U^H AU\right)^H = U^H A^H U = U^H AU = \Lambda,$$

从而$\lambda_i = \bar{\lambda}_i, i = 1, \cdots, n$. 故$\lambda_i\,(i = 1, \cdots, n)$均为实数。

(2) 若$A = -A^H$，仿(1)可推得$\bar{\lambda}_i = -\lambda_i$，即$\mathrm{Re}\,(\lambda_i) = 0, i = 1, \cdots, n$. 可见$\lambda_i$为零或纯虚数。

(3)若A为酉矩阵，则A亦正规。于是上面的(2.7)式成立，且有

$$\Lambda^H \Lambda = U^H A^H U U^H AU = U^H A^H AU = U^H U = I.$$

从而有$\lambda_i \bar{\lambda}_i = |\lambda_i|^2 = 1$，即$|\lambda_i| = 1, i = 1, \cdots, n$. 证毕。

推论 2.8 设$A = C^{n\times n}$是正规矩阵，则

(1) A是Hermite矩阵的充要条件是A的特征值均为实数；

(2) A是反Hermite矩阵的充要条件是A的特征值为零或纯虚数；

(3) A是酉矩阵的充要条件是A的特征值的模均为1。

2.3.2 正定矩阵

定义 2.10 设$A \in C^{n\times n}$为Hermite矩阵，如果对任意$0 \neq \boldsymbol{x} \in C^n$都有

$$\boldsymbol{x}^H A\boldsymbol{x} > 0 \ (\boldsymbol{x}^H A\boldsymbol{x} \geq 0)$$

则称A为Hermite正定矩阵(半正定矩阵)。

定理 2.9 设$A \in C^{n\times n}$为Hermite矩阵，则下列命题等价：

(1) A是正定矩阵；

(2) 对任意n阶可逆矩阵P，$P^H AP$为Hermite正定矩阵；

(3) A的特征值均为正数；

(4) 存在n阶可逆矩阵P，使$A = P^H P$.

证明：(1)⇒(2). 因A是Hermite阵，故P^HAP也是Hermite矩阵。对任意 $\boldsymbol{x} \in C^n$ 且 $\boldsymbol{x} \neq 0$，则因P可逆，从而$P\boldsymbol{x} \neq 0$. 于是

$$\boldsymbol{x}^H \left(P^HAP\right) \boldsymbol{x} = (P\boldsymbol{x})^H (P\boldsymbol{x}) > 0.$$

由定义 2.10知 P^HAP是Hermite正定矩阵。

(2)⇒(3). 对Hermite矩阵A，由定理 2.8知存在酉矩阵U，使

$$U^HAU = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n), \tag{2.8}$$

其中 $\lambda_1, \lambda_2, \cdots, \lambda_n$ 是 A 的特征值。由(2)知 $\operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$ 正定，从而 $\lambda_1, \cdots, \lambda_n$ 均为正数。

(3)⇒(4). 由(2.8)得

$$\begin{aligned} A &= U\operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n) U^H \\ &= U\operatorname{diag}\left(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_n}\right) \cdot \operatorname{diag}\left(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_n}\right) U^H = P^HP, \end{aligned}$$

其中$P = \operatorname{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_n}\right) U^H \in C^{n\times n}$可逆。

(4)⇒(1). 对任意$0 \neq \boldsymbol{x} \in C^n$，均有$P\boldsymbol{x} \neq 0$，从而

$$\boldsymbol{x}^HA\boldsymbol{x} = \boldsymbol{x}^HP^HP\boldsymbol{x} = (P\boldsymbol{x})^H (P\boldsymbol{x}) = (P\boldsymbol{x}, P\boldsymbol{x}) > 0.$$

故A是正定矩阵。证毕。

推论 2.9 设$A \in C^{n\times n}$为Hermite正定矩阵，则行列式 $\det A > 0$.

类似于定理 2.9，有Hermite半正定矩阵的如下结论。

定理 2.10 设$A \in C^{n\times n}$为Hermite矩阵，则下列命题等价：

(1) A是Hermite半正定矩阵；

(2) 对任意n阶可逆矩阵P，P^HAP 为Hermite半正定矩阵；

(3) A的特征值均为非负实数；

(4) 存在矩阵$P \in C^{n\times n}$，使$A = P^HP$.

Hermite正定矩阵也有类似于实对称正定矩阵的顺序主子式判别法。即

定理 2.11 设$A = (a_{ij}) \in C^{n\times n}$是Hermite矩阵，又设

$$A_k = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix}, \quad \Delta_k = \det A_k, \quad k = 1, 2, \cdots, n$$

分别称之为A的k阶顺序主子阵和顺序主子式，则A正定的充分必要条件为A的n个顺序主子式均为正数，即$\Delta_k = \det A_k > 0, k = 1, 2, \cdots, n$.

须注意的是，即使 Hermite 矩阵 A 的所有顺序主子式均非负，也不能推出 A 是 Hermite 半正定矩阵。例如

$$A=\begin{pmatrix}1 & -1 & 0\\ -1 & 1 & 0\\ 0 & 0 & -1\end{pmatrix}$$

是Hermite矩阵，易证A的3个顺序主子式均非负，但A不是Hermite半正定矩阵。不过还是有下面的结论。

定理 2.12 设$A\in C^{n\times n}$是Hermite矩阵，则A半正定的充分必要条件为A的所有主子式(所谓主子式就是行列式det$A$ 中行指标与列指标相同的子式)均非负。

下面的定理在后续内容的讨论中是有用的。

定理 2.13 设$A\in C^{m\times n}$，则

(1) A^HA 和 AA^H 的特征值均为非负实数；

(2) A^HA与AA^H的非零特征值相同；

(3) $\operatorname{rank}\left(A^HA\right)=\operatorname{rank}\left(AA^H\right)=\operatorname{rank}A$.

证明：(1) 因$\left(A^HA\right)^H=A^HA$，即 A^HA 是Hermite矩阵，且对任意 $0\neq\boldsymbol{x}\in C^n$，有

$$\boldsymbol{x}^H\left(A^HA\right)\boldsymbol{x}=(A\boldsymbol{x})^H(A\boldsymbol{x})=(A\boldsymbol{x},A\boldsymbol{x})\geq 0.$$

故由定义 2.10 知 A^HA 是Hermite半正定矩阵，再由定理 2.10 可得 A^HA 的特征值均为非负实数。同理 AA^H 的特征值均为非负实数。

(2) 设$\lambda\neq 0$为 A^HA 的特征值，$\boldsymbol{x}$ 为对应的特征向量，则$A^HA\boldsymbol{x}=\lambda x$, $Ax\neq 0$. 于是

$$AA^H(A\boldsymbol{x})=A\left(A^HA\boldsymbol{x}\right)=A(\lambda\boldsymbol{x})=\lambda(A\boldsymbol{x}).$$

即λ 是 AA^H 的非零特征值。同理 AA^H 的非零特征值也是 A^HA 的非零特征值。

(3) 由$A\boldsymbol{x}=0$，有 $A^HA\boldsymbol{x}=0$. 反之，若 $A^HA\boldsymbol{x}=0$，则

$$0=\boldsymbol{x}^HA^HA\boldsymbol{x}=(A\boldsymbol{x})^H(A\boldsymbol{x})=(A\boldsymbol{x},A\boldsymbol{x}).$$

故 $A\boldsymbol{x}=0$. 此即说明方程组 $A\boldsymbol{x}=0$ 与 $A^HA\boldsymbol{x}=0$ 同解，从而它们解空间的维数相同，即

$$n-\operatorname{rank}A=n-\operatorname{rank}\left(A^HA\right),$$

所以 $\operatorname{rank}A=\operatorname{rank}\left(A^HA\right)$. 将此式中$A$用 A^H 代替，可得

$$\operatorname{rank}\left[\left(A^H\right)^HA^H\right]=\operatorname{rank}\left(AA^H\right)=\operatorname{rank}A^H=\operatorname{rank}A.$$

综上可得$\operatorname{rank}\left(A^HA\right)=\operatorname{rank}\left(AA^H\right)=\operatorname{rank}A$. 证毕。

2.4 特征值的隔离

特征值的包含区域–盖尔(Gerschgorin)圆盘，是对特征值在复平面上的变化范围作出较精确的估计。

2.4.1 盖尔圆定理

定义 2.11 设 $A=(a_{ij})\in \boldsymbol{C}^{n\times n}$，令

$$R_i=\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}| \quad (i=1,2,\cdots,n)$$

称复平面上的圆盘

$$G_i=\{z\in C\,|\,|z-a_{ii}|\leq R_i\} \quad (i=1,2,\cdots,n)$$

为矩阵A的第i个盖尔圆，并称R_i为盖尔圆G_i的半径。

定理 2.14 设$A=(a_{ij})\in C^{n\times n}$，则$A$的所有特征值都在它的$n$个盖尔圆的并集中，即$\lambda\in\bigcup_{i=1}^{n}G_i$，其中$\lambda$为$A$的任一特征值。

证明：设λ是A的任意特征值，$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^T$ 为 A 对应于 λ 的特征向量，且$|x_i|=\|x\|_\infty=1$。由于$(A\boldsymbol{x})_i=\lambda x_i$，即

$$\sum_{j=1}^{n}a_{ij}x_j=\lambda x_i.$$

则有

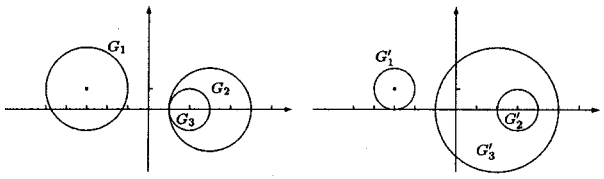
$$(\lambda-a_{ii})x_i=\sum_{\substack{j=1\\ j\neq i}}^{n}a_{ij}x_j.$$

注意到$|x_j|\leq|x_i|=1\ (j\neq i)$，则由上式得

$$|\lambda-a_{ii}|\leq\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\,|x_j|\leq\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|$$

于是$\lambda\in G_i$。证毕。

注意到A与A^T的特征值相同，于是由定理 2.14 可知 A 的全体特征值都在 A^T 的 n 个盖尔圆构成的并集中，同时称 A^T 的盖尔圆为 A 的列盖尔圆。



(a) 行盖尔圆 (b) 列盖尔圆

图 2.1: 盖尔圆

例 2.7 估计矩阵

$$A=\begin{pmatrix} -3+i & 0 & 2 \\ -1 & 3 & i \\ 0 & -1 & 2 \end{pmatrix}$$

的特征值分布范围。

解：矩阵A的3个盖尔圆为

$$G_1:|z+3-i|\leq 2,\ G_2:|z-3|\leq 2,\ G_3:|z-2|\leq 1.$$

故A的3个特征值都在$\bigcup_{i=1}^{3} G_i$之中 (见图 2.1(a))。

A 的3个列盖尔圆为

$$G_1':|z+3-i|\leq 1,\ G_2':|z-3|\leq 1,\ G_3':|z-2|\leq 3$$

故A的3个特征值也都在$\bigcup_{i=1}^{3} G_i'$之中 (见图 2.1(b))。

定理 2.14只说明A的n个特征值都在n个盖尔圆的并集中，但是并没有说明特征值的具体分布情况。进一步，可有特征值分布更精确的结果。

定义 2.12 在矩阵A的盖尔圆中，相交在一起的盖尔圆构成的最大连通区域称为一个连通部分，一个孤立的盖尔圆也是一个连通部分。

定理 2.15 若矩阵A的某一连通部分S由A的k个盖尔圆构成，则S中有且仅有A的k个特征值(盖尔圆相重时重复计数，重特征值按其重数计算)。

根据定理 2.15，例 2.7中A的特征值在G_1中有一个，而连通部分$G_2 \bigcup G_3$中有两个；或者一个在G_1'中，两个在$G_2' \bigcup G_3'$中。

值得注意的是，特征值在两个或两个以上的盖尔圆构成的连通部分中分布不一定是平均的，可能在某个盖尔圆中有几个特征值，而在某些盖尔圆中无特征值。例如

$$A=\begin{pmatrix} 10 & -8 \\ 5 & 0 \end{pmatrix}$$

其特征值为：$\lambda_1 = 5 - i\sqrt{15}$, $\lambda_2 = 5 + i\sqrt{15}$。两个盖尔圆为：$G_1 : |z - 10| \leq 8$, $G_2 : |z| \leq 5$。由定理 2.15知特征值在连通部分$G_1 \bigcup G_2$中，但实际上可以发现A 的特征值都在G_1中。

2.4.2 特征值的隔离

下面应用盖尔圆定理研究特征值的隔离问题。

由定理 2.15知，矩阵 A的每个孤立盖尔圆中恰有A的一个特征值。所以，当若干盖尔圆相交时，希望能够采用一些方法将盖尔圆隔离以获得特征值的分布。通常可采用如下两种方法：

(1) 结合矩阵A的列盖尔圆来研究A的特征值分布。

例 2.8 *应用盖尔圆定理隔离*

$$A = \begin{pmatrix} 20 & 3 & 2 \\ 2 & 10 & 4 \\ 4 & 0.5 & 0 \end{pmatrix}$$

的特征值。

解：矩阵A的3个盖尔圆为

$$G_1 : |z - 20| \leq 5, G_2 : |z - 10| \leq 6, G_3 : |z| \leq 4.5.$$

G_1, G_2, G_3相交。而A的3个列盖尔圆为

$$G_1' : |z - 20| \leq 6, G_2' : |z - 10| \leq 3.5, G_3' : |z| \leq 6$$

它们相互分离。故在G_1', G_2', G_3'中各有A的一个特征值。

(2) 利用相似变换保持特征值不变的特性，也可以实现特征值隔离。

设$A = (a_{ij}) \in C^{n\times n}$，$D = \mathrm{diag}(d_1, d_2, \cdots, d_n)$，其中$d_i > 0\ (i = 1, 2, \cdots, n)$，则$A$与$B = D^{-1}AD$相似且有相同特征值。注意到$B = D^{-1}AD$的圆盘半径为$\sum\limits_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}d_j}{d_i}\right|$，因此适当选取 D 的元素值，可以使 B 的某个圆盘半径相对减少。一般地，若要使 B 的第 i 个盖尔圆半径放大，其余盖尔圆适量缩小，可取$d_i < 1$，其他元素值取1；相反，若要使 B 的第 i 个盖尔圆半径缩小，其余盖尔圆适量放大，可取$d_i > 1$，其他元素值取 1。

例 2.9 *应用盖尔圆定理隔离*

$$A = \begin{pmatrix} 2 & 2 & -1 \\ 1 & 10 & -1 \\ 8 & 2 & 20 \end{pmatrix}$$

的特征值。

解：矩阵A的3个盖尔圆为

$$G_1:|z-2|\leq 3,\ G_2:|z-10|\leq 2,\ G_3:|z-20|\leq 10.$$

易见，G_2与G_3相交，而G_1孤立。选取$D=\mathrm{diag}(\frac{1}{2},1,1)$，则

$$B=D^{-1}AD=\begin{pmatrix}2 & 4 & -2\\ 0.5 & 10 & -1\\ 4 & 2 & 20\end{pmatrix}.$$

矩阵B的3个盖尔圆为

$$G_1':|z-2|\leq 6,\ G_2':|z-10|\leq 1.5,\ G_3':|z-20|\leq 6,$$

它们相互分离。故在G_1'、G_2'、G_3'中各有A的一个特征值。因为A的第1个盖尔圆G_1孤立，其中含A的一个特征值，所以G_1'中的特征值必在G_1之中。因此，A的三个特征值分别在盖 尔圆G_1，G_2'及G_3'之中。进一步，由于A是实矩阵，其特征值(如果是复数，则其共轭复数$\bar{\lambda}$一定也是A的特征值，而盖尔圆G_1'、G_2'、G_3'关于实轴对称，所以如果有$\lambda\in G_i'$，则也有$\bar{\lambda}\in G_i'$，因而A的特征值是实数。故在区间

$$[-1,\ 5],\ [8.5,\ 11.5],\ [14,\ 26]$$

中各有A的一个特征值。

需要说明的是，并不是任意具有互异特征值的矩阵都能用上述两种方法分隔其特征值，如对角线上有相同元素的矩阵。

2.5 特征值的幂迭代法、逆幂迭代法

应当指出，五次或五次以上的多项式方程一般是没有公式求解的。所以对于阶数较大的矩阵，其特征值的计算是非常困难的，因而就要研究特征值的各种近似求法。本节给出的幂迭代法正是求特征值近似值的基本方法。

在许多实际应用中，往往不需要知道矩阵的全部特征值和特征向量，而仅要求得到矩阵的按模最大的特征值（或称为矩阵的主特征值）和相应的特征向量，如线性方程组迭代解法的收敛性问题。幂迭代法（或简称幂法）就是解决此类问题的一种有效方法。

2.5.1 幂迭代法

幂法的思想主要基于如下结论。

定理 2.16 设$A \in R^{n\times n}$，其特征值$\lambda_i\,(i=1,2,\cdots,n)$满足

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|, \tag{2.9}$$

相应的n个线性无关的特征向量为$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$。任取一个非零向量$\boldsymbol{v}_0 \in R^n$，用$A$构造向量序列

$$\boldsymbol{v}_{k+1} = A\boldsymbol{v}_k, \quad k=0,1,2,\cdots \tag{2.10}$$

则有

$$\lim_{k\to\infty} \frac{(\boldsymbol{v}_k)_i}{(\boldsymbol{v}_{k-1})_i} = \lambda_1, \quad i=1,2,\cdots,n$$

式中，$(\boldsymbol{v}_k)_i$表示向量$\boldsymbol{v}_k$的第 i 个分量。

证明：由递推公式(2.10)，有$\boldsymbol{v}_k = A\boldsymbol{v}_{k-1} = A(A\boldsymbol{v}_{k-2}) = A^2\boldsymbol{v}_{k-2} = \cdots = A^k\boldsymbol{v}_0$。又由于$n$个线性无关的特征向量$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$构成$R^n$的一组基，所以$\boldsymbol{v}_0$可以唯一地表示为

$$\boldsymbol{v}_0 = \alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 + \cdots + \alpha_n\boldsymbol{x}_n \quad (\alpha_1 \neq 0)$$

因 $A\boldsymbol{x}_i = \lambda_i\boldsymbol{x}_i \Rightarrow A^k\boldsymbol{x}_i = \lambda_i^k\boldsymbol{x}_i\,(i=1,2,\cdots,n)$，所以

$$\begin{aligned}
\boldsymbol{v}_k &= A^k\boldsymbol{v}_0 = \alpha_1A^k\boldsymbol{x}_1 + \alpha_2A^k\boldsymbol{x}_2 + \cdots + \alpha_nA^k\boldsymbol{x}_n \\
&= \alpha_1\lambda_1^k\boldsymbol{x}_1 + \alpha_2\lambda_2^k\boldsymbol{x}_2 + \cdots + \alpha_n\lambda_n^k\boldsymbol{x}_n \\
&= \lambda_1^k(\alpha_1\boldsymbol{x}_1 + \sum_{i=2}^{n}\alpha_i\left(\frac{\lambda_i}{\lambda_1}\right)^k\boldsymbol{x}_i)
\end{aligned}$$

由于$\left|\frac{\lambda_i}{\lambda_1}\right| < 1\,(i=2,3,\cdots,n)$，故对足够大的$k$，有

$$\boldsymbol{v}_k = \lambda_1^k(\alpha_1\boldsymbol{x}_1 + \varepsilon_k),\ \boldsymbol{v}_{k-1} = \lambda_1^{k-1}(\alpha_1\boldsymbol{x}_1 + \varepsilon_{k-1})$$

式中$\varepsilon_k \to 0\ (k\to\infty)$。当$(\boldsymbol{x}_1)_i \neq 0$时，得

$$\lim_{k\to\infty}\frac{(\boldsymbol{v}_k)_i}{(\boldsymbol{v}_{k-1})_i} = \lambda_1\lim_{k\to\infty}\frac{\alpha_1(\boldsymbol{x}_1)_i + (\varepsilon_k)_i}{\alpha_1(\boldsymbol{x}_1)_i + (\varepsilon_{k-1})_i} = \lambda_1$$

证毕。

由于$\boldsymbol{v}_{k+1} = A\boldsymbol{v}_k$，而$\boldsymbol{v}_{k+1} \approx \lambda_1\boldsymbol{v}_k$，所以，$\boldsymbol{v}_k$可以作为与$\lambda_1$相应的特征向量的近似。上述这种由已知非零向量$\boldsymbol{v}_0$及矩阵的乘幂$A^k$构造向量序列$\{\boldsymbol{v}_k\}$以计算 A的按模最大的特征值及相应特征向量的方法称为乘幂法，简称为幂法。它实际上是一种迭代法，且迭代的收敛速度主要取决于$\left|\frac{\lambda_2}{\lambda_1}\right|$的大小，$\left|\frac{\lambda_2}{\lambda_1}\right|$越小时，收敛越快，越接近于1时，收敛越慢。另外，值得注意的是：

1) 因$\boldsymbol{x}_1 \neq 0$，故$\boldsymbol{x}_1$的分量不会为零，即存在$(\boldsymbol{x}_1)_i \neq 0$。至于$\alpha_1 \neq 0$的情况，通常可取$\boldsymbol{v}_0 = (1,1,\cdots,1)^T$，如果初始向量$\boldsymbol{v}_0$选择不当，以致使$\alpha_1 = 0$，上述结果理论不成立，但由于计算中的舍入误差，经过若干步以后，可有$\alpha_1 \neq 0$。

2) 条件(2.9)意味着A的主特征值是单根。若(2.9)不满足，将有下列不同的情况：

(1) 如果$\lambda_1=\lambda_2=\cdots=\lambda_r$，且$|\lambda_r|>|\lambda_{r+1}|\geq\cdots\geq|\lambda_n|$，则仿定理 2.16 的证明，对任意初始向量$\boldsymbol{v}_0=\sum_{i=1}^{n}\alpha_i\boldsymbol{x}_i$ ($\alpha_1,\cdots,\alpha_r$)，可得

$$\lim_{k\to\infty}\frac{(\boldsymbol{v}_k)_i}{(\boldsymbol{v}_{k-1})_i}=\lambda_1\,,\quad \lim_{k\to\infty}\frac{\boldsymbol{v}_k}{\lambda_1^k}=\sum_{i=1}^{r}\alpha_i\boldsymbol{x}_i$$

$\boldsymbol{v}_k$ 仍可作为与λ_1相应的特征向量的近似。

(2) 如$|\lambda_1|=|\lambda_2|$，且$\lambda_1=-\lambda_2$，$|\lambda_1|>|\lambda_3|\geq\cdots\geq|\lambda_n|$，则对任意初始向量

$$\boldsymbol{v}_0=\alpha_1\boldsymbol{x}_1+\alpha_2\boldsymbol{x}_2+\cdots+\alpha_n\boldsymbol{x}_n\ (\alpha_1\neq0\,,\ \alpha_2\neq0)$$

可得

$$\frac{(\boldsymbol{v}_{k+2})_i}{(\boldsymbol{v}_k)_i}\to\lambda_1^2\ (k\to\infty)$$

或

$$\frac{(\boldsymbol{v}_{k+1})_i}{(\boldsymbol{v}_{k-1})_i}\to\lambda_1^2\ (k\to\infty)$$

可证明对应于λ_1的A的近似特征向量为$\boldsymbol{v}_{k+1}+\lambda_1\boldsymbol{v}_k$，对应于$\lambda_2$的为$\boldsymbol{v}_{k+1}-\lambda_1\boldsymbol{v}_k$。

(3) 如$|\lambda_1|=|\lambda_2|$，但λ_1，λ_2为一对共轭复特征值，且$|\lambda_1|>|\lambda_3|\geq\cdots\geq|\lambda_n|$，则对任意初始向量

$$\boldsymbol{v}_0=\alpha_1\boldsymbol{x}_1+\alpha_2\boldsymbol{x}_2+\cdots+\alpha_n\boldsymbol{x}_n\ (\alpha_1\neq0\,,\ \alpha_2\neq0)$$

可得相继的3个迭代向量$\boldsymbol{v}_k,\boldsymbol{v}_{k+1},\boldsymbol{v}_{k+2}$近似线性相关，即有

$$\boldsymbol{v}_{k+2}+p\boldsymbol{v}_{k+1}+q\boldsymbol{v}_k\approx0.$$

选定2个分量代入上述关系，构成二阶线性方程组，解出p，q，然后再验证 p, q 对$\boldsymbol{v}_k,\boldsymbol{v}_{k+1},\boldsymbol{v}_{k+2}$的其余分量是否也满足线性关系方程，若满足，则再将 p, q 代入方程 $\lambda^2+p\lambda+q=0$就可求得λ_1,λ_2，对应于λ_1,λ_2的近似特征向量为 $\boldsymbol{v}_{k+1}-\lambda_2\boldsymbol{v}_k$，$\boldsymbol{v}_{k+1}-\lambda_1\boldsymbol{v}_k$。

根据以上的讨论，在用幂法进行计算时，当k充分大时，检查是否出现下列三种情况之一：

(Ⅰ)$\frac{(\boldsymbol{v}_{k+1})_i}{(\boldsymbol{v}_k)_i}$ 趋近于某一常数；

(Ⅱ)$\frac{(\boldsymbol{v}_{k+2})_i}{(\boldsymbol{v}_k)_i}$ 趋近于某一常数；

(Ⅲ) $\boldsymbol{v}_k$ 的波动不规律，但相继的三个向量 $\boldsymbol{v}_k,\boldsymbol{v}_{k+1},\boldsymbol{v}_{k+2}$之间满足$\boldsymbol{v}_{k+2}+p\boldsymbol{v}_{k+1}+q\boldsymbol{v}_k\approx0$。

若是，就可求出 A的按模最大的特征值和相应的特征向量。但上述三种情况也有可能均不出现，此时就需要考虑其它数值解法。

另外，由$\boldsymbol{v}_k = \lambda_1^k(\alpha_1\boldsymbol{x}_1 + \varepsilon_k)$可知，当$k \to \infty$时，若$|\lambda_1| > 1$，则$\boldsymbol{v}_k$的分量趋于无穷大；若 $|\lambda_1| < 1$，则$\boldsymbol{v}_k$的分量又会趋于零，分量变化太大，使得计算过程可能产生溢出。为了克服这一不足，通常采用规范化迭代方式。具体做法为：把迭代向量$\boldsymbol{v}_k$的最大分量化为1，于是得幂迭代法的计算公式为：

$$\begin{cases} \boldsymbol{u}_0 = \boldsymbol{v}_0 \neq 0 \\ \boldsymbol{v}_k = A\boldsymbol{u}_{k-1} \\ m_k = \max(\boldsymbol{v}_k) \\ \boldsymbol{u}_k = \boldsymbol{v}_k / m_k \qquad k = 1, 2, \cdots \end{cases}$$

式中用 $\max(\boldsymbol{v})$ 表示向量 $\boldsymbol{v}$ 的绝对值最大的分量，从而有

迭代	规范化
$\boldsymbol{v}_1 = A\boldsymbol{u}_0 = A\boldsymbol{v}_0,$	$\boldsymbol{u}_1 = \dfrac{\boldsymbol{v}_1}{\max(\boldsymbol{v}_1)} = \dfrac{A\boldsymbol{v}_0}{\max(A\boldsymbol{v}_0)}$
$\boldsymbol{v}_2 = A\boldsymbol{u}_1 = \dfrac{A^2\boldsymbol{v}_0}{\max(A\boldsymbol{v}_0)},$	$\boldsymbol{u}_2 = \dfrac{\boldsymbol{v}_2}{\max(\boldsymbol{v}_2)} = \dfrac{A^2\boldsymbol{v}_0}{\max(A^2\boldsymbol{v}_0)}$
$\vdots$	$\vdots$
$\boldsymbol{v}_k = A\boldsymbol{u}_{k-1} = \dfrac{A^k\boldsymbol{v}_0}{\max(A^{k-1}\boldsymbol{v}_0)},$	$\boldsymbol{u}_k = \dfrac{\boldsymbol{v}_k}{\max(\boldsymbol{v}_k)} = \dfrac{A^k\boldsymbol{v}_0}{\max(A^k\boldsymbol{v}_0)}$
$\vdots$	$\vdots$

于是得

$$\boldsymbol{u}_k = \frac{A^k\boldsymbol{v}_0}{\max(A^k\boldsymbol{v}_0)} = \frac{\lambda_1^k(\alpha_1 x_1 + \sum_{i=2}^{n}\alpha_i(\frac{\lambda_i}{\lambda_1})^k\boldsymbol{x}_i)}{\max(\lambda_1^k(\alpha_1\boldsymbol{x}_1 + \sum_{i=2}^{n}\alpha_i(\frac{\lambda_i}{\lambda_1})^k\boldsymbol{x}_i))}$$

$$= \frac{\alpha_1\boldsymbol{x}_1 + \sum_{i=2}^{n}\alpha_i(\frac{\lambda_i}{\lambda_1})^k\boldsymbol{x}_i}{\max(\alpha_1\boldsymbol{x}_1 + \sum_{i=2}^{n}\alpha_i(\frac{\lambda_i}{\lambda_1})^k\boldsymbol{x}_i)} \to \frac{\boldsymbol{x}_1}{\max(\boldsymbol{x}_1)} \quad (k \to \infty)$$

同理，可得

$$m_k = \max(\boldsymbol{v}_k) = \max\left[\frac{\lambda_1^k(\alpha_1\boldsymbol{x}_1 + \sum_{i=2}^{n}\alpha_i(\frac{\lambda_i}{\lambda_1})^k\boldsymbol{x}_i)}{\max(\lambda_1^{k-1}(\alpha_1\boldsymbol{x}_1 + \sum_{i=2}^{n}\alpha_i(\frac{\lambda_i}{\lambda_1})^{k-1}\boldsymbol{x}_i))}\right]$$

$$= \frac{\lambda_1\max(\alpha_1\boldsymbol{x}_1 + \sum_{i=2}^{n}\alpha_i(\frac{\lambda_i}{\lambda_1})^k\boldsymbol{x}_i)}{\max(\alpha_1\boldsymbol{x}_1 + \sum_{i=2}^{n}\alpha_i(\frac{\lambda_i}{\lambda_1})^{k-1}\boldsymbol{x}_i)} \to \lambda_1 \quad k \to \infty$$

表 2.1: 计算结果

k	$\boldsymbol{u}_k^T$	m_k
0	(0, 0, 1)	
1	(0.5, 1.0, 0.25)	4
2	(0.5, 1.0, 0.8611)	9
3	(0.5, 1.0, 0.7306)	11.44
4	(0.5, 1.0, 0.7535)	10.9224
5	(0.5, 1.0, 0.7493)	11.0140
6	(0.5, 1.0, 0.7501)	10.9972
7	(0.5, 1.0, 0.7500)	11.0004
8	(0.5, 1.0, 0.7500)	11.0000

这说明m_k收敛到λ_1，u_k收敛到规范化了的特征向量。于是有

定理 2.17 设$A \in R^{n\times n}$，其特征值$\lambda_i\,(i=1,2,\cdots,n)$满足

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$$

相应的n个线性无关的特征向量为$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$。对任意的非零向量$\boldsymbol{v}_0 = \boldsymbol{u}_0$ $(\alpha_1 \neq 0)$，按下列方法构造向量序列

$$\begin{cases} \boldsymbol{u}_0 = \boldsymbol{v}_0 \neq 0 \\ \boldsymbol{v}_k = A\boldsymbol{u}_{k-1} \\ m_k = \max(\boldsymbol{v}_k) \\ \boldsymbol{u}_k = \boldsymbol{v}_k/m_k \qquad k = 1,2,\cdots \end{cases} \tag{2.11}$$

则有

$$\lim_{k\to\infty} \max(\boldsymbol{v}_k) = \lambda_1,\ \lim_{k\to\infty} \boldsymbol{u}_k = \frac{\boldsymbol{x}_1}{\max(\boldsymbol{x}_1)}$$

例 2.10 计算矩阵

$$A = \begin{pmatrix} 2 & 3 & 2 \\ 10 & 3 & 4 \\ 3 & 6 & 1 \end{pmatrix}$$

的主特征值及对应的特征向量。当$\|\boldsymbol{u}_k - \boldsymbol{u}_{k+1}\| \leq 10^{-4}$时停止计算。

解：A 的特征值为 11，-3，-2。取 $v_0 = (0,0,1)^T$，按 (2.11) 计算，结果见表 2.1。

2.5.2 幂迭代法的加速

从上面的分析可知，序列$\{m_k\}$是线性收敛的，通常要使用加速手段来提高收敛速度。加速方法有多种，这里主要给出原点平移法。

考察矩阵$B=A-pI$，其中p为待选参数。设A的特征值为$\lambda_1,\lambda_2,\cdots,\lambda_n$，则$B$的相应特征值为$\lambda_1-p,\lambda_2-p,\cdots,\lambda_n-p$，且矩阵$A$、$B$有相同的特征向量。

如果需要计算A的按模最大特征值λ_1，就可适当选择p，使得λ_1-p仍是B的按模最大特征值，即

$$|\lambda_1-p|>|\lambda_2-p|\geq|\lambda_3-p|\geq\cdots\geq|\lambda_n-p|$$

且

$$\left|\frac{\lambda_2-p}{\lambda_1-p}\right|<\left|\frac{\lambda_2}{\lambda_1}\right|$$

这样对矩阵B应用幂法，就能较快地求得B的主特征值$\lambda_1(B)$和相应的主特征向量x_1，从而得$\lambda_1=\lambda_1(B)+p$和相应特征向量x_1。这种方法通常称为原点平移法。

由上可以看出，p的选择是加速的关键。但p的选择依赖于对A的特征值分布的大致了解，很难给出p的自动选择。而对于特征值为实数，且有

$$\lambda_1>\lambda_2\geq\lambda_3\geq\cdots\geq\lambda_n$$

时，可有最佳取值$p^*=\frac{1}{2}(\lambda_2+\lambda_n)$，使$\frac{|\lambda_2-p^*|}{|\lambda_1-p^*|}=\min\limits_{p}\max\limits_{2\leq i\leq n}\frac{|\lambda_i-p|}{|\lambda_1-p|}$。

2.5.3 逆幂迭代法

逆幂迭代法简称逆幂法，用于计算非奇异矩阵按模最小的特征值及对应的特征向量。

设矩阵$A\in R^{n\times n}$，且A非奇异，其特征值满足

$$|\lambda_1|\geq|\lambda_2|\geq\cdots\geq|\lambda_{n-1}|>|\lambda_n|>0,$$

相应的特征向量为$\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$，则$A^{-1}$的特征值$\lambda_1^{-1},\lambda_2^{-1},\cdots\lambda_n^{-1}$满足

$$\left|\lambda_n^{-1}\right|>\left|\lambda_{n-1}^{-1}\right|\geq\cdots\geq\left|\lambda_1^{-1}\right|,$$

对应的特征向量为$\boldsymbol{x}_n,\boldsymbol{x}_{n-1},\cdots,\boldsymbol{x}_1$。将幂法用于$A^{-1}$就是逆幂法，可求矩阵$A^{-1}$的主特征值$\lambda_n^{-1}$，进而可得$A$的按模最小的特征值$\lambda_n$。在计算时不必求逆矩阵，而用解方程组的办法。逆幂法的计算过程为：

$$\begin{cases}\boldsymbol{u}_0=\boldsymbol{v}_0\neq 0\ (\alpha_n\neq 0)\\ A\boldsymbol{v}_k=\boldsymbol{u}_{k-1}\\ m_k=\max(\boldsymbol{v}_k)\\ \boldsymbol{u}_k=\boldsymbol{v}_k/m_k\qquad k=1,2,\cdots\end{cases}\tag{2.12}$$

类似于幂法的分析可得，当$k\to\infty$时，

表 2.2: 计算结果

k	$\boldsymbol{u}_k^T$	m_k
0	(1, 1, 1)	
1	(0.4348, 1.0000, - 0.4783)	0.5652
2	(0.1902, 1.0000, - 0.8834)	0.9877
3	(0.1843, 1.0000, - 0.9124)	0.8245
4	(0.1831, 1.0000, - 0.9129	0.8134
5	(0.1832, 1.0000, - 0.9130	0.8134

$$\boldsymbol{u}_k \to \frac{\boldsymbol{x}_n}{\max(\boldsymbol{x}_n)},\ m_k \to \lambda_n^{-1}$$

且迭代收敛速度决定于$\left|\frac{\lambda_n}{\lambda_{n-1}}\right|$。实际计算可以先将$A$作一次$LU$分解（详见 3.1 节，必要时作列主元 LU 分解），每步解 $LUv_k = u_{k-1}$，只要解两次三角方程组。

例 2.11 用逆幂法求矩阵

$$A = \begin{pmatrix} 2 & 8 & 9 \\ 8 & 3 & 4 \\ 9 & 4 & 7 \end{pmatrix}$$

的按模最小的特征值及对应的特征向量。当$\|u_k - u_{k+1}\| \le 10^{-4}$时停止计算。

解：对A作 LU 分解，可得

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 4.5 & 1.1034 & 1 \end{pmatrix},\ U = \begin{pmatrix} 2 & 8 & 9 \\ 0 & -29 & -32 \\ 0 & 0 & 1.8103 \end{pmatrix}$$

取$\boldsymbol{v}_0 = (1,1,1)^T$，按(2.12)计算，其中

$$A\boldsymbol{v}_k = \boldsymbol{u}_{k-1} \Leftrightarrow \begin{cases} L\boldsymbol{y}_k = \boldsymbol{u}_{k-1} \\ U\boldsymbol{v}_k = \boldsymbol{y}_k \end{cases},$$

结果见表 2.2。 可得$\frac{1}{\lambda_3} \approx 0.8134$，$\lambda_3 \approx 1.2294$，其对应的特征向量为

$$x_3 \approx (0.1832,\ 1.0000,\ -0.9130)^T.$$

另一方面，如果已知参数p接近A的某一特征值λ_i，且有

$$0 < |\lambda_i - p| << |\lambda_j - p| \quad (j \ne i)$$

则可用带原点位移的逆幂迭代法求得λ_i，此即对$(A - pI)$进行逆幂法：

$$\begin{cases} (A - pI)\boldsymbol{v}_k = \boldsymbol{u}_{k-1} \\ m_k = \max(\boldsymbol{v}_k) \\ \boldsymbol{u}_k = \boldsymbol{v}_k / m_k \qquad k = 1, 2, \cdots \end{cases}$$

可得，当$k \to \infty$时，

$$\boldsymbol{u}_k \to \frac{\boldsymbol{x}_i}{\max(\boldsymbol{x}_i)},\ m_k \to (\lambda_i - p)^{-1}$$

只要p选择得好，收敛速度是很快的。

第二章习题

1. 设$A \in C^{n\times n}$，若A的每一行元素的和(简称为行和)为1，试证：
(1) 1为A的特征值；(2) 若A可逆，则A^{-1}的行和也是1；(3) 给定多项式 $f(x)$. 问 $f(A)$ 的行和是否相等？若相等，等于何值？

2. 设 $A, B \in C^{n\times n}$， A 的特征值互异。证明：$AB = BA$ 的充分必要条件是 A 与 B 同时可对角化。

3. 求下列λ-矩阵的Smith标准形：

(1) $\begin{pmatrix} \lambda^3-\lambda & 2\lambda^2 \\ \lambda^2+5\lambda & 3\lambda \end{pmatrix}$；(2) $\begin{pmatrix} 1-\lambda & \lambda^2 & \lambda \\ \lambda & \lambda & -\lambda \\ 1+\lambda^2 & \lambda^2 & -\lambda^2 \end{pmatrix}$；

(3) $\begin{pmatrix} 0 & 0 & 0 & \lambda^2 \\ 0 & 0 & \lambda^2-\lambda & 0 \\ 0 & (\lambda-1)^2 & 0 & 0 \\ \lambda^2-\lambda & 0 & 0 & 0 \end{pmatrix}$.

4. 判断 λ-矩阵

$A(\lambda) = \begin{pmatrix} 3\lambda+1 & \lambda & 4\lambda-1 \\ 1-\lambda^2 & \lambda-1 & \lambda-\lambda^2 \\ \lambda^2+\lambda+2 & \lambda & \lambda^2+2\lambda \end{pmatrix}$ 与

$B(\lambda) = \begin{pmatrix} \lambda+1 & \lambda-2 & \lambda^2-2\lambda \\ 2\lambda & 2\lambda-3 & \lambda^2-2\lambda \\ -2 & 1 & 1 \end{pmatrix}$ 是否等价。

5. 判断矩阵A与 B 是否相似：

(1) $A = \begin{pmatrix} 3 & 2 & -5 \\ 2 & 6 & -10 \\ 1 & 2 & -3 \end{pmatrix}$, $B = \begin{pmatrix} 6 & 20 & -34 \\ 6 & 32 & -51 \\ 4 & 20 & -32 \end{pmatrix}$；

(2) $A = \begin{pmatrix} 6 & 6 & -15 \\ 1 & 5 & -5 \\ 1 & 2 & -2 \end{pmatrix}$, $B = \begin{pmatrix} 37 & -20 & -4 \\ 34 & -17 & -4 \\ 119 & -70 & -11 \end{pmatrix}$.

6. 设$\varepsilon \neq 0$，证明：

(1) n阶矩阵$A = \begin{pmatrix} a & 1 & & \\ & a & \ddots & \\ & & \ddots & 1 \\ & & & a \end{pmatrix}$ 与 $B = \begin{pmatrix} a & \varepsilon & & \\ & a & \ddots & \\ & & \ddots & \varepsilon \\ & & & a \end{pmatrix}$ 相似；

(2) n阶矩阵$A=\begin{pmatrix} a & 1 & & \\ & a & \ddots & \\ & & \ddots & 1 \\ & & & a \end{pmatrix}$与$C=\begin{pmatrix} a & 1 & & \\ & a & \ddots & \\ & & \ddots & 1 \\ \varepsilon & & & a \end{pmatrix}$不相似。

7. 求下列矩阵的Jordan标准形：

(1) $\begin{pmatrix} 4 & -5 & 2 \\ 5 & -7 & 3 \\ 6 & -9 & 4 \end{pmatrix}$；(2) $\begin{pmatrix} 1 & -3 & 4 \\ 4 & -7 & 8 \\ 6 & -7 & 7 \end{pmatrix}$；(3) $\begin{pmatrix} 1 & -1 & 2 \\ 3 & -3 & 6 \\ 2 & -2 & 4 \end{pmatrix}$；

(4) $\begin{pmatrix} 3 & 0 & 8 \\ 3 & -1 & 6 \\ -2 & 0 & -5 \end{pmatrix}$；(5) $\begin{pmatrix} 2 & -1 & 1 & -1 \\ 2 & 2 & -1 & -1 \\ 1 & 2 & -1 & 2 \\ 0 & 0 & 0 & 3 \end{pmatrix}$.

8. 求下列矩阵A的Jordan标准形，并求相似变换矩阵P，使$P^{-1}AP=J$：

(1) $A=\begin{pmatrix} -1 & 0 & 0 \\ 1 & 1 & 3 \\ 0 & 2 & 2 \end{pmatrix}$；(2) $A=\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$.

9. 设$A\in \boldsymbol{C}^{n\times n}$，$\lambda_0$为$A$的$r$重特征值，试用Jordan标准形理论证明

$$\operatorname{rank}(\lambda_0 I-A)^r=n-r.$$

10. 下列矩阵A是否为正规矩阵？若是，试求酉矩阵U，使$U^{-1}AU$为对角矩阵：

(1) $A=\begin{pmatrix} -1 & i & 0 \\ -i & 0 & -i \\ 0 & i & -1 \end{pmatrix}$；(2) $A=\begin{pmatrix} 0 & i & 1 \\ -i & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$.

11. 设$A=(a_{jk})\in \boldsymbol{C}^{n\times n}$，则$A$是Hermite矩阵的充分必要条件是对任意$x\in \boldsymbol{C}^n$，$x^HAx$ 是实数。

12. 设$A\in \boldsymbol{C}^{n\times n}$是Hermite矩阵。证明：$A$是Hermite正定矩阵的充分必要条件是存在Hermite正定矩阵S，使得$A=S^2$.

13. 若A、B 均为n阶Hermite正定矩阵，且 $AB=BA$，则 AB 为正定矩阵。

14. 设$A=(a_{ij})\in \boldsymbol{C}^{n\times n}$为Hermite正定矩阵，证明：

(1) $\det A\le a_{nn}\det A_{n-1}$，其中$A_{n-1}$为$A$的$n-1$阶顺序主子阵，且等号成立当且仅当$a_{1n}=a_{2n}=\cdots=a_{n-1,n}=0$；(2) $\det A\le \prod\limits_{i=1}^{n} a_{ii}$.

15. 设$A=\begin{pmatrix} 0 & 0.8 & -1 \\ 0.1 & 2 & -0.1 \\ 0.2 & 0.7 & 3 \end{pmatrix}$，用盖尔圆定理证明$A$有3个互异实特征值。

16. 若$A=(a_{ij})_{n\times n}$严格对角占优，即$|a_{ii}|>R_i\ (i=1,2,\cdots,n)$，则$\det A\neq 0$。

17. 设矩阵$A=\begin{pmatrix} 7 & 3 & -2 \\ 3 & 4 & -1 \\ -2 & -1 & 3 \end{pmatrix}$，分别用幂迭代法和逆幂迭代法求矩阵按模最大特征值和按模最小特征值(迭代5步)。

18. 设矩阵$A=\begin{pmatrix} 3 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{pmatrix}$，(1) 用幂迭代法求矩阵$A$按模最大的特征值及其特征向量；(2)试取平移量$p=\frac{1}{2}[2+(2-\sqrt{3})]\approx 1.134$，用幂迭代法求矩阵$A$按模最大的特征值($\varepsilon=10^{-4}$)。

第 三 章　矩阵分解与广义逆矩阵

本章将介绍矩阵的一些分解方法和广义逆矩阵。矩阵分解是求解线性方程组、处理各类最小二乘问题和最优化问题的主要工具。广义逆矩阵的提出起源于线性方程组的求解，它是数理统计、最优化理论、现代控制理论、网络系统等科学的重要理论基础，并已成为实际工程中广泛使用的重要计算工具。本章主要介绍几种常用的矩阵分解方法和Moore-Penrose广义逆矩阵的基本概念、性质及其计算方法。

3.1　三角分解、满秩分解和奇异值分解

矩阵分解是将矩阵分解成两个或三个在形式上、性质上比较简单的矩阵的乘积。如果矩阵能够分解成为三角矩阵的乘积，那么三角矩阵的特殊形状及良好的运算性质将对求解矩阵行列式、矩阵的逆及线性方程组等问题有很好的帮助。

定义 3.1 设 $A \in C^{n\times n}$，如果存在下三角矩阵 $L \in C^{n\times n}$ 和上三角矩阵 $U \in C^{n\times n}$ 使得 $A = LU$，则称 A 可以做三角分解。

若定义 3.1 中的 L 是对角元素为1的下三角矩阵（单位下三角矩阵），U 为上三角矩阵，则称该三角分解为 Doolittle 分解（或LU分解）。下面我们给出 Doolittle 分解的存在性和唯一性结论。

3.1.1　Doolittle分解

定理 3.1 设$A \in C^{n\times n}$，若其前$n-1$阶顺序主子式均不为0，则A存在Doolittle分解$A = LU$，其中L为单位下三角矩阵，U为上三角矩阵。此外，若A非奇异，则分解唯一。

证明：设 $A = (a_{ij}) \in C^{n\times n}$。利用消元法思想，取

$$L_k = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{k+1,k} & 1 & \\ & & \vdots & & \ddots \\ & & -l_{n,k} & & 1 \end{pmatrix} \quad (k = 1, 2, \cdots, n-1)$$

其中$l_{i,k}=\frac{a_{ik}}{a_{kk}}\ (i=k+1,\cdots,n)$。

则有 $L_{n-1}L_{n-2}\cdots L_1A$为上三角矩阵，记为$U$。 进一步令$L=L_1^{-1}\cdots L_{n-1}^{-1}$，则$L$为单位下三角矩阵，且 $A=LU$。

下证唯一性。设A有两种上述分解，即$A=LU=\bar{L}\bar{U}$。由于A非奇异，故$L,\bar{L},U,\bar{U}$均可逆。 于是有$L^{-1}\bar{L}=U\bar{U}^{-1}$。注意到左端单位下三角矩阵乘积仍是单位下三角矩阵， 右端上三角矩阵乘积仍是上三角矩阵，因此必有$L^{-1}\bar{L}=U\bar{U}^{-1}=I$，即$L=\bar{L},U=\bar{U}$。 证毕。

设矩阵A有唯一的Doolittle分解，即

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} 1 & & & \\ l_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ & u_{22} & \cdots & u_{2n} \\ & & \ddots & \vdots \\ & & & u_{nn} \end{pmatrix}.$$

比较等式两边矩阵的元素，得

$$\begin{cases} a_{kj}=\sum\limits_{t=1}^{k-1} l_{kt}u_{tj}+u_{kj} & (k=1,2,\cdots,n,\ j=k,\cdots,n) \\ a_{ik}=\sum\limits_{t=1}^{k} l_{it}u_{tk} & (i=k+1,\cdots,n). \end{cases}$$

由上可导出Doolittle计算格式

$$\begin{cases} u_{kj}=a_{kj}-\sum\limits_{t=1}^{k-1} l_{kt}u_{tj}, & (k=1,2,\cdots,n,\ j=k,\cdots,n) \\ l_{ik}=\left(a_{ik}-\sum\limits_{t=1}^{k-1} l_{it}u_{tk}\right)\Big/ u_{kk}, & (i=k+1,\cdots,n). \end{cases} \tag{3.1}$$

在进行计算时，也可按下表先行后列逐次交替进行，第一行$u_{1j}=a_{1j}$不用计算，最后$k=n$时，只计算u_{nn}， 且在实际计算过程中，为节省存储空间，得到的l_{ij},u_{ij} 可以存放在 A 的相应元素位置上。

$$\begin{array}{c|c|c|c|c} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ \hline l_{21} & u_{22} & u_{23} & \cdots & u_{2n} \\ \hline l_{31} & l_{32} & u_{33} & \cdots & u_{3n} \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline l_{n1} & l_{n2} & \cdots & \cdots & u_{nn} \end{array}$$

例 3.1 求矩阵

$$A=\begin{pmatrix} 2 & 5 & -6 \\ 4 & 13 & -19 \\ -6 & -3 & -6 \end{pmatrix}$$

的Doolittle分解。

解： 由式(3.1)得

$$\begin{aligned} &u_{11}=a_{11}=2, u_{12}=a_{12}=5, u_{13}=a_{13}=-6 \\ &l_{21}=a_{21}/u_{11}=2, l_{31}=a_{31}/u_{11}=-3 \\ &u_{22}=a_{22}-l_{21}u_{12}=3, u_{23}=a_{23}-l_{21}u_{13}=-7 \\ &l_{32}=(a_{32}-l_{31}u_{12})/u_{22}=4 \\ &u_{33}=a_{33}-l_{31}u_{13}-l_{32}u_{23}=4 \end{aligned}$$

则A的Doolittle分解为

$$A=LU=\begin{pmatrix} 1 & & \\ 2 & 1 & \\ -3 & 4 & 1 \end{pmatrix}\begin{pmatrix} 2 & 5 & -6 \\ & 3 & -7 \\ & & 4 \end{pmatrix}.$$

3.1.2 选列主元的Doolittle分解

若 $u_{kk}=0$ 或 $|u_{kk}|$ 很小时，在前一小节的三角分解过程中，将出现中断或溢出。为克服上述缺点，我们需要在分解过程中加入选列主元步骤。

定义 3.2 以n阶单位矩阵I_n的n个列向量$e_1, e_2, \cdots, e_n$为列构成的 n 阶方阵 $P=(e_{i_1}\ e_{i_2}\ \cdots e_{i_n})$ 称为 n 阶置换矩阵，这里 $i_1, i_2, \cdots, i_n$ 是 $1, 2, \cdots, n$的一个全排列。

定理 3.2 设A为非奇异矩阵，则存在置换矩阵P，使得 $PA=LU$，其中L为单位下三角矩阵，U为上三角矩阵。

定理3.2给出了选列主元三角分解的可行性保证。下面给出具体分解及其存储

过程。设A已经过$k-1$步列主元分解，即有

$$A \longrightarrow \bar{A} = \begin{pmatrix} u_{11} & \cdots & u_{1,k-1} & u_{1,k} & \cdots & u_{1,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{k-1,1} & \cdots & u_{k-1,k-1} & u_{k-1,k} & \cdots & u_{k-1,n} \\ l_{k,1} & \cdots & l_{k,k-1} & \bar{a}_{k,k} & \cdots & \bar{a}_{k,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ l_{i,1} & \cdots & l_{i,k-1} & \bar{a}_{i,k} & \cdots & \bar{a}_{i,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{n,1} & \cdots & l_{n,k-1} & \bar{a}_{n,k} & \cdots & \bar{a}_{n,n} \end{pmatrix}.$$

第k步时，令

$$\bar{a}_{ik} \longleftarrow c_i = \bar{a}_{ik} - \sum_{t=1}^{k-1} \bar{a}_{it}\bar{a}_{tk} = \bar{a}_{ik} - \sum_{t=1}^{k-1} l_{it}u_{tk} \quad (i = k, k+1, \cdots n). \tag{3.2}$$

此时,增加选列主元步骤，取

$$|c_{i_k}| = \max_{k \le i \le n} |c_i|.$$

如果$i_k \neq k$，则交换上述矩阵$\bar{A}$中的第k行与第i_k行，即

$$\bar{a}_{kj} \longleftrightarrow \bar{a}_{i_k j}, \quad (j = 1, 2, \cdots, n). \tag{3.3}$$

得

$$\tilde{A} = \begin{pmatrix} u_{11} & \cdots & u_{1,k-1} & u_{1,k} & \cdots & u_{1,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{k-1,1} & \cdots & u_{k-1,k-1} & u_{k-1,k} & \cdots & u_{k-1,n} \\ l_{i_k,1} & \cdots & l_{i_k,k-1} & c_{i_k} & \cdots & \bar{a}_{i_k,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{i,1} & \cdots & l_{i,k-1} & c_i & \cdots & \bar{a}_{i,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{k,1} & \cdots & l_{k,k-1} & c_k & \cdots & \bar{a}_{k,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{n,1} & \cdots & l_{n,k-1} & c_n & \cdots & \bar{a}_{n,n} \end{pmatrix}.$$

由此给出计算l_{ik} $(i = k+1, \cdots, n)$及u_{kj} $(j = k, k+1, \cdots, n)$的公式及其存储位置。

$$\begin{cases} \tilde{a}_{kk} \longleftarrow u_{kk} = \tilde{a}_{kk}, \\ \tilde{a}_{ik} \longleftarrow l_{ik} = \dfrac{\tilde{a}_{ik}}{\tilde{a}_{kk}} \quad (i = k+1, \cdots, n), \\ \tilde{a}_{kj} \longleftarrow u_{kj} = \tilde{a}_{kj} - \displaystyle\sum_{t=1}^{k-1} \tilde{a}_{kt}\tilde{a}_{tj} \quad (j = k+1, \cdots, n). \end{cases} \tag{3.4}$$

其中

$$\tilde{a}_{kk} = c_{i_k},$$

$$\tilde{a}_{i_k,k} = c_k,$$

$$\tilde{a}_{ik} = c_i \quad (i = k+1, \cdots, n; i \neq k, i \neq i_k),$$

$$\tilde{a}_{kj} = \bar{a}_{i_k,j} \quad (j = k+1, \cdots, n).$$

按式(3.2)-(3.4)依次计算n步，可得到矩阵选列主元三角分解的最终结果。而且L和U的元素存储于原始矩阵A的相应位置上，对角线以下元素为L的数值，对角线及其以上元素为U的数值。

例 3.2 用选列主元的Doolittle分解法分解矩阵

$$A = \begin{pmatrix} 0.5 & 1 & 0 \\ 2 & 1.5 & 1 \\ 0.2 & 1 & 2.5 \end{pmatrix}.$$

解：根据选列主元的Doolittle分解法，有

$$A = \begin{pmatrix} 0.5 & 1 & 0 \\ 2 & 1.5 & 1 \\ 0.2 & 1 & 2.5 \end{pmatrix}$$

$$\xrightarrow{k=1} \begin{pmatrix} 2 & 1.5 & 1 \\ 0.5 & 1 & 0 \\ 0.2 & 1 & 2.5 \end{pmatrix} \xrightarrow{k=1} A^{(1)} = \begin{pmatrix} 2 & 1.5 & 1 \\ 0.25 & 1 & 0 \\ 0.1 & 1 & 2.5 \end{pmatrix}$$

$$\xrightarrow{k=2} \begin{pmatrix} 2 & 1.5 & 1 \\ 0.1 & 0.85 & 2.5 \\ 0.25 & 0.625 & 0 \end{pmatrix} \xrightarrow{k=2} A^{(2)} = \begin{pmatrix} 2 & 1.5 & 1 \\ 0.1 & 0.85 & 2.4 \\ 0.25 & \frac{25}{34} & 0 \end{pmatrix}$$

$$\xrightarrow{k=3} A^{(3)} = \begin{pmatrix} 2 & 1.5 & 1 \\ 0.1 & 0.85 & 2.4 \\ 0.25 & \frac{25}{34} & -\frac{137}{68} \end{pmatrix}$$

具体过程说明如下，

$k=1$: $c_1 = a_{11} = 0.5$, $c_2 = a_{21} = 2$, $c_3 = 0.2; |c_2| = \max\limits_{1\leq i\leq 3} |c_i|$, 交换第一，二行; $u_{11}^{(1)} = 2$, $u_{12}^{(1)} = 1.5$ $u_{13}^{(1)} = 1; l_{21}^{(1)} = \frac{0.5}{2} = 0.25$, $l_{31}^{(1)} = \frac{0.2}{2} = 0.1$.

$k=2$: $c_2 = a_{22}^{(1)} - a_{21}^{(1)} a_{12}^{(1)} = 1 - 0.25 \times 1.5 = 0.625, c_3 = a_{32}^{(1)} - a_{31}^{(1)} a_{12}^{(1)} = 1 - 0.1 \times 1.5 = 0.85; |c_3| = \max\limits_{2\leq i\leq 3} |c_i|$, 交换第二，三行; $u_{22}^{(2)} = 0.85$, $u_{23}^{(2)} = 2.5 - 0.1 \times 1 = 2.4$, $l_{32}^{(2)} = \frac{0.625}{0.85} = \frac{25}{34}$.

$k=3:\ c_3=a_{33}^{(2)}-a_{31}^{(2)}a_{13}^{(2)}-a_{32}^{(2)}a_{23}^{(2)}=0-0.25\times 1-\frac{25}{34}\times 2.4=-\frac{137}{68};u_{33}^{(3)}=c_3=-\frac{137}{68}.$

于是得

$$P=\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix},\quad L=\begin{pmatrix}1&0&0\\0.1&1&0\\0.25&\frac{25}{34}&1\end{pmatrix},\quad U=\begin{pmatrix}2&1.5&1\\0&0.85&2.4\\0&0&-\frac{137}{68}\end{pmatrix}.$$

3.1.3 Cholesky分解

当矩阵A为Hermite正定阵时，A的LU分解更为简单，且无需选主元。

定理 3.3 设$A\in C^{n\times n}$是Hermite正定矩阵，则存在唯一分解$A=LL^H$，其中$L\in C^{n\times n}$为具有主对角元素为正数的下三角矩阵。

证明：由于A是Hermite正定矩阵，则A的所有顺序主子式$\triangle_i(i=1,\cdots,n)$皆大于0。所以由定理3.1知存在唯一分解$A=L_1U_1$，其中$L_1$为单位下三角矩阵，$U_1$为上三角矩阵，且其主对角元素设为$d_i$，有$d_i>0(i=1,\cdots,n)$。因而对$U_1$可进一步分解，有$U_1=DU_2$，其中$D=\mathrm{diag}(d_1,\cdots,d_n)$，$U_2$为单位上三角矩阵。同时，注意到$A$为Hermite矩阵，所以有$U_2=L_1^H$。因而得

$$A=L_1\mathrm{diag}(\sqrt{d_1},\cdots,\sqrt{d_n})\mathrm{diag}(\sqrt{d_1},\cdots,\sqrt{d_n})L_1^H.$$

令

$$L=L_1\mathrm{diag}(\sqrt{d_1},\cdots,\sqrt{d_n}),$$

则L为下三角阵，且$A=LL^H$。 证毕。

称分解式$A=LL^H$为矩阵A的Cholesky分解。设

$$L=\begin{pmatrix}l_{11}&&&\\l_{21}&l_{22}&&\\\cdots&\cdots&\cdots&\\l_{n1}&l_{n2}&\cdots&l_{nn}\end{pmatrix},$$

给出其元素的计算公式如下：

$$\begin{cases}l_{jj}=\left(a_{jj}-\sum\limits_{k=1}^{j-1}|l_{jk}|^2\right)^{\frac{1}{2}},\quad (j=1,2,\cdots,n),\\ l_{ij}=\left(a_{ij}-\sum\limits_{k=1}^{j-1}l_{ik}\bar{l}_{jk}\right)/l_{jj},\quad (i=j+1,\cdots,n).\end{cases}\tag{3.5}$$

例 3.3 已知矩阵$A=\begin{pmatrix}4&-1&1\\-1&2&-2\\1&-2&3\end{pmatrix}$，求$A$的Cholesky分解。

解：

$$l_{11}=\sqrt{a_{11}}=2,\ l_{21}=\frac{a_{21}}{l_{11}}=-\frac{1}{2},\ l_{31}=\frac{a_{31}}{l_{11}}=\frac{1}{2},$$

$$l_{22}=\sqrt{a_{22}-|l_{21}|^2}=\frac{\sqrt{7}}{2},\ l_{32}=(a_{32}-l_{31}\bar{l}_{21})=-\frac{\sqrt{7}}{2},$$

$$l_{33}=\sqrt{a_{33}-|l_{31}|^2-|l_{32}|^2}=1.$$

故A的Cholesky分解为

$$A=\begin{pmatrix}2&0&0\\-\frac{1}{2}&\frac{\sqrt{7}}{2}&0\\\frac{1}{2}&-\frac{\sqrt{7}}{2}&1\end{pmatrix}\begin{pmatrix}2&-\frac{1}{2}&\frac{1}{2}\\0&\frac{\sqrt{7}}{2}&-\frac{\sqrt{7}}{2}\\0&0&1\end{pmatrix}.$$

3.1.4 矩阵的 QR 分解

定义 3.3 设 $A\in C^{n\times n}$。如果存在 n 阶酉矩阵 Q 和 n 阶上三角矩阵R，使得 $A=QR$，则称其为 A 的QR分解或酉三角分解。当 $A\in R^{n\times n}$，则称其为 A 的正交三角分解。

不加证明地引入如下结论。

定理 3.4 任意 $A\in C^{n\times n}$ 都可以进行 QR 分解。

用 Gram-Schmidt 正交化方法可将一般的长方阵作QR分解。

定理 3.5 设 $A\in C^{m\times n}$，那么存在有标准正交列的矩阵 $Q\in C^{m\times n}$ 和上三角矩阵 $R\in C^{n\times n}$，使得 $A=QR$。

3.1.5 矩阵的满秩分解

满秩分解是将非零矩阵分解为一个列满秩矩阵与一个行满秩矩阵之积，它是研究和计算矩阵广义逆等问题的有力工具。

定义 3.4 设 $A\in C_r^{m\times n}$ $(r>0)$，如果存在 $F\in C_r^{m\times r}$ 和 $G_r^{r\times n}$ 使得 $A=FG$，则称为 A 的满秩分解。

在介绍满秩分解存在性和方法之前，先给出一些相关的基础知识。

定义 3.5 设 $A,B\in C^{m\times n}$，如果 A 经过有限次初等变换化成矩阵 B，则称矩阵A与B等价。

定理 3.6 设 $A\in C_r^{m\times n}$ $(r>0)$，则存在 $S\in C_m^{m\times m}$，$T\in C_n^{n\times n}$ 使得

$$SAT=\begin{pmatrix} I_r & 0\\ 0 & 0\end{pmatrix}.$$

其中右端项称为矩阵 A 的等价标准形。

由定理3.6，可得到满秩分解存在性结论:

定理 3.7 设$A\in C_r^{m\times n}(r>0)$,则总存在$F\in C_r^{m\times r}$和$G\in C_r^{r\times n}$使得$A=FG$，即$A$的满秩分解总是存在的。

证明：当$r=m$时，令$F=I_m,G=A$，则$A=I_mA$为A的一个满秩分解。当$r=n$时，令$F=A,G=I_n$，则$A=AI_n$为一个满秩分解。下面考虑 $0<r<\min\{m,n\}$的情形。由定理3.6知存在$S\in C_m^{m\times m},T\in C_m^{m\times m}$使得

$$SAT=\begin{pmatrix} I_r & 0\\ 0 & 0\end{pmatrix}.$$

于是

$$A=S^{-1}\begin{pmatrix} I_r & 0\\ 0 & 0\end{pmatrix}T^{-1}=S^{-1}\begin{pmatrix} I_r\\ 0\end{pmatrix}\begin{pmatrix} I_r & 0\end{pmatrix}T^{-1}.$$

令$F=S^{-1}\begin{pmatrix} I_r\\ 0\end{pmatrix}\in C_r^{m\times r},G=\begin{pmatrix} I_r & 0\end{pmatrix}T^{-1}\in C_r^{r\times n}$，则$A=FG$为$A$的一个满秩分解。 证毕。

需要注意的是，满秩分解并不唯一。若$A=FG$为A的一个满秩分解，设$B\in C_r^{r\times r}$为一可逆矩阵，且令$\bar{F}=FB,\bar{G}=B^{-1}G$，则$A=\bar{F}\bar{G}$构成$A$的另一个满秩分解。

满秩分解存在性定理 3.7 的证明过程也同时给出了求解满秩分解的一种方法，它要求已知矩阵 S 和 T。下面讨论求解 S,T 的方法。

如果只对A作初等行变换，有下面定理成立。

定理 3.8 设$A\in C_r^{m\times n}(r>0)$，则$A$可通过初等行变换化为满足如下条件的矩阵 H:

(1) H 的前 r 行中每一行至少含一个非零元素，且第一个非零元素是 1，而后 $m-r$ 行元素仍为 0;

(2) 若 H 中第 i 行的第一个非零元素 1 位于第 j_i 列 $(i=1,2,\cdots,r)$，则 $j_1<j_2<\cdots<j_r$;

(3) H 的第 $j_1,j_2,\cdots,j_r$ 列为 I_m 的前 r 列，即有

$$H=\begin{pmatrix}
0\cdots 0\,1\,*\cdots * & 0\,*\cdots 0\,*\cdots * \\
0\cdots 0\,0\,0\cdots 0\,1\,*\cdots & \vdots\quad\vdots \\
\vdots\quad\vdots\quad\vdots & 0\,*\cdots * \\
0\cdots 0\,0\,0\cdots 0\,0\,0\cdots & 1\,*\cdots * \\
0\cdots 0\,0\,0\cdots 0\,0\,0\cdots & 0\,0\cdots 0 \\
\vdots\quad\vdots\quad\vdots & \vdots\quad\vdots\quad\vdots \\
0\cdots 0\,0\,0\cdots 0\,0\,0\cdots & 0\,0\cdots 0
\end{pmatrix}\Bigg\}(r\text{行})$$

$$\underbrace{\qquad}_{j_1}\quad\underbrace{\qquad}_{j_2}\quad\cdots\quad\underbrace{\qquad}_{j_r}$$

称 H 为 A 的Hermite标准形或行最简形。也就是说，存在 $S\in C_m^{m\times m}$ 使得 $SA=H$.

由上定理可知，对 $m\times(m+n)$ 矩阵 (A,I_m)，有

$$S(A,I_m)=(SA,S)=(H,S),$$

即

$$(A,I_m)\xrightarrow{\text{初等行变换}}(H,S). \tag{3.6}$$

则便得到矩阵 S 和Hermite标准形 H，且相应元素分别保存在原单位矩阵和矩阵 A 的相应位置上。

如果将 A 的Hermite标准形 H 的第 $j_1,j_2,\cdots,j_r$ 列依次调换到前 r 列位置上，即用置换矩阵 $P=(e_{j_1},e_{j_2},\cdots,e_{j_n})$ 右乘矩阵 H，有

定理 3.9 设 $A\in C_r^{m\times n}$ $(r>0)$，则存在 $S\in C_m^{m\times m}$ 和 n 阶置换矩阵 P 使得

$$SAP=\begin{pmatrix}I_r & K\\ 0 & 0\end{pmatrix}\qquad(K\in C^{r\times(n-r)}). \tag{3.7}$$

若对上面右端项的矩阵继续进行初等列变换，便可得到 A 的等价标准形 $\begin{pmatrix}I_r & 0\\ 0 & 0\end{pmatrix}$。于是，可得求 T 的过程，即

$$\begin{pmatrix}H\\ I_n\end{pmatrix}\xrightarrow{\text{初等列变换}}\begin{pmatrix}\begin{pmatrix}I_r & 0\\ 0 & 0\end{pmatrix}\\ T\end{pmatrix}. \tag{3.8}$$

利用变换过程 (3.6) 和 (3.8)，求得 S,T，继而求得 S^{-1},T^{-1}，最后取 S^{-1} 的前 r 列得到 F，取 T^{-1} 的前 r 行得到 G。

例 3.4 设 $A=\begin{pmatrix}1&-1&-1&4\\0&0&2&2\\-1&1&5&0\end{pmatrix}$

(1) 求A的Hermite标准形H及矩阵S; (2) 求置换矩阵P化A为(3.7)的形式; (3) 求化A为等价标准形所需的矩阵S,T; (4) 求矩阵A的满秩分解。

解: (1) 因为

$$(A,I_3)=\left(\begin{array}{cccc|ccc}1&-1&-1&4&1&0&0\\0&0&2&2&0&1&0\\-1&1&5&0&0&0&1\end{array}\right)\xrightarrow{r_3+r_1}\left(\begin{array}{cccc|ccc}1&-1&-1&4&1&0&0\\0&0&2&2&0&1&0\\0&0&4&4&1&0&1\end{array}\right)$$

$$\xrightarrow{r_2\times\frac{1}{2}}\left(\begin{array}{cccc|ccc}1&-1&-1&4&1&0&0\\0&0&1&1&0&\frac{1}{2}&0\\0&0&4&4&1&0&1\end{array}\right)\xrightarrow{r_3-4\times r_2}\left(\begin{array}{cccc|ccc}1&-1&-1&4&1&0&0\\0&1&1&1&0&\frac{1}{2}&0\\0&0&0&0&1&-2&1\end{array}\right)$$

$$\xrightarrow{r_1+r_2}\left(\begin{array}{cccc|ccc}1&-1&0&5&1&\frac{1}{2}&0\\0&0&1&1&0&\frac{1}{2}&0\\0&0&0&0&1&-2&1\end{array}\right),$$

所以

$$H=\begin{pmatrix}1&-1&0&5\\0&0&1&1\\0&0&0&0\end{pmatrix}\quad S=\begin{pmatrix}1&\frac{1}{2}&0\\0&\frac{1}{2}&0\\1&-2&1\end{pmatrix}.$$

(2) 取4阶置换矩阵 $P=(e_1,e_3,e_2,e_4)$, 则

$$SAP=\begin{pmatrix}I_2&K\\0&0\end{pmatrix}=\begin{pmatrix}1&0&-1&5\\0&1&0&1\\0&0&0&0\end{pmatrix}.$$

(3) 又

$$\left(\begin{array}{cccc}1&-1&0&5\\0&0&1&1\\0&0&0&0\\\hline1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{array}\right)\xrightarrow{c_2\leftrightarrow c_3}\left(\begin{array}{cccc}1&0&-1&5\\0&1&0&1\\0&0&0&0\\\hline1&0&0&0\\0&0&1&0\\0&1&0&0\\0&0&0&1\end{array}\right)\xrightarrow[c_4-c_2]{\substack{c_3+c_1\\c_4+c_1\times(-5)}}\left(\begin{array}{cccc}1&0&0&0\\0&1&0&0\\0&0&0&0\\\hline1&0&1&-5\\0&0&1&0\\0&1&0&-1\\0&0&0&1\end{array}\right),$$

所以

$$T=\begin{pmatrix}1&0&1&-5\\0&0&1&0\\0&1&0&-1\\0&0&0&1\end{pmatrix}.$$

由(1)得

$$S=\begin{pmatrix}1&\frac{1}{2}&0\\0&\frac{1}{2}&0\\1&-2&1\end{pmatrix}.$$

(4) 由(3)，得

$$S^{-1}=\begin{pmatrix}1&-1&0\\0&2&0\\-1&5&1\end{pmatrix}\quad T^{-1}=\begin{pmatrix}1&-1&0&5\\0&0&1&1\\0&1&0&0\\0&0&0&1\end{pmatrix}.$$

由(1)可知,$r_A=2$，故可取

$$F=\begin{pmatrix}1&-1\\0&2\\-1&5\end{pmatrix}\quad G=\begin{pmatrix}1&-1&0&5\\0&0&1&1\end{pmatrix}$$

而 $A=FG$ 为 A 的一个满秩分解。

通过定理3.7所给方法，可以得到矩阵的一个满秩分解， 但这种方法的计算量较大，为克服该缺点，下面给出一种较为简单的改进方法。

定理 3.10 设 $A\in C_r^{m\times n}$ $(r>0)$，且A的Hermite标准形为H，取 A 的第 $j_1,j_2,\cdots,j_r$ 列构成 F， 取H的前r行构成G， 则$A=FG$ 即为A的一个满秩分解。

证明：取 $P=(e_{j_1},e_{j_2},\cdots,e_{j_r})$, 则 $GP=I_r$. 从式 $A=FG$ 两边右乘 P，得 $F=AP$， 可见 F 由 A 的第 $j_1,j_2,\cdots,j_r$ 列构成。又 $r=\mathrm{rank}A=\mathrm{rank}(FG)\le \mathrm{rank}F\le r$， 所以$F\in C_r^{m\times r}$。显然, $G\in C_r^{r\times n}$。 证毕。

由定理 3.10 所叙述方法求满秩分解时，只需求出 A 的Hermite标准形 H， 因此计算量简化很多。

例 3.5 求矩阵 $A=\begin{pmatrix}1&-1&-1&4\\0&0&2&2\\-1&1&5&0\end{pmatrix}$ 的满秩分解。

解： 由例 3.4，得

$$A \xrightarrow{\text{初等行变换}} H = \begin{pmatrix} 1 & -1 & 0 & 5 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

则得 $j_1 = 1, j_2 = 3$，由此取 A 的第 1 列和第3 列元素构成 F，H 的前两行元素构成 G，即

$$A = \begin{pmatrix} 1 & -1 \\ 0 & 2 \\ -1 & 5 \end{pmatrix} \begin{pmatrix} 1 & -1 & 0 & 5 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

3.1.6 矩阵的奇异值分解

矩阵的奇异值分解在广义逆矩阵和最小二乘问题及统计等理论方面有很重要的应用。首先给出相关的定义。

定义 3.6 设 $A \in C_r^{m\times n}$ $(r > 0)$, A^HA的特征值为

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0,$$

则称$\sigma_i = \sqrt{\lambda_i}$ $(i = 1, 2, \cdots, n)$ 为A的奇异值.

定义 3.7 设$A, B \in C^{m\times n}$， 若存在m 阶酉矩阵U和n阶酉矩阵 V， 使得$U^HAV = B$， 则称A与B酉等价。

对于酉等价的矩阵有如下基本性质：

定理 3.11 酉等价矩阵有相同的奇异值。

证明：设$U^HAV = B$， 则

$$B^HB = (U^HAV)^H(U^HAV) = V^HA^HAV$$

即B^HB与A^HA相似, 从而它们有相同的特征值，所以A与B 有相同的奇异值。 证毕。

下面给出矩阵奇异值分解定理：

定理 3.12 设 $A \in C_r^{m\times n}$ $(r > 0)$， 则存在 m 阶酉矩阵 U 和 n 阶酉矩阵 V 使得

$$U^HAV = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix},$$

其中 $\Sigma=\mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_r)$, σ_i为 A的非零奇异值。 而

$$A=U\begin{pmatrix}\Sigma & 0\\ 0 & 0\end{pmatrix}V^H \tag{3.9}$$

称为 A 的奇异值分解。

证明：由于A^HA为Hermite矩阵，则存在n阶酉矩阵V使得

$$V^HA^HAV=diag(\lambda_1,\lambda_2,\cdots,\lambda_n)=\begin{pmatrix}\Sigma^2 & 0\\ 0 & 0\end{pmatrix}.$$

将V分块为

$$V=(V_1,V_2),\quad (V_1\in C^{n\times r},\ V_2\in C^{n\times(n-r)}),$$

得

$$V_1^HA^HAV_1=\Sigma^2, V_2^HA^HAV_2=0$$

于是

$$\Sigma^{-1}V_1^HA^HAV_1\Sigma^{-1}=I_r,\ \ (AV_2)^H(AV_2)=0$$

从而 $AV_2=0$。 又记 $U_1=AV_1\Sigma^{-1}$，则$U_1^HU_1=\mathrm{I_r}$，即 U_1 的 r 个列是两两正交的单位向量。取$U_2\in C^{m\times(m-r)}$ 使 $U=(U_1,U_2)$ 为 m 阶酉矩阵, 即 $U_2^HU_1=0$，$U_2^HU_2=\mathrm{I_{m-r}}$。 则有

$$U^HAV=\begin{pmatrix}U_1^H\\ U_2^H\end{pmatrix}A(V_1,V_2)=\begin{pmatrix}U_1^HAV_1 & U_1^HAV_2\\ U_2^HAV_1 & U_2^HAV_2\end{pmatrix}$$

$$=\begin{pmatrix}U_1^H(U_1\Sigma) & 0\\ U_2^H(U_1\Sigma) & 0\end{pmatrix}=\begin{pmatrix}\Sigma & 0\\ 0 & 0\end{pmatrix}.$$

证毕。

由上可以看出，奇异值分解本质为矩阵在酉等价下的一种标准形。

推论 3.1 设$A\in C_n^{n\times n}$, 则存在n阶酉矩阵U 和 V使得

$$U^HAV=\mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_n)$$

其中$\sigma_i>0$ $(i=1,2,\cdots,n)$为A的奇异值。

由定理3.12的证明，可得 A 的奇异值分解的一种方法，其步骤如下：

(1) 将A^HA酉对角化，即求得酉矩阵 V 使 $V^HA^HAV=\begin{pmatrix}\Sigma^2 & 0\\ 0 & 0\end{pmatrix}$;

(2) 将V分块为 $V=(V_1,V_2)$, $(V_1\in C^{n\times r},\ V_2\in C^{n\times(n-r)})$,

计算$U_1=AV_1\Sigma^{-1}$;

(3) 取 U_2 使 $U=(U_1,U_2)$ 为 m 阶酉矩阵，则 $A=U\begin{pmatrix}\Sigma & 0\\ 0 & 0\end{pmatrix}V^H$ 为 A 的奇异值分解。

例 3.6 求矩阵$A=\begin{pmatrix}1&0&1\\1&1&0\end{pmatrix}$的奇异值分解。

解：因为 $A^TA=\begin{pmatrix}2&1&1\\1&1&0\\1&0&1\end{pmatrix}$，所以$A^TA$的特征值为 $\lambda_1=3,\lambda_2=1,\lambda_3=0$，对应的特征向量为

$$p_1=\begin{pmatrix}2\\1\\1\end{pmatrix},\quad p_2=\begin{pmatrix}0\\-1\\1\end{pmatrix},\quad p_3=\begin{pmatrix}-1\\1\\1\end{pmatrix}$$

标准化得

$$V=\begin{pmatrix}\frac{2}{\sqrt{6}}&0&-\frac{1}{\sqrt{3}}\\ \frac{1}{\sqrt{6}}&-\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{3}}\\ \frac{1}{\sqrt{6}}&\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{3}}\end{pmatrix}$$

使得 $V^HA^HAV=\begin{pmatrix}3&&\\&1&\\&&0\end{pmatrix}=\begin{pmatrix}\Sigma^2&\\&0\end{pmatrix}$。

计算

$$U_1=AV_1\Sigma^{-1}=\begin{pmatrix}1&0&1\\1&1&0\end{pmatrix}\begin{pmatrix}\frac{2}{\sqrt{6}}&0\\ \frac{1}{\sqrt{6}}&-\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{6}}&\frac{1}{\sqrt{2}}\end{pmatrix}\begin{pmatrix}\frac{1}{\sqrt{3}}&0\\0&1\end{pmatrix}$$

$$=\begin{pmatrix}\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}&-\frac{1}{\sqrt{2}}\end{pmatrix}.$$

则 $U=U_1$是酉矩阵。故A的奇异值分解为

$$A=U\begin{pmatrix}\Sigma&0\end{pmatrix}V^H$$

$$=\begin{pmatrix}\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}&-\frac{1}{\sqrt{2}}\end{pmatrix}\begin{pmatrix}\sqrt{3}&0&0\\0&1&0\end{pmatrix}\begin{pmatrix}\frac{2}{\sqrt{6}}&\frac{1}{\sqrt{6}}&\frac{1}{\sqrt{6}}\\ 0&-\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}\\ -\frac{1}{\sqrt{3}}&\frac{1}{\sqrt{3}}&\frac{1}{\sqrt{3}}\end{pmatrix}.$$

3.2 Penrose方程及其Moore-Penrose逆的计算

3.2.1 Penrose方程

考虑下面的线性方程组的求解问题：

$$A\boldsymbol{x} = \boldsymbol{b} \tag{3.10}$$

其中$A \in \boldsymbol{C}^{m\times n}$，$\boldsymbol{x} \in \boldsymbol{C}^{n}$，$\boldsymbol{b} \in \boldsymbol{C}^{m}$。

当 $m = n$，且 A 可逆时，方程组(3.10)有唯一解，即

$$\boldsymbol{x} = A^{-1}\boldsymbol{b}. \tag{3.11}$$

然而，在很多实际问题中方程组(3.10)的系数矩阵 A 不可逆，或者甚至根本就不是方阵，但是方程组的解却是存在的。那么如何用类似于(3.11)的形式来表示方程组的解呢？由于此时系数矩阵 A 在通常意义下是不可逆的，因此就需要对通常意义下的“逆矩阵”概念进行推广，这就是广义逆矩阵。

广义逆矩阵的概念是在 1920 年由美国学者 E.H.Moore 首先提出的，他利用正交投影算子首次给出广义逆矩阵的定义，但却持续三十多年未能引起人们的重视，直到 1955 年英国学者 Penrose 利用4个矩阵方程（现在称之为Penrose方程组）给出了广义逆矩阵的简洁而实用的新定义之后，广义逆矩阵的理论与应用才进入了迅速发展的时期。

1955 年，Penrose 定义了 A 的广义逆矩阵如下。

定义 3.8 *对任意矩阵 $A \in \boldsymbol{C}^{m\times n}$，若矩阵 $X \in \boldsymbol{C}^{n\times m}$ 满足下面4个方程*

$$\begin{aligned} &(1)\ AXA = A; \qquad (2)\ XAX = X;\\ &(3)\ (AX)^H = AX; \quad (4)\ (XA)^H = XA; \end{aligned} \tag{3.12}$$

中的任意一个或者几个，则矩阵 X 就称为矩阵 A 的广义逆矩阵。

这4个方程称为 Penrose 方程，由于 X 只需要满足4个方程中的任意一个或某几个，因此通常按照 A 的广义逆矩阵 X 所满足的方程对广义逆矩阵进行分类。

定义 3.9 *对于矩阵 $A \in \boldsymbol{C}^{m\times n}$，若矩阵 $X \in \boldsymbol{C}^{n\times m}$ 满足 Penrose 方程中的第 $(i),(j),\cdots,(l)$ 个方程，则称 X 为 A 的 $\{i,j,\ldots,l\}$-逆，记为 $A^{(i,j,\cdots,l)}$；所有的 $A^{(i,j,\cdots,l)}$ 构成一类广义逆，记为 $A\{i,j,\cdots,l\}$。*

按照这样的分类方式，矩阵 A 的广义逆矩阵总共可分为 15 类，其中应用较多的有下面几类广义逆矩阵：

(1) $A\{1\}$: 矩阵的 $\{1\}$-逆是最基本的广义逆矩阵，通常记为 A^-. 它与相容线性方程组 $Ax=b$ 的解有着密切联系；

(2) $A\{1,2\}$: 矩阵的 $\{1,2\}$-逆被称为自反广义逆矩阵。此时矩阵 A 和 X 的地位完全一样，它们互为 $\{1,2\}$-逆；

(3) $A\{1,3\}$: 矩阵的 $\{1,3\}$-逆被称为最小二乘广义逆矩阵。在实际问题中许多线性方程组没有解，此时可以求方程组的最小二乘解， 而矩阵的 $\{1,3\}$-逆在最小二乘解的求解中起着非常重要的作用；

(4) $A\{1,4\}$: 矩阵的 $\{1,4\}$-逆也被称为最小范数广义逆矩阵。当相容线性方程组 $Ax=b$ 有无穷多个解的情况下，人们往往需要找到范数最小的解， A 的 $\{1,4\}$-逆在最小范数解的求解中起着十分重要的作用；

(5) $A\{1,2,3,4\}$: 矩阵的 $\{1,2,3,4\}$-逆也被称为 Moore-Penrose 逆，通常记为A^+，Moore-Penrose 逆 A^+ 是应用最多、最广泛的广义逆矩阵。

在这些常用的广义逆矩阵中，矩阵的 $\{1\}$-逆是最基本的，而矩阵的 Moore-Penrose 逆 A^+ 同时满足4个 Penrose方程，它满足所有广义逆矩阵的所有性质，是应用最多、最广泛的广义逆矩阵。本节中主要介绍矩阵 A 的 Moore-Penrose 逆。

3.2.2 Moore-Penrose逆的计算

目前，有很多关于 A^+ 的计算方法，这里主要介绍利用矩阵的奇异值分解及满秩分解的直接计算方法和几种计算广义逆矩阵的迭代方法。

1. 计算A^+的直接方法

(1) 利用 A 的奇异值分解计算 A^+

定理 3.13 *对于任意矩阵 $A\in C_r^{m\times n}$，都存在酉矩阵 $U\in C_m^{m\times m}$ 和 $V\in C_n^{n\times n}$，得到 A 的奇异值分解为*

$$A=U\begin{pmatrix}\Sigma & 0\\ 0 & 0\end{pmatrix}_{m\times n}V^H$$

其中 $\Sigma=\mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_r)$，$\sigma_1,\sigma_2,\cdots,\sigma_r$为 A 的 r 个非零奇异值。则

$$A^+=V\begin{pmatrix}\Sigma^{-1} & 0\\ 0 & 0\end{pmatrix}_{n\times m}U^H. \tag{3.13}$$

证明： 不妨记

$$D=\begin{pmatrix}\Sigma & 0\\ 0 & 0\end{pmatrix}_{m\times n},\ \tilde{D}=\begin{pmatrix}\Sigma^{-1} & 0\\ 0 & 0\end{pmatrix}_{n\times m},\ X=V\tilde{D}U^H.$$

则有

$$\begin{aligned}
AXA &= (UDV^H)(V\tilde{D}U^H)(UDV^H) = UD\tilde{D}DV^H = UDV^H = A,\\
XAX &= (V\tilde{D}U^H)(UDV^H)(V\tilde{D}U^H) = V\tilde{D}D\tilde{D}U^H = X,\\
(AX)^H &= [(UDV^H)(V\tilde{D}U^H)]^H = (UD\tilde{D}U^H)]^H = UD\tilde{D}U^H\\
&= (UDV^H)(V\tilde{D}U^H) = AX,\\
(XA)^H &= [(V\tilde{D}U^H)(UDV^H)]^H = (V\tilde{D}DV^H)^H = V\tilde{D}DV^H\\
&= (V\tilde{D}U^H)(UDV^H) = XA.
\end{aligned}$$

所以 X 同时满足 A 的4个 Penrose 方程，因此 $X = A^+$。 证毕

例 3.7 设 $A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$，利用矩阵的奇异值分解求 A^+。

解：根据计算 A 的奇异值分解的方法可得 A 的奇异值分解为 $A = UDV^H$，其中

$$U = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{pmatrix},\ D = \begin{pmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},\ V^H = \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{pmatrix}.$$

因此

$$\begin{aligned}
A^+ &= \begin{pmatrix} \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} \\ \frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{pmatrix}\\
&= \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix}.
\end{aligned}$$

(2) 利用 A 的满秩分解计算 A^+

定理 3.14 设 $A \in \boldsymbol{C}_r^{m\times n}$ 的一个满秩分解为 $A = FG$，其中 $F \in \boldsymbol{C}_r^{m\times r}$，$G \in \boldsymbol{C}_r^{r\times n}$。则

$$A^+ = G^H(GG^H)^{-1}(F^HF)^{-1}F^H = G^H(F^HAG^H)^{-1}F^H. \tag{3.14}$$

证明：由于 $F \in \boldsymbol{C}_r^{m\times r}$，$G \in \boldsymbol{C}_r^{r\times n}$，因此 F^HF、GG^H 和 $F^HAG^H \in \boldsymbol{C}_r^{r\times r}$ 都是 r 阶可逆方阵，直接根据 A^+ 的定义可以验证 (3.14) 满足 A 的4个 Penrose 方程。证毕

根据这个定理，可以得到下面的两个推论:

推论 3.2 设 $A\in C^{m\times n}$，则有

(1) 当 $\mathrm{rank}A=m$ 时，$A^{+}=A^{H}(AA^{H})^{-1}$；

(2) 当 $\mathrm{rank}A=n$ 时，$A^{+}=(A^{H}A)^{-1}A^{H}$.

证明：(1) 当 $\mathrm{rank}A=m$ 时，$A=I_mA$，所以 $A^{+}=A^{H}(AA^{H})^{-1}(I_m^{H}I_m)^{-1}I_m^{H}=A^{H}(AA^{H})^{-1}$. (2) 当 $\mathrm{rank}A=n$ 时，$A=AI_n$，所以 $A^{+}=I_n^{H}(I_nI_n^{H})^{-1}(A^{H}A)^{-1}A^{H}=(A^{H}A)^{-1}A^{H}$. 因此推论结论成立。 证毕

推论 3.3 设 $\boldsymbol{a},\boldsymbol{b}\in C^{n}$，且 $\boldsymbol{a}\neq 0,\boldsymbol{b}\neq 0$，则

(1) $(\boldsymbol{ab}^{\mathrm{H}})^{+}=\frac{1}{\boldsymbol{a}^{\mathrm{H}}\boldsymbol{ab}^{\mathrm{H}}\boldsymbol{b}}\boldsymbol{ba}^{\mathrm{H}}$;

(2) $\boldsymbol{a}^{+}=\frac{1}{\boldsymbol{a}^{\mathrm{H}}\boldsymbol{a}}\boldsymbol{a}^{\mathrm{H}}$，$(\boldsymbol{b}^{\mathrm{H}})^{+}=\frac{1}{\boldsymbol{b}^{\mathrm{H}}\boldsymbol{b}}\boldsymbol{b}$.

证明：(1) $(\boldsymbol{ab}^{\mathrm{H}})^{+}=\boldsymbol{b}(\boldsymbol{b}^{\mathrm{H}}\boldsymbol{b})^{-1}(\boldsymbol{a}^{\mathrm{H}}\boldsymbol{a})^{-1}\boldsymbol{a}^{\mathrm{H}}=\frac{1}{\boldsymbol{a}^{\mathrm{H}}\boldsymbol{ab}^{\mathrm{H}}\boldsymbol{b}}\boldsymbol{ba}^{\mathrm{H}}$。 (2) $\boldsymbol{a}\in\boldsymbol{C}^{n}$，$\mathrm{rank}\boldsymbol{a}=1$，由推论 3.2，$\boldsymbol{a}^{+}=\frac{1}{\boldsymbol{a}^{\mathrm{H}}\boldsymbol{a}}\boldsymbol{a}^{\mathrm{H}}$。同理，$\boldsymbol{b}\in\boldsymbol{C}^{n}$，$\mathrm{rank}\boldsymbol{b}^{\mathrm{H}}=1$，由推论 3.2，$(\boldsymbol{b}^{\mathrm{H}})^{+}=\frac{1}{\boldsymbol{b}^{\mathrm{H}}\boldsymbol{b}}\boldsymbol{b}$。 证毕

例 3.8 设 $A=\begin{pmatrix}1&0&0\\0&1&-1\\1&0&0\\2&1&-1\end{pmatrix}$，利用矩阵的满秩分解求 A^{+}。

解：矩阵 A 的一个满秩分解为

$$A=FG=\begin{pmatrix}1&0\\0&1\\1&0\\2&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&-1\end{pmatrix}.$$

因此

$$A^{+}=G^{H}(F^{H}AG^{H})^{-1}F^{H}$$

$$=\begin{pmatrix}1&0\\0&1\\0&-1\end{pmatrix}\left(\begin{pmatrix}1&0&1&2\\0&1&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&-1\\1&0&0\\2&1&-1\end{pmatrix}\begin{pmatrix}1&0\\0&1\\0&-1\end{pmatrix}\right)^{-1}\begin{pmatrix}1&0&1&2\\0&1&0&1\end{pmatrix}$$

$$=\begin{pmatrix}1&0\\0&1\\0&-1\end{pmatrix}\begin{pmatrix}6&4\\2&4\end{pmatrix}^{-1}\begin{pmatrix}1&0&1&2\\0&1&0&1\end{pmatrix}=\frac{1}{8}\begin{pmatrix}2&-2&2&2\\-1&3&-1&1\\1&-3&1&-1\end{pmatrix}.$$

在直接解法中，往往需要对 A 进行各种形式的分解，但实际计算中这种分解是

比较困难的。因此，在实际计算中，常常是用迭代法来计算 A^+ 的近似矩阵。

2. 计算 A^+ 的迭代方法

(1) 计算 A^+ 的矩阵迭代方法

定理 3.15 对于任意矩阵 $A \in C_r^{m\times n}$，设 $\sigma_1,\sigma_2,\cdots\sigma_r$ 为 A 的 r 个非零奇异值，记 $c=\max\{\sigma_1^2,\sigma_2^2,\cdots\sigma_r^2\}$，则当 $0<\alpha<\frac{2}{c}$ 时，由迭代格式

$$\begin{cases} X_0=\alpha A^H \\ X_k=X_{k-1}+\alpha(I-\alpha A^HA)^kA^H,\ k=1,2,\cdots \end{cases} \tag{3.15}$$

得到的迭代序列 $\{X_k\}$ 收敛于 A^+，即

$$A^+=\sum_{k=0}^{+\infty}\alpha(I-\alpha A^HA)^kA^H. \tag{3.16}$$

证明： 由迭代格式 (3.15) 可知

$$\lim_{k\to+\infty}X_k=\sum_{k=0}^{+\infty}\alpha(I-\alpha A^HA)^kA^H$$

对于 $A\in C_r^{m\times n}$，存在酉矩阵 $U\in C_m^{m\times m}$ 和 $V\in C_n^{n\times n}$，形成 A 的奇异值分解

$$A=UDV^H$$

式中 $D=\begin{pmatrix}\sum & 0\\ 0 & 0\end{pmatrix}$，$\sum=\mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_r)$，$\sigma_1,\sigma_2,\cdots,\sigma_r$ 为 A 的 r 个非零奇异值。因此

$$A^HA=(UDV^H)^H(UDV^H)=VD^HDV^H.$$

从而有

$$\begin{aligned}\alpha(I-\alpha A^HA)^kA^H&=\alpha[V(I-\alpha D^HD)V^H]^k(UDV^H)^H\\&=V[\alpha(I-\alpha D^HD)^kD^H]U^H,\end{aligned}$$

即

$$\sum_{k=0}^{+\infty}\alpha(I-\alpha A^HA)^kA^H=V\left[\sum_{k=0}^{+\infty}\alpha(I-\alpha D^HD)^kD^H\right]U^H.$$

记

$$W_l=\sum_{k=0}^{l}\alpha(I-\alpha D^HD)^kD^H=\begin{pmatrix}B_l & 0\\ 0 & 0\end{pmatrix}_{n\times m}$$

其中 $B_l = \sum^{-1} - \sum^{-1}\mathrm{diag}((1-\alpha\sigma_1^2)^l, (1-\alpha\sigma_2^2)^l, \cdots (1-\alpha\sigma_r^2)^l)$. 如果记 $c = \max\{\sigma_1^2, \sigma_2^2, \cdots \sigma_r^2\}$，则当 $0 < \alpha < \frac{2}{c}$ 时，有 $|1-\alpha\sigma_i^2| < 1, (i = 1, 2, \cdots, r)$，从而有

$$\lim_{l\to+\infty} W_l = W = \begin{pmatrix} \sum^{-1} & 0 \\ 0 & 0 \end{pmatrix}_{n\times m}.$$

所以

$$\sum_{k=0}^{+\infty} \alpha(I-\alpha A^H A)^k A^H = V \begin{pmatrix} \sum^{-1} & 0 \\ 0 & 0 \end{pmatrix}_{n\times m} U^H.$$

根据基于奇异值分解的 A^+ 的计算公式 (3.13)，可得

$$\sum_{k=0}^{\infty} \alpha(I-\alpha A^H A)^k A^H = A^+,$$

即迭代格式 (3.15) 得到的迭代序列 $\{X_k\}$ 收敛于 A^+。 证毕

根据迭代格式(3.15) 还可以得到下面的一些计算 A^+ 的其他矩阵迭代方法：

推论 3.4 *在定理 3.15 的条件下，由迭代格式*

$$\begin{cases} \tilde{X}_0 = \alpha A^H \\ \tilde{X}_k = \tilde{X}_{k-1} + \alpha A^H (I-\alpha A A^H)^k, \ k = 1, 2, \cdots \end{cases} \tag{3.17}$$

得到的迭代序列$\{\tilde{X}_k\}$收敛于A^+，即

$$A^+ = \sum_{k=0}^{+\infty} \alpha A^H (I-\alpha A A^H)^k \tag{3.18}$$

证明： 由定理3.15可得

$$(A^H)^+ = \sum_{k=0}^{+\infty} \alpha(I-\alpha A A^H)^k A.$$

容易验证 $(A^H)^+ = (A^+)^H$ （只需验证 $(A^+)^H$ 满足 A^H 的4个 Penrose 方程即可），从而有

$$A^+ = [(A^+)^H]^H = [(A^H)^+]^H = \sum_{k=0}^{\infty} \alpha A^H (I-\alpha A A^H)^k.$$

证毕

推论 3.5 *在定理 3.15 的条件下，由迭代格式*

$$\begin{cases} X_0 = \alpha A^H \\ X_k = X_{k-1} + X_0(I - A X_{k-1}), \ k = 1, 2, \cdots \end{cases} \tag{3.19}$$

得到的迭代序列 $\{X_k\}$ 也收敛于 A^+。

证明：事实上，若记 $R=I-\alpha A^HA$，则迭代格式(3.15)可表示为

$$X_k=(I+R+R^2+\cdots+R^k)X_0.$$

由于

$$(I-R)(I+R+R^2+\cdots+R^k)X_0=(I-R^{k+1})X_0$$

即

$$(I-R)X_k=X_0-R^{k+1}X_0,$$

所以

$$R^{k+1}X_0=X_0-(I-R)X_k=X_0-\alpha A^HAX_k=X_0-X_0AX_k.$$

从而

$$\begin{aligned}X_{k+1}&=(I+R+R^2+\cdots+R^{k+1})X_0\\&=X_k+R^{k+1}X_0=X_k+X_0(I-AX_k),\end{aligned}$$

即迭代格式(3.15)也可表示为

$$\begin{cases}X_0=\alpha A^H\\X_k=X_{k-1}+X_0(I-AX_{k-1}),\ k=1,2,\cdots\end{cases}.$$

故迭代格式 (3.15) 和 迭代格式 (3.19) 是完全等价的。在定理 3.15 的条件下，迭代格式 (3.15) 收敛，因此迭代格式 (3.19) 也收敛。 证毕

定理 3.16 *在定理 3.15 的条件下，由迭代格式*

$$\begin{cases}Z_0=\alpha A^H\\Z_{k+1}=Z_k(2I-AZ_k),\ k=0,1,2,\cdots\end{cases}\tag{3.20}$$

得到的迭代序列 $\{Z_k\}$ 也收敛于 A^+。

证明： 事实上，迭代格式(3.19)得到的迭代序列 $\{X_k\}$ 和迭代格式 (3.20) 得到的迭代序列$\{Z_k\}$有如下关系

$$Z_k=X_{2^k-1},k=0,1,2,\cdots$$

可用数学归纳法进行证明：

(1) 当 $k=0$ 时，$Z_0=X_0$ 成立;

(2) 当 $k=1$ 时，$Z_1=Z_0(2I-AZ_0)=X_0+X_0(I-AX_0)=X_1$;

(3) 当 $k=2$ 时，利用推论 3.5 证明中的记法 $R=I-\alpha A^HA$，有

$$\begin{aligned}Z_2&=Z_1(2I-AZ_1)=Z_1+Z_1(I-AZ_1)\\&=X_1+X_1(I-AX_1)=(I+R)X_0+(I+R)X_0(I-AX_1)\\&=(I+R)X_0+(I+R)R^2X_0=(I+R+R^2+R^3)X_0=X_3=X_{2^2-1},\end{aligned}$$

即 $k=2$ 时，$Z_k=X_{2^k-1}$ 也成立；

(4) 假设 $Z_k=X_{2^k-1}$ 成立，则对于 Z_{k+1} 有

$$\begin{aligned}
&Z_{k+1}=Z_k(2I-AZ_k)=X_{2^k-1}+X_{2^k-1}(I-AX_{2^k-1})\\
&=(I+R+\cdots+R^{2^k-1})X_0+(I+R+\cdots+R^{2^k-1})X_0(I-AX_{2^k-1})\\
&=(I+R+R^2+\cdots+R^{2^k-1})X_0+(I+R+R^2+\cdots+R^{2^k-1})R^{2^k}X_0\\
&=(I+R+R^2+\cdots+R^{2^{(k+1)}-1})X_0=X_{2^{(k+1)}-1}.
\end{aligned}$$

由数学归纳法，$Z_k=X_{2^k-1}$成立。

这表明，迭代格式 (3.20) 得到的序列 $\{Z_k\}$ 实际上是迭代格式 (3.19) 得到的序列 $\{X_k\}$ 的一个子序列，因此在满足定理 3.15 的条件下，由迭代格式 (3.20) 得到的序列 $\{Z_k\}$ 也收敛于 A^+。 证毕

该定理的证明还表明，迭代格式 (3.20) 第 k 次迭代得到的 A^+ 的近似矩阵等于迭代格式 (3.19) 迭代 2^k-1 次迭代得到的 A^+ 的近似矩阵，显然迭代格式 (3.20) 的收敛速度比迭代格式(3.19)要快得多。

(2) 计算 A^+ 的 Greville 递推法

计算 A^+ 的矩阵迭代格式都需要计算矩阵的乘法，当矩阵规模较大时，这类方法的计算量非常大。1960年，Greville 提出了一种只需要计算矩阵和向量乘法的递推方法。

定理 3.17 设 $A=(\boldsymbol{a}_1,\cdots,\boldsymbol{a}_n)\in C^{m\times n}$，记 $A_k\in C^{m\times k}$ 为 A 的前 k 列组成的矩阵，并将 A_k 分块为

$$A_k=(A_{k-1},\boldsymbol{a}_k),\quad k=1,2,\cdots,n$$

其中 $A_1=\boldsymbol{a}_1$，$A_n=A$。则有

$$A_k^+=\begin{pmatrix}A_{k-1}^+-\boldsymbol{d}_k\boldsymbol{b}_k^H\\ \boldsymbol{b}_k^H\end{pmatrix},\quad k=2,\cdots,n \tag{3.21}$$

其中

$$\boldsymbol{d}_k=A_{k-1}^+\boldsymbol{a}_k,\quad \boldsymbol{c}_k=\boldsymbol{a}_k-A_{k-1}\boldsymbol{d}_k,$$

$$\boldsymbol{b}_k^H=\begin{cases}\boldsymbol{c}_k^+, & \boldsymbol{c}_k\neq 0\\ \dfrac{\boldsymbol{d}_k^HA_{k-1}^+}{1+\boldsymbol{d}_k^H\boldsymbol{d}_k}, & \boldsymbol{c}_k=0.\end{cases}$$

证明：设 $X_k=\begin{pmatrix}A_{k-1}^+-\boldsymbol{d}_k\boldsymbol{b}_k^H\\ \boldsymbol{b}_k^H\end{pmatrix}$

(1) 当 $\boldsymbol{c}_k = 0$ 时，有 $\boldsymbol{a}_k = A_{k-1}\boldsymbol{d}_k$，则

$$\begin{aligned} A_k X_k A_k &= (A_{k-1}, \boldsymbol{a}_k)\begin{pmatrix} A_{k-1}^+ - \boldsymbol{d}_k \boldsymbol{b}_k^H \\ \boldsymbol{b}_k^H \end{pmatrix}(A_{k-1}, \boldsymbol{a}_k) \\ &= (A_{k-1}A_{k-1}^+ - A_{k-1}\boldsymbol{d}_k\boldsymbol{b}_k^H + \boldsymbol{a}_k\boldsymbol{b}_k^H)(A_{k-1}, \boldsymbol{a}_k) \\ &= (A_{k-1}A_{k-1}^+ A_{k-1}, A_{k-1}A_{k-1}^+ A_{k-1}\boldsymbol{d}_k) \\ &= (A_{k-1}, A_{k-1}\boldsymbol{d}_k) = (A_{k-1}, \boldsymbol{a}_k) = A_k. \end{aligned}$$

同样可以验证 X_k 满足 A_k 的其它3个 Penrose 方程，因此 $X_k = A_k^+$。

(2) 当 $\boldsymbol{c}_k \neq 0$ 时，$\boldsymbol{b}_k^H = \boldsymbol{c}_k^+$。由于 $A_{k-1}^+\boldsymbol{c}_k = A_{k-1}^+(\boldsymbol{a}_k - A_{k-1}\boldsymbol{d}_k) = A_{k-1}^+\boldsymbol{a}_k - A_{k-1}^+A_{k-1}A_{k-1}^+\boldsymbol{a}_k = 0$，所以

$$\begin{aligned} 0 &= A_{k-1}^H A_{k-1}A_{k-1}^+\boldsymbol{c}_k\boldsymbol{c}_k^+(\boldsymbol{c}_k^+)^H = A_{k-1}^H(A_{k-1}A_{k-1}^+)^H(\boldsymbol{c}_k\boldsymbol{c}_k^+)^H(\boldsymbol{c}_k^+)^H \\ &= A_{k-1}^H(A_{k-1}^+)^H(A_{k-1})^H(\boldsymbol{c}_k^+)^H(\boldsymbol{c}_k)^H(\boldsymbol{c}_k^+)^H \\ &= (A_{k-1}A_{k-1}^+A_{k-1})^H(\boldsymbol{c}_k^+\boldsymbol{c}_k\boldsymbol{c}_k^+)^H = A_{k-1}^H(\boldsymbol{c}_k^+)^H = (\boldsymbol{c}_k^+A_{k-1})^H. \end{aligned}$$

因此 $\boldsymbol{c}_k^+A_{k-1} = 0$，从而有 $\boldsymbol{c}_k = \boldsymbol{c}_k\boldsymbol{c}_k^+\boldsymbol{c}_k = \boldsymbol{c}_k\boldsymbol{c}_k^+(\boldsymbol{a}_k - A_{k-1}\boldsymbol{d}_k) = \boldsymbol{c}_k\boldsymbol{c}_k^+\boldsymbol{a}_k$。根据 (3.14) 可知：$\boldsymbol{c}_k^+ = \frac{\boldsymbol{c}_k^H}{\boldsymbol{c}_k^H\boldsymbol{c}_k}$，因此有 $\boldsymbol{c}_k^+\boldsymbol{c}_k = 1$，从而 $\boldsymbol{c}_k^+\boldsymbol{a}_k = \boldsymbol{c}_k^+\boldsymbol{c}_k\boldsymbol{c}_k^+\boldsymbol{a}_k = \boldsymbol{c}_k^+\boldsymbol{c}_k = 1$，则

$$\begin{aligned} A_k X_k A_k &= (A_{k-1}, \boldsymbol{a}_k)\begin{pmatrix} A_{k-1}^+ - \boldsymbol{d}_k\boldsymbol{c}_k^+ \\ \boldsymbol{c}_k^+ \end{pmatrix}(A_{k-1}, \boldsymbol{a}_k) \\ &= (A_{k-1}A_{k-1}^+ + (\boldsymbol{a}_k - A_{k-1}\boldsymbol{d}_k)\boldsymbol{c}_k^+)(A_{k-1}, \boldsymbol{a}_k) \\ &= (A_{k-1} + \boldsymbol{c}_k(\boldsymbol{c}_k^+A_{k-1}), A_{k-1}A_{k-1}^+\boldsymbol{a}_k - A_{k-1}\boldsymbol{d}_k(\boldsymbol{c}_k^+\boldsymbol{a}_k) + \boldsymbol{a}_k(\boldsymbol{c}_k^+\boldsymbol{a}_k)) \\ &= (A_{k-1}, A_{k-1}\boldsymbol{d}_k - A_{k-1}\boldsymbol{d}_k + \boldsymbol{a}_k) = (A_{k-1}, \boldsymbol{a}_k) = A_k. \end{aligned}$$

同样可以验证 X_k 满足 A_k 的其它3个 Penrose 方程，因此 $X_k = A_k^+$。故 (3.21) 成立，从而有 $A^+ = A_n^+$。 证毕

Greville 提出的这种方法实际上是一种以迭代形式实现的计算 A^+ 的直接方法，在理论上由式 (3.21) 迭代 n 步就可以精确得到 A^+。从计算量的角度，这种方法只用到了矩阵和向量的乘法，并没有用到矩阵和矩阵的乘法，因此计算量要比其他的矩阵迭代方法小得多。

3.3 Moore-Penrose 逆的性质

矩阵 $A \in \boldsymbol{C}^{m\times n}$ 的 Moore-Penrose 逆 A^+ 同时满足4个 Penrose 方程，它具有所有广义逆矩阵的所有性质，同时它也具有一些自己特殊的性质，比如当A为满秩方

阵时， 容易验证 A^{-1} 同时满足4个 Penrose 方程，即此时 $A^{+}=A^{-1}$。 除此之外 A^{+} 还具有下面一些基本性质：

定理 3.18 设$A\in C^{m\times n}$，则 A 的 Moore-Penrose 逆 A^{+} 存在且唯一。

证明： 首先，由于 A 的奇异值分解是一定存在的，因此根据基于 A 的奇异值分解计算 A^{+} 的方法可知，A^{+}是一定存在的。 下面，证明唯一性。

假设 X,Y 均为 A 的 Moore-Penrose 逆，即 X,Y 都满足4个 Penrose 方程，则有

$$\begin{aligned}X&=XAX=X(AX)^{H}=X[(AYA)X]^{H}=X(AX)^{H}(AY)^{H}\\&=XAXAY=XAY=XA(YAY)=(XA)^{H}(YA)^{H}Y\\&=(YAXA)^{H}Y=(YA)^{H}Y=YAY=Y\end{aligned}$$

所以 A 的 Moore-Penrose 逆 A^{+} 是唯一的。 证毕

由于 A^{+} 同时满足所有4个Penrose方程，因此在15类广义逆矩阵中的每一类都包含 A^{+}，这也表明任意矩阵的15类广义逆矩阵都是存在的。 但需要指出的是，在所有 15 类广义逆矩阵中，除了 Moore-Penrose 逆是唯一的以外， 其余 14类广义逆矩阵都是不唯一的。

定理 3.19 设$A\in C^{m\times n}$，则 $A^{+}\in C^{n\times m}$ 具有下列一些性质：

(1) rankA^{+} = rankA;

(2) $(A^{+})^{+}=A$;

(3) $(A^{H})^{+}=(A^{+})^{H}$;

(4) $(A^{H}A)^{+}=A^{+}(A^{H})^{+}$, $(AA^{H})^{+}=(A^{H})^{+}A^{+}$;

(5) $A^{+}=(A^{H}A)^{+}A^{H}=A^{H}(AA^{H})^{+}$;

证明： (1) 由于 $AA^{+}A=A$，$A^{+}AA^{+}=A^{+}$，所以有 rank$A^{+}\geq$ rank$A$，且 rank$A\geq$ rankA^{+}，即 rankA^{+} = rankA。

(2) 只需验证 A 满足A^{+}的4个Penrose方程：

$A^{+}AA^{+}=A^{+}(AA^{+}A)A^{+}=(A^{+}AA^{+})AA^{+}=A^{+}AA^{+}=A^{+}$;

$AA^{+}A=A(A^{+}AA^{+})A=(AA^{+}A)A^{+}A=AA^{+}A=A$;

$(A^{+}A)^{H}=A^{+}A$ (因为A^{+}满足A的第4个Penrose方程);

$(AA^{+})^{H}=AA^{+}$ (因为A^{+}满足A的第3个Penrose方程),

因此$(A^{+})^{+}=A$。

(3) 同上，只需验证 A^{H} 满足 $(A^{+})^{H}$ 的4个Penrose方程即可。

(4) 同上，只需验证 $A^{H}A$ 满足 $A^{+}(A^{H})^{+}$ 的4个 Penrose 方程， AA^{H} 满足

$(A^H)^+A^+$ 的4个 Penrose 方程即可。

(5) 假设A的奇异值分解为

$$A = U\begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}_{m\times n} V^H$$

其中$\sum = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_r)$， $\sigma_1, \sigma_2, \cdots, \sigma_r$ 为 A的非零奇异值。 则有

$$A^HA = V\begin{pmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{pmatrix} V^H.$$

容易验证

$$(A^HA)^+ = V\begin{pmatrix} (\Sigma^2)^{-1} & 0 \\ 0 & 0 \end{pmatrix} V^H.$$

因此

$$\begin{aligned}(A^HA)^+A^H &= V\begin{pmatrix} (\Sigma^2)^{-1} & 0 \\ 0 & 0 \end{pmatrix} V^HV\begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} U^H \\ &= V\begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^H = A^+.\end{aligned}$$

同理可证 $A^H(AA^H)^+ = A^+$，因此 $A^+ = (A^HA)^+A^H = A^H(AA^H)^+$ 成立。 证毕

该定理也可直接得出推论 3.2，它是计算 Moore-Penrose 逆 A^+ 的一种直接方法。

通常意义下的逆矩阵 A^{-1} 实际上是 A^+ 的一种特殊情况， 凡是 A^+ 满足的性质，A^{-1} 肯定满足，但是反过来，A^{-1} 满足的性质， 对于一般情形下的 A^+ 并不一定成立。

例 3.9 设 $A = (1\ \ 1)$， $B = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$，求 $(AB)^+$ 和 B^+A^+。

解： $AB = (1)$，因此 $(AB)^+ = [(1)^H(1)]^{-1}(1)^H = (1)$;

由于A为行满秩矩阵，因此

$$A^+ = A^H(AA^H)^{-1} = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

而 B 为列满秩矩阵，因此

$$B^+ = (B^HB)^{-1}B^H = (1\ \ 0),$$

从而

$$B^+A^+ = (\frac{1}{2}) \neq (AB)^+.$$

这个例子说明，对于 A^{-1} 满足的性质 $(AB)^{-1}=B^{-1}A^{-1}$对于一般情况下的 A^{+} 并不一定成立。

除此之外，通常意义有 $AA^{-1}=A^{-1}A$，而对 $A\in C^{m\times n}$，$A^{+}\in C^{n\times m}$，因此 $AA^{+}\in C^{m\times m}$，而 $A^{+}A\in C^{n\times n}$，显然 $AA^{+}\neq A^{+}A$。

定理 3.20 设 $A\in C^{m\times n}$, $b\in C^{m}$, 则 $Ax=b$ 有解的充分必要条件是

$$AA^{+}b=b.$$

若$Ax=b$有解，其通解为

$$x=A^{+}b+(I-A^{+}A)y,\qquad (y\in C^{n}任意). \tag{3.22}$$

若方程解不唯一，则 $x_0=A^{+}b$ 为唯一的极小范数解，即

$$||x_0||_2=\min_{Ax=b}\|x\|_2.$$

证明： 首先，证明 $Ax=b$ 有解的充分必要条件是$AA^{+}b=b$.

一方面，若$AA^{+}b=b$ 成立，则显然$x=A^{+}b$ 是方程$Ax=b$的解。 另一方面，若$Ax=b$ 有解，则

$$b=Ax=AA^{+}Ax=AA^{+}b$$

因此，$AA^{+}b=b$是$Ax=b$有解的充分必要条件。

其次，在有解的情况下，证明式 (3.22) 是方程 $Ax=b$的通解。

一方面，将式 (3.22) 代入 Ax 可算出它等于 b. 这说明凡是用 (3.22) 表示的 x 都是 $Ax=b$ 的解；另一方面，设 x_0 是 $Ax=b$ 的任一解， 则有

$$\begin{aligned}x_0&=A^{+}b+x_0-A^{+}b\\&=A^{+}b+x_0-A^{+}Ax_0\\&=A^{+}b+(I-A^{+}A)x_0\end{aligned}$$

即 x_0 可表示为 (3.22) 的形式，故 (3.22) 为 $Ax=b$ 的通解。

下面证明方程的解不唯一时，$x_0=A^{+}b$ 为唯一的极小范数解。

首先，对于方程 $A\boldsymbol{x}=\boldsymbol{b}$ 的解 $\boldsymbol{x}$，有

$$\begin{aligned}||\boldsymbol{x}||_2&=\boldsymbol{x}^H\boldsymbol{x}\\&=[A^+\boldsymbol{b}+(I-A^+A)\boldsymbol{y}]^H[A^+\boldsymbol{b}+(I-A^+A)\boldsymbol{y}]\\&=||A^+\boldsymbol{b}||_2^2+||(I-A^+A)\boldsymbol{y}||_2^2\\&\quad+(A^+\boldsymbol{b})^H(I-A^+A)\boldsymbol{y}+[(I-A^+A)\boldsymbol{y}]^HA^+\boldsymbol{b}\\&=||A^+\boldsymbol{b}||_2^2+||(I-A^+A)\boldsymbol{y}||_2^2\\&\quad+\boldsymbol{b}^H[(A^+)^H-(A^+)^HA^+A]\boldsymbol{y}+\boldsymbol{y}^H[A^+-(A^+A)^HA^+]\boldsymbol{b}\\&=||A^+\boldsymbol{b}||_2^2+||(I-A^+A)\boldsymbol{y}||_2^2.\end{aligned}$$

对于任意 $\boldsymbol{y}\in\boldsymbol{C}^n$，都有 $||\boldsymbol{x}||_2\geq||A^+\boldsymbol{b}||_2$，因此 $\boldsymbol{x}_0=A^+\boldsymbol{b}$ 为极小范数解。

然后证明唯一性。设除了 $\boldsymbol{x}_0=A^+\boldsymbol{b}$ 为极小范数解，还有 $\boldsymbol{x}_1$ 为极小范数解，则显然有 $||\boldsymbol{x}_0||_2=||\boldsymbol{x}_1||_2=||A^+\boldsymbol{b}||_2$. 对于 $\boldsymbol{x}_1$，一定存在 $\boldsymbol{y}_1\in\boldsymbol{C}^n$ 使得

$$\boldsymbol{x}_1=A^+\boldsymbol{b}+(\mathrm{I}-A^+A)\boldsymbol{y}_1.$$

根据前面的推导过程可知

$$||\boldsymbol{x}_1||_2^2=||A^+\boldsymbol{b}||_2^2+||(I-A^+A)\boldsymbol{y}_1||_2^2.$$

要使 $||\boldsymbol{x}_0||_2=||\boldsymbol{x}_1||_2=||A^+\boldsymbol{b}||_2$，只有 $(I-A^+A)\boldsymbol{y}_1=0$，从而

$$\boldsymbol{x}_1=A^+\boldsymbol{b}=\boldsymbol{x}_0.$$

因此 $\boldsymbol{x}_0=A^+\boldsymbol{b}$ 是唯一的极小范数解。 证毕

例 3.10 用广义逆矩阵A^+方法判断线性方程组

$$\begin{cases}x_1+x_2-x_3=1\\2x_1+x_2+x_3=2\\3x_1+2x_2=3\end{cases}$$

是否有解？若有解，求出其极小范数解。

解：令

$$A=\begin{pmatrix}1&1&-1\\2&1&1\\3&2&0\end{pmatrix},\ \boldsymbol{b}=\begin{pmatrix}1\\2\\3\end{pmatrix}.$$

容易求得A的满秩分解为

$$A=FG=\begin{pmatrix}1&1\\2&1\\3&2\end{pmatrix}\begin{pmatrix}1&0&2\\0&1&-3\end{pmatrix}$$

所以

$$A^{+}=G^{H}(GG^{H})^{-1}(F^{H}F)^{-1}F^{H}=\frac{1}{42}\begin{pmatrix}0&6&6\\7&-2&5\\-21&18&-3\end{pmatrix}$$

可以验证

$$AA^{+}b=(1,2,3)^{T}=\boldsymbol{b},$$

所以原方程组有解，且极小范数解为

$$\boldsymbol{x}_0=A^{+}\boldsymbol{b}=\frac{1}{42}\begin{pmatrix}0&6&6\\7&-2&5\\-21&18&-3\end{pmatrix}\begin{pmatrix}1\\2\\3\end{pmatrix}=\frac{1}{7}\begin{pmatrix}5\\3\\1\end{pmatrix}$$

第三章习题

1. 求$A=\begin{pmatrix}2&1&-3\\2&5&-4\\6&7&3\end{pmatrix}$的Doolittle分解。

2. 求$A=\begin{pmatrix}1&2&-1\\2&\frac{2}{3}&\frac{1}{3}\\3&1&1\end{pmatrix}$的选列主元Doolittle分解。

3. 求$A=\begin{pmatrix}2.25&-3&4.5\\-3&5&-10\\4.5&-10&34\end{pmatrix}$的Cholesky分解。

4. 求下列矩阵的Hermite标准形和变换矩阵S，并求满秩分解

$$A=\begin{pmatrix}2&4&1&1\\1&2&-1&2\\-1&-2&-2&1\end{pmatrix},\qquad B=\begin{pmatrix}1&2&3&6\\2&4&3&6\\1&2&3&6\\2&4&6&12\end{pmatrix}.$$

5. 设$A=\begin{pmatrix}1&0&1\\0&1&1\\0&0&0\end{pmatrix}$，

   (1) 求 A 的奇异值分解；

   (2) U 为 AA^H 的正交单位化特征向量组成的酉矩阵，V 为 A^HA 正交单位化特征向量组成的酉矩阵， 试验证能否由此构成 A 的奇异值分解。

6. 设$A\in C^{m\times n}$，证明：

   (a) $(A^H)^+=(A^+)^H$；

   (b) $(A^HA)^+=A^+(A^H)^+$，$(AA^H)^+=(A^H)^+A^+$.

7. 设 A 是一个正规矩阵，证明：

   (a) $AA^+=A^+A$.

   (b) 对于任意自然数 k，有 $(A^k)^+=(A^+)^k$.

8. 设$A=\begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_s \end{pmatrix}$，证明：$A^+=\begin{pmatrix} A_1^+ & & & \\ & A_2^+ & & \\ & & \ddots & \\ & & & A_s^+ \end{pmatrix}$.

9. 利用矩阵的奇异值分解求下面矩阵的 Moore-Penrose逆 A^+:

(1) $A=\begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$; (2) $A=\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$

(3) $A=\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$.

10. 利用矩阵的满秩分解求下面矩阵的 Moore-Penrose逆 A^+:

(1) $A=\begin{pmatrix} 1 & 0 & 2 \\ 2 & 1 & 5 \\ 0 & 1 & -1 \\ 1 & 3 & -1 \end{pmatrix}$; (2) $A=\begin{pmatrix} 4 & -1 & -3 & -2 \\ -2 & 5 & -1 & -3 \\ 2 & 13 & -9 & -5 \end{pmatrix}$;

(3) $A=\begin{pmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 5 \\ -1 & -3 & 3 & 0 \end{pmatrix}$; (4) $A=\begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 2 & 1 & 1 \end{pmatrix}$.

11. 用广义逆矩阵A^+方法判断线性方程组

$$\begin{cases} 2x_1+4x_2+x_3+x_4=3 \\ x_1+2x_2-x_3+2x_4=0 \\ -x_1-2x_2-2x_3+x_4=3 \end{cases}$$

是否有解？如果有解，求通解和极小范数解。

第 四 章　线性方程组数值解法

在线性代数中一个基本问题是求线性方程组

$$Ax = b \tag{4.1}$$

的解 $x \in C^n$， 其中 $A \in C^{m\times n}$, $b \in C^m$ 已知。本章讨论(4.1)的数值解法。求解线性方程组是科学与工程计算的中心问题。很多科学技术问题及其相关的数学问题，例如微分方程和非线性代数方程组的数值求解都需要求解线性方程组，而且相应方程通常是要求未知变量很多的大型线性方程组。 所以有必要寻求高效的适合计算机运算的数值解法。

在线性代数课程中，曾经学习过用 Gauss 消去法求解 (4.1)，在没有舍入误差的假设下，这种方法在有限步产生方程的解，我们称之为直接方法。由于方程组的系数矩阵 A 通常阶数很大，有的结构具有特殊的性质，我们还可以采用更加高效的解法。本章将进一步介绍直接方法， 并介绍迭代法、极小化方法等求解方法。

4.1　线性方程组的直接解法

线性方程组的直接解法主要是 Gauss 消去法及其变种，包括主元素法和三角分解方法等。

4.1.1　Gauss 消去法

设 $A = (a_{ij}) \in C^{n\times n}$， $b = (b_1, b_2, \cdots, b_n)^T \in C^n$， 求解线性方程组

$$Ax = b. \tag{4.2}$$

如果 A 是一个下三角矩阵，即对于 $j > i$, $a_{ij} = 0$, 而且 $a_{ii} \neq 0, i = 1, 2, \cdots, n$，我们可以从第一个方程求得 x_1, 代入第二个方程求得 x_2，这样逐次向前代入求出 x 的所有分量。如果 A 是一个上三角矩阵，即对于 $i > j$, $a_{ij} = 0$, 而且 $a_{ii} \neq 0, i = 1, 2, \cdots, n$， 我们可以从最后一个方程求得 x_n, 代入第 $n-1$ 个方程求得 x_{n-1}，这样逐次向后回代求出 x 的所有分量。Gauss 消去法的思想就是将系数矩阵 A 逐次消去成上三角矩阵， 再逐次向后回代求出方程组的解。

第一步：从第二个到第 n 个方程消去未知量 x_1，其具体作法是：设 $a_{11}\neq 0$，令 $l_{i1}=-\frac{a_{i1}}{a_{11}}$ $(i=2,\cdots,n)$，计算

$$\begin{cases} a_{i1}:=0; \\ a_{ij}:=a_{ij}+l_{i1}a_{1j}, \quad j=2,3,\cdots,n; \\ b_i:=b_i+l_{i1}b_1. \end{cases}$$

用矩阵运算表示，完成这一步相当于对（4.2）的增广矩阵 $(A^{(1)},\boldsymbol{b}^{(1)})=(A,\boldsymbol{b})$进行初等行变换，分别让其第 i $(i=2,3,\cdots,n)$ 行加上第一行的 l_{i1}倍，即

$$\left(A^{(1)},\boldsymbol{b}^{(1)}\right)\longrightarrow\left(A^{(2)},\boldsymbol{b}^{(2)}\right)=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ 0 & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2} & \cdots & a_{nn} & b_n \end{pmatrix}.$$

注意 a_{ij} 和 b_i 在计算过程中已改变。

第一步也等价于用初等矩阵

$$L_1=\begin{pmatrix} 1 & & & \\ l_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & 0 & \cdots & 1 \end{pmatrix}$$

左乘(4.2)的两端得到

$$L_1A\boldsymbol{x}=L_1\boldsymbol{b}.$$

这样，（4.2）变成如下等价的线性方程组

$$A^{(2)}\boldsymbol{x}=\boldsymbol{b}^{(2)}. \tag{4.3}$$

第二步：设方程组 (4.3)中 $a_{22}\neq 0$，从第三个方程到第 n 个方程消去未知量 x_2：

$$\begin{cases} l_{i2}:=-\frac{a_{i2}}{a_{22}}; \\ a_{i2}:=0; \\ a_{ij}:=a_{ij}+l_{i2}a_{2j}; \\ b_i:=b_i+l_{i2}b_2, \quad i,j=3,4,\cdots,n. \end{cases}$$

完成这一步相当于对（4.3）的增广矩阵进行初等行变换，分别让其第 i $(i=3,4\cdots,n)$ 行加上第二行的 l_{i2}倍，即

$$\left(A^{(2)},\boldsymbol{b}^{(2)}\right)\longrightarrow\left(A^{(3)},\boldsymbol{b}^{(3)}\right)=\begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} & b_1 \\ 0 & a_{22} & a_{23} & \cdots & a_{2n} & b_2 \\ 0 & 0 & a_{33} & \cdots & a_{3n} & b_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & a_{n3} & \cdots & a_{nn} & b_n \end{pmatrix}.$$

第二步也等价于用初等矩阵

$$L_2=\begin{pmatrix}1&&&&\\0&1&&&\\0&l_{32}&1&&\\\vdots&\vdots&\vdots&\ddots&\\0&l_{n2}&0&\cdots&1\end{pmatrix}$$

左乘(4.3)的两端得到

$$L_2L_1A\boldsymbol{x}=L_2L_1\boldsymbol{b}.$$

这样，(4.2）变成等价的线性方程组

$$A^{(3)}\boldsymbol{x}=\boldsymbol{b}^{(3)}.$$

一般地，第 k 步：设$a_{kk}\neq 0$，从第 $k+1$个方程到第 n 个方程消去未知量 x_k：

$$\begin{cases}l_{ik}:=-\frac{a_{ik}}{a_{kk}};\\a_{ik}:=0;\\a_{ij}:=a_{ij}+l_{ik}a_{kj};\\b_i:=b_i+l_{ik}b_k,\quad i,j=k+1,\cdots,n.\end{cases}$$

完成这一步相当于对第 $k-1$ 步所形成的方程组的增广矩阵进行初等行变换， 分别让其第 i $(i=k+1,k+2\cdots,n)$ 行加上第 k 行的 l_{ik}倍。即

$$L_kL_{k-1}\cdots L_2L_1A\boldsymbol{x}=L_kL_{k-1}\cdots L_2L_1\boldsymbol{b},$$

其中

$$L_k=\begin{pmatrix}1&&&&&&\\&\ddots&&&&&\\0&\cdots&1&&&&\\0&\cdots&l_{k+1,k}&1&&&\\&&\vdots&&\ddots&&\\0&\cdots&l_{n,k}&0&\cdots&1\end{pmatrix}.$$

在完成 $n-1$ 步后，方程组（4.2）就化成与其等价的上三角方程组

$$A^{(n)}\boldsymbol{x}=\boldsymbol{b}^{(n)} \tag{4.4}$$

即

$$\left.\begin{aligned}a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n&=b_1\\+a_{22}x_2+\cdots+a_{2n}x_n&=b_2\\&\vdots\\a_{nn}x_n&=b_n\end{aligned}\right\}.$$

我们把这个过程称为 Gauss 消元过程。再利用回代过程从下到上逐次把 x_n, $x_{n-1},\cdots$, x_1 计算出来。

上述 Gauss 消去法中，未知量是按其出现于方程式中的自然顺序消去的，用来消去其它方程式中未知量的过程也是按自然顺序进行的，故称为顺序消去法。

总之，Gauss 消去法分两大过程进行：

（1） 消元过程：经过 $n-1$ 步消元将原方程组化成等价的上三角形方程组

$$A^{(n)}\boldsymbol{x}=\boldsymbol{b}^{(n)}.$$

其算法可描述为：

对 $k=1,2,\cdots,n-1$, 设$a_{kk}^{(k)}\neq 0$，对 $i=k+1,\cdots,n$ 计算

$$\begin{cases} l_{ik}=-a_{ik}^{(k)}/a_{kk}^{(k)};\\ a_{ij}^{(k+1)}=a_{ij}^{(k)}+l_{ik}a_{kj}^{(k)};\\ b_i^{(k+1)}=b_i^{(k)}+l_{ik}b_k^{(k)}, \quad j=k+1,\cdots,n. \end{cases}$$

(2) 回代过程：利用逐步回代，求解三角形方程组 (4.4)。算法可描述为：

$$\begin{cases} x_n=b_n^{(n)}/a_{nn}^{(n)};\\ x_k=\left(b_k^{(n)}-\sum\limits_{j=k+1}^{n}a_{kj}^{(k)}x_j\right)\Big/a_{kk}^{(k)}, \quad k=n-1,n-2,\cdots,1. \end{cases}$$

Gauss 消去法总的乘除法次数为

$$\frac{n^3}{3}+n^2-\frac{n}{3}\approx\frac{n^3}{3},$$

加减法次数为

$$\frac{n^3}{3}+\frac{n^2}{2}-\frac{5n}{6}\approx\frac{n^3}{3}.$$

如果我们用克莱姆法则计算（4.2）的解，要计算 $n+1$ 个 n 阶行列式并作 n 次除法。而计算每个行列式，若用子式展开的方法，则有 $n!$ 次乘法，所以用克莱姆法则大概需要 $(n+1)!$ 次乘除法运算。例如，当 $n=10$ 时，约需 4×10^7 次乘除法运算, 而用 Gauss 消去法只需 430次乘除法运算。

从上面消去法过程可以看出，Gauss 消去步骤能够顺序进行的条件是主元素 $a_{11}^{(1)},a_{22}^{(2)},\cdots,a_{n-1,n-1}^{(n-1)}$ 全不为0，回代步骤还要求 $a_{nn}^{(n)}\neq 0$. 若用 $\triangle_i$ 表示 A 的顺序主子式，即

$$\triangle_i=\begin{vmatrix} a_{11} & \cdots & a_{1i}\\ \vdots & & \vdots\\ a_{i1} & \cdots & a_{ii} \end{vmatrix}, \quad i=1,2,\cdots,n,$$

有下面定理：

定理 4.1 $a_{ii}^{(i)}(i=1,2,\cdots,k)$ 全不为 0 的充要条件是 A 的顺序主子式 $\triangle_i \neq 0$, $i=1,2,\cdots,k$，其中 $k \leq n$。

定理 4.2 对于方程组 (4.2), 若 A 的的顺序主子式均不为零，则可用 Gauss 消去法求出方程组的解。

如果 A 非奇异，方程组(4.2)就有唯一解。但 A 的顺序主子式不一定都非零。例如，若 $a_{11}=0$, 消去法的第一步就不能进行。但我们可在 A 的第1列找出一个非零元 a_{i1}，先进行将第1行与第 i 行交换，然后作第一步消去法。其它各步类似处理。

有时虽然 $a_{ii}^{(i)} \neq 0$，但 $|a_{ii}^{(i)}|$ 很小，消去法可以计算下去，但计算过程中的舍入误差可能导致结果不可靠。

例 4.1 用3位十进制浮点运算求解

$$\begin{cases} 1.00 \times 10^{-5}x_1 + 1.00x_2 = 1.00 \\ 1.00x_1 + 1.00x_2 = 2.00 \end{cases}$$

解： 这个方程组的准确解显然接近 $(1.00, 1.00)^T$。 但是系数 $a_{11} = 1.00 \times 10^{-5}$ 与其它系数比较是个较小的数， 如果我们用顺序Gauss 消去法求解，则有

$$\begin{aligned} l_{21} &= \frac{a_{21}}{a_{11}} = 1.00 \times 10^5 \\ a_{22}^{(2)} &= a_{22} - l_{21}a_{12} = 1.00 - 1.00 \times 10^5 \\ b_2^{(2)} &= b_2 - l_{21}b_1 = 2.00 - 1.00 \times 10^5 \end{aligned}$$

在 3 位十进制运算的限制下，得到

$$x_2 = b_2^{(2)}/a_{22}^{(2)} = 1.00$$

回代得 $x_1 = 0.00$， 显然这是不正确的解。因为用小数 a_{11} 做除数，使 l_{21} 是个大数，在计算 $a_{22}^{(2)}$ 的过程中 a_{22} 的值完全被掩盖了。如果先交换方程组的两个方程的顺序，再用 Gauss 消去法，就不会出现上述问题，解得 $x_1 = 1.00$, $x_2 = 1.00$。

为了保证计算过程对于舍入误差的数值稳定性，我们介绍主元素方法。

主元素法的基本思想是在消元的各步中，主元素不要太小，与该步的其它系数比较，选取绝对值最大的元素作为主元，常用的有部分主元素法和总体主元素法两种。

部分主元素消去法又称列主元素消去法，未知数仍然是按自然顺序消去，但是把各方程中要消去的那个未知数的系数中按模最大的作为主元。具体计算过程如下：

第一步：在 x_1 的系数中，即系数矩阵 A 的第一列中按模最大者作为主元，设

$$|a_{r1}| = \max_{1 \leq i \leq n} |a_{i1}|,$$

若有 k 个数 a_{r1} 都是按模最大者，则取 a_{r1} 中 r 值最小者作为主元，交换第一个方程和第 r 个方程，即把 a_{r1} 换在 a_{11} 的位置上，再按 Gauss 顺序消去法第一步消元。

第 k 步：在第 k 到第 n 个方程中找出未知数 x_k 系数中按模最大者，设 $|a_{r_k k}| = \max\limits_{k\le i\le n} |a_{ik}|$，交换第 k 个方程和第 r_k 个方程，按 Gauss 顺序消去法第 k 步进行消元。

由于这种选主元每一步是在"余下"方程组的第一列中选主元，故称列主元素法。

设 A 是 $n\times n$ 矩阵，ρ_k 是行交换指标，部分主元素算法如下：对于 $k = 1,2,\cdots,n$，

1）找 $\rho_k \ge k$，使得 $|a_{\rho_k k}| = \max\limits_{k\le i\le n} |a_{ik}|$.

2）如果 $a_{\rho_k k} = 0$，停止计算，否则

3）$a_{kj} \leftrightarrow a_{\rho_k j}$，$j = k, k+1,\cdots,n$，

$b_k \leftrightarrow b_{\rho_k}$，

4）计算 $l_{ik} = a_{ik}/a_{kk}$，$i = k+1, k+2,\cdots,n$，

5）$a_{ij} = a_{ij} - l_{ik}a_{kj}$，$i,j = k+1,k+2,\cdots,n$，

$b_i = b_i - l_{ik}b_k$，$i = k+1,k+2,\cdots,n$。

总体主元素消去法，未知数不再按顺序消去。每消去一个未知数之前，例如要消去第 k 个未知数，先在第 k 个至第 n 个方程中及在第 k 个至第 n 个未知数的系数中找出按模最大者作为主元。然后交换行和列的位置，再按 Gauss 顺序消去法第 k 步消元。总体主元素消去法比列主元素消去法运算量大得多，可以证明列主元素消去法的舍入误差一般已较小，所以实际计算中多用列主元素消去法。

例 4.2 用列主元素法解方程组

$$\begin{cases} -3x_1 + 2x_2 + 6x_3 = 4 \\ 10x_1 - 7x_2 = 7 \\ 5x_1 - x_2 + 5x_3 = 6. \end{cases}$$

解：对增广矩阵按列选取主元进行 Gauss 消元法

$$\begin{pmatrix} -3 & 2 & 6 & 4 \\ 10 & -7 & 0 & 7 \\ 5 & -1 & 5 & 6 \end{pmatrix} \to \begin{pmatrix} 10 & -7 & 0 & 7 \\ -3 & 2 & 6 & 4 \\ 5 & -1 & 5 & 6 \end{pmatrix} \to \begin{pmatrix} 10 & -7 & 0 & 7 \\ 0 & -\frac{1}{10} & 6 & \frac{61}{10} \\ 0 & \frac{5}{2} & 5 & \frac{5}{2} \end{pmatrix} \to$$

$$\begin{pmatrix} 10 & -7 & 0 & 7 \\ 0 & \frac{5}{2} & 5 & \frac{5}{2} \\ 0 & -\frac{1}{10} & 6 & \frac{61}{10} \end{pmatrix} \to \begin{pmatrix} 10 & -7 & 0 & 7 \\ 0 & \frac{5}{2} & 5 & \frac{5}{2} \\ 0 & 0 & \frac{31}{5} & \frac{31}{5} \end{pmatrix}.$$

回代得解 $x_3 = 1,\ x_2 = -1,\ x_1 = 0$。

4.1.2 直接三角分解解法

从上面 Gauss 顺序消去法的过程，我们看到

$$L_{n-1}L_{n-2}\cdots L_2L_1A = A^{(n)}$$

即

$$A = L_1^{-1}L_2^{-1}\cdots L_{n-2}^{-1}L_{n-1}^{-1}A^{(n)}.$$

令 $L = L_1^{-1}L_2^{-1}\cdots L_{n-2}^{-1}L_{n-1}^{-1},\ R = A^{(n)}$，则 L 是单位下三角矩阵，R 是上三角矩阵，

$$A = LR$$

是矩阵 A 的Doolittle分解。

将 A 分解成 一个上三角矩阵与一个下三角矩阵的乘积

$$A = LR$$

让线性方程组 (4.2)的求解变成两个三角形方程组

$$\begin{cases} L\boldsymbol{y} = \boldsymbol{b} \\ R\boldsymbol{x} = \boldsymbol{y} \end{cases} \tag{4.5}$$

的求解问题，我们称之为线性方程组 (4.2) 的直接三角分解解法。 直接三角分解解法是 Gauss 消去法的“高级”代数描述。把一个用标量语言描述的矩阵算法用矩阵分解的语言来描述，有助于将算法推广到一般情形而且便于展示算法间的关系。

由于矩阵 A 的性质不同和 L 与 R 的取法不同，有各种具体的解法。

解线性方程组的 Doolittle 方法

首先利用 3.1.1 节的方法得到 A 的Doolittle分解

$$A = LR$$

其中$L = (l_{ij})$ 是单位下三角矩阵，$R = (r_{ij})$ 是上三角矩阵，再求解(4.5)。第一步解 $L\boldsymbol{y} = \boldsymbol{b}$，因为 L 是下三角矩阵，采用逐次向前代入的方法；第二步解 $R\boldsymbol{x} = \boldsymbol{y}$，

因为 R 是上三角矩阵，采用逐次向后回代的方法。设 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^T$, $\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^T$, $\boldsymbol{b}=(b_1,b_2,\cdots,b_n)^T$。计算公式是

$$y_i=b_i-\sum_{k=1}^{i-1}l_{ik}y_k,\quad i=1,2,\cdots,n, \tag{4.6}$$

$$x_i=\left(y_i-\sum_{k=i+1}^{n}u_{ik}x_k\right)\Big/ r_{ii},\quad i=n,n-1,\cdots,1. \tag{4.7}$$

以上方法称为解线性方程组的 Doolittle 方法。Doolittle 方法的计算工作量等同 Gauss 消去法。在上述计算过程中元素 r_{ii} 不能为零，否则计算无法进行。若 r_{ii} 的绝对值很小，也会产生较大的舍入误差，使得计算机运算溢出。为此需要对 A 进行选列主元的 Doolittle 分解，并采用部分主元消去法类似的办法解方程组。但在某些特定情况下不必选主元，从而提高运算效率。为举例说明选主元可安全略去，我们考虑对角占优矩阵。如果

$$|a_{ii}|>\sum_{j=1,j\neq i}^{n}|a_{ij}|,\quad i=1,2,\cdots,n,$$

则我们称 $A\in R^{n\times n}$ 是严格对角占优的。以下定理表明此性质可保证不用选主元的 LU 分解。

定理 4.3 *如果 A^T 是严格对角占优的，则 A 有 LU 分解且 $|l_{ij}|\leq 1$. 换句话说，如果用选列主元的 Doolittle 分解方法，定理 3.2 中置换矩阵 $P=I$, $A=LU$.*

例 4.3 用Doolittle方法求解

$$\begin{pmatrix}6&2&1&-1\\2&4&1&0\\1&1&4&-1\\-1&0&-1&3\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}=\begin{pmatrix}6\\-1\\5\\-5\end{pmatrix}.$$

解： 第1步，用 3.1.1 节的方法得到系数矩阵的Doolittle分解，

$$L=\begin{pmatrix}1&&&\\\frac{1}{3}&1&&\\\frac{1}{6}&\frac{1}{5}&1&\\-\frac{1}{6}&\frac{1}{10}&-\frac{9}{37}&1\end{pmatrix},\quad U=\begin{pmatrix}6&2&1&-1\\&\frac{10}{3}&\frac{2}{3}&\frac{1}{3}\\&&\frac{37}{10}&-\frac{9}{10}\\&&&\frac{191}{74}\end{pmatrix}.$$

第2步，由（4.6）得

$$\boldsymbol{y}=(6,3,23/5,-191/74)^T.$$

第3步，由（4.7）得

$$\boldsymbol{x}=(1,-1,1,-1)^T.$$

对称正定方程组的 Cholesky 法

设方程组（4.2）的系数矩阵 $A \in C^{n\times n}$ 是 Hermite 正定矩阵，由 3.1.3 节：存在唯一的对角元素为正的下三角矩阵 L，使 A 有 Cholesky 分解

$$A = LL^H. \tag{4.8}$$

利用 A 的 Cholesky 分解来求解线性方程组（4.2）的方法称为 Cholesky 方法或平方根法。这样求解线性方程组 $A\boldsymbol{x} = \boldsymbol{b}$ 转化为求解下三角方程组 $L\boldsymbol{y} = \boldsymbol{b}$ 和上三角方程组 $L^H\boldsymbol{x} = \boldsymbol{y}$.

Cholesky 方法的原理基于矩阵的三角（Cholesky）分解，所以它也是 Gauss 消去法的变种。但它运用了矩阵 Hermite 正定的性质，运算次数比标准的 Gauss 消去法小得多。从 $A = LL^H$ 元素的计算公式 (3.5) 可得

$$a_{ii} = \sum_{k=1}^{i} l_{ik}^2.$$

由此可推出

$$|l_{ik}| \le \sqrt{a_{ii}},\ k = 1, 2, \cdots, i.$$

这表明 L 的元素有界，即 l_{ik} 完全可以控制，故舍入误差的增长也是可以控制的，因而计算过程稳定。

可以证明，若 A 是 Hermite 正定的，用顺序 Gauss 消去法求解 (4.2)，则有

$$\max_{1\le i,j\le n} |a_{ij}^{(k)}| \le \max_{1\le i,j\le n} |a_{ij}|, \quad k = 1, 2, \cdots, n,$$

其中 $a_{ij}^{(k)}$ 为 $A^{(k)}$ 中的元素，即顺序 Gauss 消去法的中间量也得以控制，不必加入选主元的步骤。

Cholesky 分解 $A = LL^H$ 也可以修改为 $A = LDL^H$，其中 L 为单位下三角矩阵，D 为对角线元素大于零的对角矩阵。事实上令

$$A = \begin{pmatrix} 1 & & & 0 \\ l_{21} & 1 & & \\ \vdots & & \ddots & \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} d_{11} & & & 0 \\ & d_{22} & & \\ & & \ddots & \\ 0 & & & d_{nn} \end{pmatrix} \begin{pmatrix} 1 & \bar{l}_{21} & \cdots & \bar{l}_{n1} \\ & 1 & \cdots & \bar{l}_{n2} \\ & & \ddots & \vdots \\ 0 & & & 1 \end{pmatrix}$$

$$= \begin{pmatrix} d_{11} & & & 0 \\ s_{21} & d_{22} & & \\ \vdots & & \ddots & \\ s_{n1} & s_{n2} & \cdots & d_{nn} \end{pmatrix} \begin{pmatrix} 1 & \bar{l}_{21} & \cdots & \bar{l}_{n1} \\ & 1 & \cdots & \bar{l}_{n2} \\ & & \ddots & \vdots \\ 0 & & & 1 \end{pmatrix},$$

其中 $s_{ij}=l_{ij}d_{jj}$. 比较元素得

$$\begin{cases} d_{11}=a_{11}; \\ s_{ij}=a_{ij}-\sum\limits_{k=1}^{j-1}s_{ik}\bar{l}_{jk}; \\ l_{ij}=s_{ij}/d_{jj}, \qquad j=1,2,\cdots,i-1; \\ d_{ii}=a_{ii}-\sum\limits_{k=1}^{i-1}s_{ik}\bar{l}_{ik}. \end{cases} \tag{4.9}$$

解方程组按如下步骤进行：

$$A\boldsymbol{x}=\boldsymbol{b}\to LDL^H\boldsymbol{x}=\boldsymbol{b}\to\begin{cases} L\boldsymbol{y}=\boldsymbol{b}; \\ L^H\boldsymbol{x}=D^{-1}\boldsymbol{y}. \end{cases}$$

所以

$$\begin{cases} y_i=b_i-\sum\limits_{k=1}^{i-1}l_{ik}y_k \qquad i=1,2,\cdots,n; \\ x_i=y_i/d_{ii}-\sum\limits_{k=i+1}^{n}\bar{l}_{ki}x_k, \qquad i=n,n-1,\cdots,1. \end{cases} \tag{4.10}$$

这一方法称为改进平方根法。

例 4.4 *用改进平方根法求解方程组* $A\boldsymbol{x}=\boldsymbol{b}$, *其中*

$$A=\begin{pmatrix} 1 & 2 & 1 & -3 \\ 2 & 5 & 0 & -5 \\ 1 & 0 & 14 & 1 \\ -3 & -5 & 1 & 15 \end{pmatrix}, \quad \boldsymbol{b}=\begin{pmatrix} 1 \\ 2 \\ 16 \\ 8 \end{pmatrix}.$$

解： 按 (4.9) 计算分解式

$$\begin{aligned} &i=1: \ d_{11}=1; \\ &i=2: \ s_{21}=2, \quad l_{21}=2, \quad d_{22}=1; \\ &i=3: \ s_{31}=1, \quad l_{31}=1, \\ &\qquad s_{32}=-2, \ l_{32}=-2, \ d_{33}=9; \end{aligned}$$

$$\begin{aligned} &i=4: \ s_{41}=-3, \ s_{41}=-3, \\ &\qquad s_{42}=1, \quad l_{42}=1, \\ &\qquad s_{43}=6, \quad l_{43}=\tfrac{2}{3}, \quad d_{44}=1. \end{aligned}$$

依照 (4.10) 解方程组得

$y_1=1,\ y_2=0,\ y_3=15,\ y_4=1; \quad x_4=1,\ x_3=1,\ x_2=1,\ x_1=1.$

三对角方程组的追赶法

很多数学物理问题的数值求解都归结为解具有带状系数矩阵的线性方程组。如差分法解偏微分方程，有限元法解数学物理问题。 数值逼近中用样条插值则要求解三对角方程组。

定义 4.1 设 $A\in C^{n\times n}$ 是 n 阶方阵，如果对 $|i-j|>m$ 时，有 $a_{ij}=0$，我们就称 A 是带宽为 $2m+1$ 的带状矩阵。带宽为 3 的带状矩阵称为三对角矩阵，形如

$$A=\begin{pmatrix} a_1 & c_1 & & & 0\\ u_2 & a_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & u_{n-1} & a_{n-1} & c_{n-1}\\ 0 & & & u_n & a_n \end{pmatrix}. \tag{4.11}$$

可以证明：若 n 阶方阵 A 的带宽为 $2m+1$，假设 A 可以分解为下三角阵 L 和上三角阵 R 的乘积，即 $A=LR$，那么 L 和 R 同样具有带宽 $2m+1$.

在解线性方程组时，若系数矩阵是带状矩阵，充分利用系数矩阵的特殊结构，既可以节省计算量又可以节省存储单元。

以下我们具体讨论三对角方程组的解法。

设 A 满足 (4.11)，如果 A 的顺序主子式皆非零，则 A 有如下三角分解

$$A=PQ=\begin{pmatrix} 1 & & & 0\\ p_2 & 1 & & \\ & \ddots & \ddots & \\ 0 & & p_n & 1 \end{pmatrix}\begin{pmatrix} q_1 & r_1 & & 0\\ & q_2 & \ddots & \\ & & \ddots & r_{n-1}\\ 0 & & & q_n \end{pmatrix}. \tag{4.12}$$

当 $i=1,2,\cdots,n-1$ 时，由于

$$a_{i,i+1}=c_i=(0,0,\cdots,p_i,1,0,\cdots,0)(0,0,\cdots,r_i,q_{i+1},0,\cdots,0)^T,$$

所以

$$c_i=r_i,\quad i=1,2,\cdots,n-1.$$

再由

$$\begin{aligned} a_{i,i-1}&=u_i\\ &=(0,0,\cdots,p_i,1,0,\cdots,0)(0,0,\cdots,r_{i-2},q_{i-1},0,\cdots,0)^T\\ &=p_iq_{i-1}, \end{aligned}$$

及

$$\begin{aligned} a_{i,i} &= a_i \\ &= (0,0,\cdots,p_i,1,0,\cdots,0)(0,0,\cdots,r_{i-1},q_i,0,\cdots,0)^T \\ &= p_i r_{i-1} + q_i, \end{aligned}$$

得到

$$\begin{cases} q_i = a_i - p_i c_{i-1}, \quad i = 1,2,\cdots,n; \\ p_i = u_i/q_{i-1}, \quad i = 2,3,\cdots,n; \\ p_1 = 0. \end{cases} \tag{4.13}$$

从 $P\boldsymbol{y} = \boldsymbol{b}$ 及 $Q\boldsymbol{x} = \boldsymbol{y}$ 得

$$\begin{cases} y_i = b_i - p_i y_{i-1}, \quad i = 1,2,\cdots,n; \\ x_i = (y_i - c_i x_{i+1})/q_i, \quad i = n,n-1,\cdots,1; x_{n+1} = 0. \end{cases} \tag{4.14}$$

按 (4.13) 及 (4.14) 求解 $\boldsymbol{x}$ 的方法称为追赶法，第一个过程称为“追”过程，第二个过程称为“赶”过程，“追赶法”由此而得名。

求解三对角型方程组的追赶法可以描述为：

For $i = 1,2,\cdots,n-1, \quad c_i := r_i$.

$p_1 = 0$.

For $i = 1,2,\cdots,n, \quad q_i := a_i - p_i r_{i-1}$; $p_i := u_i/q_{i-1}$.

$c_n := 0$.

For $i = 1,2,\cdots,n, \quad y_i = b_i - p_i y_{i-1}$.

For $i = n,n-1,\cdots,1, \quad x_i = (y_i - r_i x_{i+1})/q_i$.

定理 4.4 *设三对角方程组的系数 (4.11) 满足*

$$\begin{aligned} &|a_1| > |c_1| > 0, \quad |a_n| > |c_n| > 0; \\ &|a_i| > |c_i| + |u_i| > 0,\ u_i c_i \neq 0,\ i = 2,3,\cdots,n-1. \end{aligned}$$

则 A 非奇异，且有 (4.12) 成立，其中

$$\begin{aligned} &q_i \neq 0,\ i = 2,3,\cdots,n; \\ &0 < \tfrac{|c_i|}{|q_i|} < 1,\ i = 2,3,\cdots,n-1; \\ &|a_i| - |u_i| < |q_i| < |a_i| + |u_i|,\ i = 2,3,\cdots,n-1. \end{aligned}$$

定理保证了追赶法能进行计算。

例 4.5 用追赶法解三对角方程组

$$\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ & & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

解： 设系数矩阵有分解

$$\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ & & -1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & & & \\ p_2 & 1 & & \\ & p_3 & 1 & \\ & & p_4 & 1 \end{pmatrix} \begin{pmatrix} q_1 & r_1 & & \\ & q_2 & r_2 & \\ & & q_3 & r_3 \\ & & & q_4 \end{pmatrix}$$

由公式 (4.13) 得

$$r_1 = c_1 = -1,\ r_2 = c_2 = -1,\ r_3 = c_3 = -1,$$

$$q_1 = a_1 = 2,\ p_2 = \frac{u_2}{q_1} = -\frac{1}{2},\ q_2 = a_2 - p_2 c_1 = \frac{3}{2},$$

$$p_3 = \frac{u_3}{q_2} = -\frac{2}{3},\ q_3 = a_3 - p_3 c_2 = \frac{4}{3},$$

$$p_4 = \frac{u_4}{q_3} = -\frac{3}{4},\ q_4 = a_4 - p_4 c_3 = \frac{5}{4}.$$

依照 (4.14) 得方程组的解

$$y_1 = 1,\ y_2 = \frac{1}{2},\ y_3 = \frac{1}{3},\ y_4 = \frac{5}{4};\quad x_4 = 1,\ x_3 = 1,\ x_2 = 1,\ x_1 = 1.$$

4.2 广义逆矩阵求解矛盾方程组

本节讨论系数矩阵是长方阵的线性方程组的数值解法。设

$$A\boldsymbol{x} = \boldsymbol{b} \tag{4.15}$$

其中 $A \in \boldsymbol{C}^{m\times n}$, $\boldsymbol{b} \in \boldsymbol{C}^m$。 (4.15) 的解存在的充要条件是 $\boldsymbol{b}$ 在 A 的值域中。在许多实际问题中此条件不能满足， 所以对任意 $\boldsymbol{x} \in \boldsymbol{C}^n$，都有 $A\boldsymbol{x} - \boldsymbol{b} \neq 0$，此时称 (4.15) 为矛盾方程组。对于矛盾方程组，我们希望找出这样的向量 $\boldsymbol{x}_0 \in \boldsymbol{C}^n$，它使 $\|A\boldsymbol{x} - \boldsymbol{b}\|_2$ 达到最小，即

$$\|A\boldsymbol{x}_0 - \boldsymbol{b}\|_2 = \min_{\boldsymbol{x}\in\boldsymbol{C}^n} \|A\boldsymbol{x} - \boldsymbol{b}\|_2 \tag{4.16}$$

称此问题为线性方程组 (4.15) 的最小二乘问题（或 LS 问题），称 $\boldsymbol{x}_0$ 为矛盾方程组 (4.15) 的最小二乘解。

最小二乘问题出现在许多场合中，特别是在最小二乘曲线拟合和多元线性回归分析中常常要计算矛盾方程组的最小二乘解。本章用广义逆矩阵的理论讨论 LS 问题的性质及其解法。

4.2.1 基本理论结果

本节将用广义逆矩阵的工具讨论最小二乘问题。

定理 4.5 设 $A \in \boldsymbol{C}^{m\times n}$, $\boldsymbol{b} \in \boldsymbol{C}^m$，则 $\boldsymbol{z}$ 是矛盾方程组 $A\boldsymbol{x} = \boldsymbol{b}$ 的最小二乘解的充要条件为 $\boldsymbol{z}$ 是方程组 $A\boldsymbol{x} = AA^+\boldsymbol{b}$ 的解。

证明：对任意 $\boldsymbol{x} \in \boldsymbol{C}^n$，可证

$$\begin{aligned}\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 &= \|A\boldsymbol{x} - AA^+\boldsymbol{b} + AA^+\boldsymbol{b} - \boldsymbol{b}\|_2^2 \\ &= \|A\boldsymbol{x} - AA^+\boldsymbol{b}\|_2^2 + \|AA^+\boldsymbol{b} - \boldsymbol{b}\|_2^2,\end{aligned}$$

于是

$$\|A\boldsymbol{x} - \boldsymbol{b}\|_2 \geq \|AA^+\boldsymbol{b} - \boldsymbol{b}\|_2.$$

当 $A\boldsymbol{x} = AA^+\boldsymbol{b}$ 时，$\|A\boldsymbol{x} - \boldsymbol{b}\|_2$ 达到最小值 $\|AA^+\boldsymbol{b} - \boldsymbol{b}\|_2$。

反之，若 $\boldsymbol{z}_0$ 是 $A\boldsymbol{x} = \boldsymbol{b}$ 的任意最小二乘解，则

$$\|A\boldsymbol{z}_0 - \boldsymbol{b}\|_2 = \|AA^+\boldsymbol{b} - \boldsymbol{b}\|_2.$$

与前面类似，有

$$\|A\boldsymbol{z}_0 - \boldsymbol{b}\|_2^2 = \|A\boldsymbol{z}_0 - AA^+\boldsymbol{b}\|_2^2 + \|AA^+\boldsymbol{b} - \boldsymbol{b}\|_2^2.$$

从而 $\|A\boldsymbol{z}_0 - AA^+\boldsymbol{b}\|_2 = 0$，即 $A\boldsymbol{z}_0 = AA^+\boldsymbol{b}$。所以 $\boldsymbol{z}_0$ 是 $A\boldsymbol{x} = AA^+\boldsymbol{b}$ 的解。证毕。

定理 4.6 设 $A \in \boldsymbol{C}^{m\times n}$, $\boldsymbol{b} \in \boldsymbol{C}^m$，则 $\boldsymbol{z}$ 是矛盾方程组 (4.15) 的最小二乘解的充要条件为 $\boldsymbol{z}$ 是方程组

$$A^HA\boldsymbol{x} = A^H\boldsymbol{b} \tag{4.17}$$

的解。

证明：若 $\boldsymbol{z}$ 是 $A\boldsymbol{x} = \boldsymbol{b}$ 的最小二乘解，由 定理 4.5 知，$\boldsymbol{z}$ 是 $A\boldsymbol{x} = AA^+\boldsymbol{b}$ 的解，于是

$$A^HA\boldsymbol{z} = A^H(AA^+\boldsymbol{b}) = A^H(AA^+)^H\boldsymbol{b} = (AA^+A)^H\boldsymbol{b} = A^H\boldsymbol{b},$$

即 z 是 $A^HAx = A^Hb$ 的解。

反之，若 z 是 $A^HAx = A^Hb$ 的解，则有

$$\begin{aligned} Az &= AA^+Az = (AA^+)^HAz = (A^+)^H(A^HAz) \\ &= (A^+)^HA^Hb = (AA^+)^Hb = AA^+b, \end{aligned}$$

可见 z 是 $Ax = AA^+b$ 的解，从而是 $Ax = b$ 的最小二乘解。 证毕。

推论 4.1 若 $A = FG$ 分解中 G 为行满秩，则 (4.17) 等价于方程组

$$F^HAx = F^Hb.$$

证明：由于 $A^H = G^HF^H$，而 G^H 列满秩，则

$$A^HAx = A^Hb \Leftrightarrow G^HF^H(Ax - b) = 0 \Leftrightarrow F^H(Ax - b) = 0.$$

证毕。

推论 4.2 向量 x 是 (4.17) 的解当且仅当有向量 $r \in C^m$ 使

$$\begin{pmatrix} -I & A \\ A^H & 0 \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}.$$

定理 4.7 设 $A \in C^{m\times n}$, $b \in C^m$，矛盾方程组 $Ax = b$ 的全部最小二乘解为

$$z = A^+b + (I - A^+A)y, \quad y \in C^n \text{ 任意.} \tag{4.18}$$

证明：因为 $(AA^+)(AA^+b) = AA^+b$，由定理 3.20 知，$Ax = AA^+b$ 有解，且其通解为

$$z = A^+b + (I - A^+A)y, \quad y \in C^n \text{ 任意.}$$

又由定理 4.5 知，此即为矛盾方程组 $Ax = b$ 的全部最小二乘解。 证毕。

由式 (4.18) 知，矛盾方程组 $Ax = b$ 有唯一最小二乘解的充分必要条件是 $A^+A = I$，即 A 列满秩。最小二乘解一般不唯一，在所有最小二乘解中范数最小的二乘解称为矛盾方程组 $Ax = b$ 的极小范数最小二乘解。

定理 4.8 设 $A \in C^{m\times n}$, $b \in C^m$，矛盾方程组 $Ax = b$ 的唯一极小范数最小二乘解为 $x_0 = A^+b$。

证明：由 定理 4.5 知矛盾线性方程组 (4.15) 的唯一极小范数最小二乘解为 $Ax = AA^+b$ 的唯一极小范数解。根据定理 3.20 知此解为

$$x_0 = A^+(AA^+b) = A^+b.$$

证毕。

由以上结论，如果我们能计算出 A^+ 即可求解 LS 问题。如果 A 是小型矩阵，可用满秩分解的方法计算 A^+。

例 4.6 *用满秩分解方法求矛盾方程组*

$$\begin{cases} 2x_1+4x_2+x_3+x_4=10 \\ x_1+2x_2-x_3+2x_4=6 \\ -x_1-2x_2-2x_3+x_4=-7 \end{cases}$$

的全部最小二乘解和极小范数最小二乘解。

解：将线性方程组写成矩阵形式 $A\boldsymbol{x}=\boldsymbol{b}$，其中

$$A=\begin{pmatrix} 2 & 4 & 1 & 1 \\ 1 & 2 & -1 & 2 \\ -1 & -2 & -2 & 1 \end{pmatrix},\quad \boldsymbol{b}=\begin{pmatrix} 10 \\ 6 \\ -7 \end{pmatrix}.$$

计算 A 的满秩分解得

$$A=FG=\begin{pmatrix} 2 & 1 \\ 1 & -1 \\ -1 & -2 \end{pmatrix}\begin{pmatrix} 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix},$$

于是

$$\begin{aligned} A^+ &= G^T(GG^T)^{-1}(F^TF)^{-1}F^T \\ &= \frac{1}{33}\begin{pmatrix} 2 & 1 & -1 \\ 4 & 2 & -2 \\ 1 & -5 & -6 \\ 1 & 6 & 5 \end{pmatrix}. \end{aligned}$$

全部最小二乘解为

$$\begin{aligned} \boldsymbol{x} &= A^+\boldsymbol{b}+(I-A^+A)\boldsymbol{y} \\ &= \begin{pmatrix} 1 \\ 2 \\ \frac{2}{3} \\ \frac{1}{3} \end{pmatrix}+\frac{1}{11}\begin{pmatrix} 9 & -4 & -1 & -1 \\ -4 & 3 & -2 & -2 \\ -1 & -2 & 5 & 5 \\ -1 & -2 & 5 & 5 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} \end{aligned}$$

其中 $y_1, y_2, y_3, y_4 \in \boldsymbol{C}$ 任意。极小范数最小二乘解为

$$\boldsymbol{x}=A^+\boldsymbol{b}=\left(1,2,\frac{1}{3},\frac{2}{3}\right)^T.$$

4.2.2 列满秩 LS 问题

设 $A \in \boldsymbol{C}^{m\times n}$, $\boldsymbol{b} \in \boldsymbol{C}^m$，且 A 是列满秩矩阵，称对应的 LS 问题为列满秩 LS 问题。列满秩 LS 问题一般采用法方程组和 QR 分解的方法计算。

法方程组的方法

求解列满秩 LS 问题应用最广的方法是求解法方程组的方法。设 $A \in \boldsymbol{C}^{m\times n}$, $\boldsymbol{b} \in \boldsymbol{C}^m$，且 A 是列满秩矩阵。根据定理 4.6，矛盾方程组 $A\boldsymbol{x} = \boldsymbol{b}$ 的最小二乘解 $\boldsymbol{x}_{LS}$ 满足

$$A^H A\boldsymbol{x}_{LS} = A^H\boldsymbol{b}. \tag{4.19}$$

因为 A 是列满秩矩阵，所以 A^HA 为 Hermite 正定矩阵，可进行 Cholesky 分解。方程组 (4.19) 有唯一的解，算法如下：

1. 计算 $G = A^HA$;
2. 计算 $\boldsymbol{d} = A^H\boldsymbol{b}$;
3. 计算 Cholesky 分解 $G = LL^H$;
4. 解 $L\boldsymbol{y} = \boldsymbol{d}$ 和 $L^H\boldsymbol{x}_{LS} = \boldsymbol{y}$.

用 QR 分解求解列满秩的 LS 问题

本节再介绍用 QR 分解求解列满秩 LS 问题的一种方法。设 $A \in \boldsymbol{C}^{m\times n}$, $\boldsymbol{b} \in \boldsymbol{C}^m$，且 A 是列满秩矩阵。根据定理 3.5，可将 A 分解为列正交矩阵 Q 与上三角矩阵 R 的乘积。将 $A = QR$ 代入方程 $A^HA\boldsymbol{x} = A^H\boldsymbol{b}$，并注意到 $Q^HQ = I$，有

$$R^HQ^HQR\boldsymbol{x} = R^HQ^H\boldsymbol{b}.$$

由于 R 非奇异，用 $(R^H)^{-1}$ 左乘上式两端得

$$R\boldsymbol{x} = Q^H\boldsymbol{b}. \tag{4.20}$$

所以，求得 R 和 $Q^H\boldsymbol{b}$ 后，仍需要解三角形方程组 (4.20) 方可求得最小二乘解 $\boldsymbol{x}_{LS}$。

从理论上讲计算出的 Q 是列正交的，但由于舍入误差的影响，得到的 Q 并不一定列正交，有时正交性很差。由于数值计算稳定性的要求，目前采用精度高得多的所谓“修改的 Gram-Schmidt 正交化过程”。具体计算步骤如下：

(1) 将 $A^{(1)}=A=(a_1^{(1)},a_2^{(1)},\cdots,a_n^{(1)})$ 的第一列 $a_1^{(1)}$ 规一化为 Q 的第一列 q_1:

$$\begin{cases} q_1=a_1^{(1)}/r_{11}; \\ r_{11}=\sqrt{(a_1^{(1)},a_1^{(1)})}. \end{cases}$$

然后将 $A^{(1)}$ 的第 2 列至第 n 列分别与 q_1 正交化，得出向量组 $a_2^{(2)}$, $a_3^{(2)}$, $\cdots$, $a_n^{(2)}$:

$$\begin{cases} a_j^{(2)}=a_j^{(1)}-r_{1j}\cdot q_1; \\ r_{1j}=(q_1,a_j^{(1)}), \quad j=2,3,\cdots,n; \end{cases}$$

(2) 对于 $k=2,3,\cdots,n$，执行下列运算:

$$\begin{cases} q_k=a_k^{(k)}/r_{kk}; \\ r_{kk}=\sqrt{(a_k^{(k)},a_k^{(k)})}; \\ a_j^{(k+1)}=a_j^{(k)}-r_{kj}\cdot q_k; \\ r_{kj}=(q_k,a_j^{(k)}), \quad j=k+1,k+2,\cdots,n. \end{cases}$$

便得到正交规一化的向量组 $\{q_k\}_1^n$;

(3) 利用 $Q=(q_1,q_2,\cdots,q_n)$，$R=(r_{ij})$，计算 $Q^H\boldsymbol{b}$，解三角形方程组

$$R\boldsymbol{x}=Q^H\boldsymbol{b},$$

即得解 $\boldsymbol{x}_{LS}=R^{-1}Q^H\boldsymbol{b}$。

注 4.1 设 $A\in C^{m\times n}$ 列满秩，令条件数 $\mathrm{cond}_2(A)=\|A\|_2\|A^+\|_2$，最小二乘解的余量 $\rho_{LS}=\|A\boldsymbol{x}_{LS}-\boldsymbol{b}\|_2$。易证 $\mathrm{cond}_2(A)=\sqrt{\|A^HA\|_2\|(A^HA)^{-1}\|_2}$。下面给出将 LS 问题的法方程组法和 QR 方法比较的一些结论以供参考。

(1) 法方程组法得到的解 $\boldsymbol{x}_{LS}$ 的相对误差依赖于条件数的平方。

(2) QR 方法解一个近似的 LS 问题的解的相对误差约为 $\mu(\mathrm{cond}_2(A)+\rho_{LS}(\mathrm{cond}_2(A))^2)$。

(3) 法方程组法在 $m\gg n$ 时只需一半的运算，且不需要太大的存储空间。

因此，如果 ρ_{LS} 很小且 $\mathrm{cond}_2(A)$ 很大，则法方程组法通常给出比稳定的 QR 方法精度更低的解。相反，当用来解大余量和坏条件数问题时，两种方法得到差不多的不精确的解。总之，很难选择一种完全令人满意的解法。

4.2.3 秩亏损的 LS 问题

设 $A \in C^{m\times n}$，且 $\mathrm{rank}(A) = r < n$, 则称 A 是秩亏损的矩阵。由定理 4.7 知，秩亏损 LS 问题有无穷多解。对于大规模的秩亏损的 LS 问题，我们可以采用对 A 完全正交分解的方法来求解 LS 问题。具体地说，我们求酉矩阵 Q 和 Z 使得

$$Q^H A Z = T = \begin{pmatrix} T_{11} & 0 \\ 0 & 0 \end{pmatrix},$$

其中 $T_{11} \in C_r^{r\times r}$, 则

$$\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 = \|(Q^H A Z)Z^H\boldsymbol{x} - Q^H\boldsymbol{b}\|_2^2 = \|T_{11}\boldsymbol{w} - \boldsymbol{c}\|_2^2 + \|\boldsymbol{d}\|_2^2,$$

其中

$$Z^H\boldsymbol{x} = \begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{y} \end{pmatrix}, \quad Q^H\boldsymbol{b} = \begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{d} \end{pmatrix}$$

其中 $\boldsymbol{w}, \boldsymbol{c} \in C^r$，$\boldsymbol{y} \in C^{n-r}$，$\boldsymbol{d} \in C^{m-r}$。又注意到

$$\|\boldsymbol{x}\|_2^2 = \|Z^H\boldsymbol{x}\|_2^2 = \|\boldsymbol{w}\|_2^2 + \|\boldsymbol{y}\|_2^2,$$

则当 $T_{11}\boldsymbol{w} - \boldsymbol{c} = 0$ 及 $\boldsymbol{y} = 0$ 时，得到 LS 问题的极小范数最小二乘解

$$\boldsymbol{x}_{LS} = Z\begin{pmatrix} T_{11}^{-1}\boldsymbol{c} \\ 0 \end{pmatrix}.$$

我们可以对 A 进行一系列行和列 Householder 变换及列置换得到 A 的完全正交分解。

定理 4.9 *给定 $A \in C_r^{m\times n}$，则存在 r 个*Householder *矩阵之积 U, V 以及互换矩阵之积 P，使有*

$$UAPV = UAK = \begin{pmatrix} G & 0 \\ 0 & 0 \end{pmatrix}$$

其中 $G \in C_r^{r\times r}$ 为上三角矩阵，$V, P, K \in C^{n\times n}$，$U \in C^{m\times m}$ 均为酉矩阵。

A 的奇异值分解是一种特殊的完全正交分解。若已知 A 的奇异值分解，则可以得到 $\boldsymbol{x}_{LS}$ 的一个简洁表达式。

定理 4.10 *给定 $A \in C_r^{m\times n}$, $\boldsymbol{b} \in C^m$，如果 A 的奇异值分解为*

$$U^H A V = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}$$

其中 $\sum = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_r)$，而 σ_i $(i = 1, 2, \cdots, r)$ 为 A 的非零奇异值，酉矩阵 $U = (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_m)$ 和 $V = (\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n)$ 是按列划分的，则 LS 问题的极小范数最小二乘解

$$\boldsymbol{x}_{LS} = \sum_{i=1}^{r} \frac{\boldsymbol{u}_i^H \boldsymbol{b}}{\sigma_i} \boldsymbol{v}_i.$$

而且误差余量

$$\rho_{LS}^2 = \|A\boldsymbol{x}_{LS} - \boldsymbol{b}\|_2^2 = \sum_{i=r+1}^{m} (\boldsymbol{u}_i^H \boldsymbol{b})^2.$$

证明：因为 A 的奇异值分解为

$$U^H A V = \begin{bmatrix} \sum & 0 \\ 0 & 0 \end{bmatrix},$$

所以

$$A = U \begin{bmatrix} \sum & 0 \\ 0 & 0 \end{bmatrix} V^H,$$

$$A^+ = V \begin{bmatrix} \sum^- & 0 \\ 0 & 0 \end{bmatrix} U^H.$$

由定理4.8，LS 问题的极小范数最小二乘解为

$$\boldsymbol{x}_{LS} = A^+ \boldsymbol{b} = V \begin{bmatrix} \sum^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^H \boldsymbol{b} = \sum_{i=1}^{r} \frac{\boldsymbol{u}_i^H \boldsymbol{b}}{\sigma_i} \boldsymbol{v}_i.$$

此时误差余量

$$\begin{aligned} \rho_{LS}^2 &= \|A\boldsymbol{x}_{LS} - \boldsymbol{b}\|_2^2 = \|AA^+\boldsymbol{b} - \boldsymbol{b}\|_2^2 \\ &= \left\| U \begin{bmatrix} \sum & 0 \\ 0 & 0 \end{bmatrix} V^H V \begin{bmatrix} \sum^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^H \boldsymbol{b} - \boldsymbol{b} \right\|_2^2 \\ &= \left\| \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} U^H \boldsymbol{b} - U^H \boldsymbol{b} \right\|_2^2 \\ &= \sum_{i=r+1}^{m} (\boldsymbol{u}_i^H \boldsymbol{b})^2. \end{aligned}$$

证毕。

4.3 线性方程组的迭代解法

本节讨论线性方程组的另一类求解方法—迭代解法。迭代法与直接法不同，迭代法一般不能通过有限次的算术运算求得方程的精确解，而是从初始向量出发，用

设计好的步骤逐次算出近似解。 大型线性方程组的求解是大规模科学与工程计算的核心。随着计算机的飞速发展，需求解的问题规模越来越大， 迭代法已取代直接法成为求解大型线性方程组的最重要一类方法。

4.3.1 迭代法的一般概念

设$A \in \boldsymbol{R}_n^{n\times n}$, $\boldsymbol{b} \in \boldsymbol{R}^n$，考虑线性方程组

$$A\boldsymbol{x} = \boldsymbol{b}. \tag{4.21}$$

一般地，先将 (4.21) 变为同解方程组

$$\boldsymbol{x} = B\boldsymbol{x} + f, \tag{4.22}$$

形成迭代格式

$$\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + f, \quad k = 0, 1, \cdots \tag{4.23}$$

其中 B 称为迭代矩阵，$\boldsymbol{x}^{(0)}$ 是任给的初始向量。由于 $\boldsymbol{x}^{(k+1)}$ 是 $\boldsymbol{x}^{(k)}$ 的线性函数，(4.23) 称为线性迭代。

定义 4.2 *如果迭代公式 (4.23) 产生的序列 $\{\boldsymbol{x}^{(k)}\}$ 满足*

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}^*, \quad \forall \boldsymbol{x}^{(0)} \in \boldsymbol{R}^n,$$

则称迭代法 (4.23) 是收敛的。

实际使用的迭代法应该是收敛的迭代法，这里我们要给出判别迭代法收敛的条件。设 $\boldsymbol{x}^*$ 是 (4.22) 的解，即

$$\boldsymbol{x}^* = B\boldsymbol{x}^* + f.$$

以 (4.23) 式减去上式，并记误差向量为

$$\boldsymbol{e}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^*,$$

则有

$$\boldsymbol{e}^{(k+1)} = B\boldsymbol{e}^{(k)}, k = 0, 1, \cdots,$$

由此可推得

$$\boldsymbol{e}^{(k)} = B^k\boldsymbol{e}^{(0)},$$

其中 $\boldsymbol{e}^{(0)} = \boldsymbol{x}^{(0)} - \boldsymbol{x}^*$ 与 k 无关，所以 (4.23) 收敛就意味着

$$\lim_{k\to\infty} \boldsymbol{e}^{(k)} = \lim_{k\to\infty} B^k\boldsymbol{e}^{(0)} = 0, \ \forall \boldsymbol{e}^{(0)} \in \boldsymbol{R}^n.$$

定理 4.11 设 $A^{(k)}$ 是 $\boldsymbol{R}^{n\times n}$ 上的矩阵序列。则 $\lim\limits_{k\to\infty} A^{(k)}=0$ 的充分必要条件是

$$\lim_{k\to\infty} A^{(k)}\boldsymbol{x}=0,\ \forall \boldsymbol{x}\in\boldsymbol{R}^n. \tag{4.24}$$

证明：对于任意相容的向量与矩阵范数有

$$\|A^{(k)}\boldsymbol{x}\|\le\|A^{(k)}\|\|\boldsymbol{x}\|,$$

从而可证必要性。若取 $\boldsymbol{x}$ 为第 j 个单位向量 $\boldsymbol{e}_j$，则 $\lim\limits_{k\to\infty} A^{(k)}\boldsymbol{e}_j=0$，意味着 $A^{(k)}$ 的第 j 列各元素极限为零，取 $j=1,2,\cdots,n$，充分性得证。 证毕。

下面给出迭代法收敛的充分必要条件：

定理 4.12 下面 3 个命题是等价的：

(1) 迭代法 $\boldsymbol{x}^{(k+1)}=B\boldsymbol{x}^{(k)}+f$ 收敛；

(2) $\rho(B)<1$；

(3) 至少存在一个矩阵范数 $\|\cdot\|$，使 $\|B\|<1$.

证明：从以上分析，命题 (1) 中迭代法收敛等价于

$$\lim_{k\to\infty} B^k e^{(0)}=\lim_{k\to\infty} e^{(k)}=0,\ \forall e^{(0)}\in\boldsymbol{R}^n.$$

由定理 4.11，上式成立的充要条件是 $\lim\limits_{k\to\infty} B^k=0$，即 B 是收敛矩阵。由谱半径性质的定理 1.25，定理 1.26及定理 1.27，可得结论。 证毕。

有时实际判别一个迭代法是否收敛，条件 $\rho(B)<1$ 较难验证。但 $\|B\|_1$, $\|B\|_\infty$, $\|B\|_F$ 可以用 B 的元素表示，所以用 $\|B\|_1<1$, $\|B\|_\infty<1$, $\|B\|_F<1$ 作为收敛的充分条件较为方便。

下面讨论收敛速度的问题：

定理 4.13 设 $\boldsymbol{x}^*$ 是方程 $\boldsymbol{x}=B\boldsymbol{x}+f$ 的唯一解，$\|B\|<1$，则由 (4.23) 产生的向量序列 $\boldsymbol{x}^{(k)}$ 满足

$$\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|\le\frac{\|B\|}{1-\|B\|}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|; \tag{4.25}$$

$$\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|\le\frac{\|B\|^k}{1-\|B\|}(\|f\|+2\|\boldsymbol{x}^{(0)}\|)\|. \tag{4.26}$$

证明：不难验证

$$\begin{aligned}\boldsymbol{x}^{(k)}-\boldsymbol{x}^* &= B(\boldsymbol{x}^{(k-1)}-\boldsymbol{x}^*)\\ &= B(\boldsymbol{x}^{(k-1)}-\boldsymbol{x}^{(k)})+B(\boldsymbol{x}^{(k)}-\boldsymbol{x}^*),\end{aligned}$$

所以

$$(I-B)(\boldsymbol{x}^{(k)}-\boldsymbol{x}^{*})=B(\boldsymbol{x}^{(k-1)}-\boldsymbol{x}^{(k)}).$$

因为 $\|B\|<1$，由定理 4.12 知迭代法收敛，$\lim\limits_{k\to\infty}\boldsymbol{x}^{(k)}=\boldsymbol{x}^{*}$. 又由推论 1.34 知 $I-B$ 非奇异，且

$$\begin{aligned}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{*}\|&=\|(I-B)^{-1}B(\boldsymbol{x}^{(k-1)}-\boldsymbol{x}^{(k)}\|\\&\leq\frac{\|B\|}{1-\|B\|}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|,\end{aligned}$$

即得 (4.25)。再反复运用

$$\begin{aligned}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|&=\|B(\boldsymbol{x}^{(k-1)}-\boldsymbol{x}^{(k-2)})\|\\&\leq\|B\|\|\boldsymbol{x}^{(k-1)}-\boldsymbol{x}^{(k-2)}\|,\end{aligned}$$

即得

$$\begin{aligned}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{*}\|&\leq\frac{\|B\|^{k}}{1-\|B\|}\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(0)}\|\\&=\frac{\|B\|^{k}}{1-\|B\|}\|B\boldsymbol{x}^{(0)}+f-\boldsymbol{x}^{(0)}\|\\&\leq\frac{\|B\|^{k}}{1-\|B\|}(\|f\|+2\|\boldsymbol{x}^{(0)}\|).\end{aligned}$$

证毕。

式 (4.26) 可用来估计达到指定精度需要的迭代次数，若要使 $\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{*}\|\leq\varepsilon$，只要

$$\frac{\|B\|^{k}}{1-\|B\|}(\|f\|+2\|\boldsymbol{x}^{(0)}\|)\leq\varepsilon,$$

即

$$k\geq\ln\left(\frac{\varepsilon(1-\|B\|)}{\|f\|+2\|\boldsymbol{x}^{(0)}\|}\right)/\ln\|B\|.$$

顺便指出，当 $\|B\|$ 不是很接近 1 时，可用相邻两次迭代向量的差的范数 $\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|\leq\varepsilon$ 来作为计算终止的标志。

在线性迭代法中，我们用 $R(B)=-\ln\rho(B)$ 表示迭代法的收敛速度。

考察误差向量 $\boldsymbol{e}^{(k)}=\boldsymbol{x}^{(k)}-\boldsymbol{x}^{*}=B^{k}\boldsymbol{e}^{(0)}$. 设 B 有 n 个线性无关的特征向量 $\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_n$，相应的特征值为 $\lambda_1,\lambda_2,\cdots,\lambda_n$. 由

$$\boldsymbol{e}^{(0)}=\sum_{i=1}^{n}a_i\boldsymbol{u}_i,$$

得

$$\boldsymbol{e}^{(k)}=B^{k}\boldsymbol{e}^{(0)}=\sum_{i=1}^{n}a_iB^{k}\boldsymbol{u}_i=\sum_{i=1}^{n}a_i\lambda_i^{k}\boldsymbol{u}_i$$

当 $\rho(B)<1$ 时，如 $\lambda_i^k \to 0\ (i=1,2,\cdots,n;\ k\to\infty)$ 越快，则 $\boldsymbol{e}^{(k)}\to 0$ 越快。由此说明可用 $\rho(B)$ 来刻划迭代法的收敛快慢。

现在来确定迭代次数 k，使

$$[\rho(B)]^k < 10^{-\varepsilon},$$

取对数得

$$k \geq \frac{\varepsilon \ln 10}{-\ln \rho(B)}.$$

由此看出，当 $R(B)=-\ln\rho(B)$ 越大，上式成立所需迭代次数越少，表明迭代次数与收敛速度成反比。

4.3.2 J 迭代法和 G-S 迭代法

J 迭代法和 G-S 迭代法的构造

从方程组 (4.21) 出发可以由不同的路径得到不同的方程组 (4.22)，从而得到不同的迭代法 (4.23)。例如，设 A 可以分裂为

$$A = M - N \tag{4.27}$$

其中 M 非奇异。则由 (4.21) 可得

$$\boldsymbol{x} = M^{-1}N\boldsymbol{x} + M^{-1}\boldsymbol{b}. \tag{4.28}$$

令

$$B = M^{-1}N = I - M^{-1}A,$$
$$f = M^{-1}\boldsymbol{b},$$

就可以得到 (4.22) 的形式。

设 (4.21) 中 $A=(a_{ij})$，可以把 A 分裂为

$$A = D + L + U$$

其中 $D=\operatorname{diag}(a_{11},a_{22},\cdots,a_{nn})$，

$$L=\begin{bmatrix} 0 & & & 0 \\ a_{21} & 0 & & \\ \vdots & & \ddots & \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix},\quad U=\begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ & 0 & \cdots & \vdots \\ & & \ddots & a_{n-1,n} \\ 0 & & & 0 \end{bmatrix}$$

其中 L 和 U 为 A 的严格上、下三角部分（不包括对角线)。现设 D 非奇异，即 $a_{ii} \neq 0,\ i = 1, \cdots, n$. 对应于 (4.27) 形式的一般分裂式，令 $M = D,\ N = -L - U$，则 (4.28) 等价于

$$\boldsymbol{x} = -D^{-1}(L+U)\boldsymbol{x} + D^{-1}\boldsymbol{b}.$$

由此构造迭代法

$$\boldsymbol{x}^{(k+1)} = B_J\boldsymbol{x}^{(k)} + f, \quad k = 0, 1, \cdots, \tag{4.29}$$

其中 向量 f 和迭代矩阵 B_J 为

$$\begin{aligned} f &= D^{-1}\boldsymbol{b}, \\ B_J &= -D^{-1}(L+U) = I - D^{-1}A. \end{aligned} \tag{4.30}$$

(4.29) 称为 Jacobi 迭代法(简称 J 迭代法)。

Jacobi 迭代算法可简单描述为： 任给初始向量 $\boldsymbol{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})^T$， 对 $k = 0, 1, 2, \cdots,\ i = 1, 2, \cdots, n$

$$x_i^{(k+1)} = \left(b_i - \sum_{j=1, j\neq i}^{n} a_{ij}x_j^{(k)}\right) \cdot \frac{1}{a_{ii}}.$$

如果 $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\| \leq \varepsilon$，停止 ，否则转下一个 k。

如果令 $M = D + L,\ N = -U$，对应于 (4.27) 的分裂式有

$$\begin{aligned} f &= (D+L)^{-1}\boldsymbol{b}, \\ B_G &= -(D+L)^{-1}U = I - (D+L)^{-1}A, \end{aligned} \tag{4.31}$$

便得到 Gauss—Seidel 迭代法（简称 G-S 迭代法）

$$\boldsymbol{x}^{(k+1)} = B_G\boldsymbol{x}^{(k)} + f, \quad k = 0, 1, \cdots. \tag{4.32}$$

G-S 迭代算法可简单描述为： 任给初始向量 $\boldsymbol{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})^T$， 对 $k = 0, 1, 2, \cdots,\ i = 1, 2, \cdots, n$

$$x_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right) \cdot \frac{1}{a_{ii}}.$$

如果 $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\| \leq \varepsilon$，停止 ，否则转下一个 k。

例 4.7 用 Jacobi 迭代法求解方程组

$$\begin{cases} x_1 + 2x_2 - 2x_3 &= 1 \\ x_1 + x_2 + x_3 &= 3 \\ 2x_1 + 2x_2 + x_3 &= 5. \end{cases}$$

解： 取初始向量 $\boldsymbol{x}^{(0)}=(0,0,0)^T$， Jacobi 迭代格式为

$$\begin{cases} x_1^{(k)}=1-2x_2^{(k-1)}+2x_3^{(k-1)} \\ x_2^{(k)}=3-x_1^{(k-1)}-x_3^{(k-1)} \\ x_3^{(k)}=5-2x_1^{(k-1)}-2x_2^{(k-1)}. \end{cases}$$

可得， $\boldsymbol{x}^{(1)}=(1,3,5)^T$, $\boldsymbol{x}^{(2)}=(5,-3,-3)^T$, $\boldsymbol{x}^{(3)}=(1,1,1)^T$, $\boldsymbol{x}^{(4)}=(1,1,1)^T$。 所以方程组的解为 $\boldsymbol{x}=(1,1,1)^T$。

例 4.8 用 Gauss-Seidel 迭代法求解方程组

$$\begin{cases} 9x_1-x_2-x_3 &=7 \\ -x_1+8x_2 &=7 \\ -x_1+9x_3 &=8 \end{cases}$$

解： 取初始向量 $\boldsymbol{x}^{(0)}=(0,0,0)^T$， Gauss-Seidel 迭代格式为

$$\begin{cases} x_1^{(k)}=\frac{1}{9}(7+x_2^{(k-1)}+x_3^{(k-1)}) \\ x_2^{(k)}=\frac{1}{8}(7+x_1^{(k)}) \\ x_3^{(k)}=\frac{1}{9}(8+x_1^{(k)}). \end{cases}$$

可得， $\boldsymbol{x}^{(1)}=(0.7778,0.9722,0.9753)^T$, $\boldsymbol{x}^{(2)}=(0.9942,0.9993,0.9994)^T$, $\boldsymbol{x}^{(3)}=(0.9999,0.9999,0.9999)^T$, $\boldsymbol{x}^{(4)}=(1,1,1)^T$, $\boldsymbol{x}^{(5)}=(1,1,1)^T$。 所以方程组的解为 $\boldsymbol{x}=(1,1,1)^T$。

上面我们从 A 的不同分裂式 $A=M-N$ 得到两种不同的迭代法。一般地，迭代式 (4.23) 可以看成解方程组 $(I-B)\boldsymbol{x}=f$ 的一种自然的安排。因为 $B=I-M^{-1}A$, 这个方程组就是

$$M^{-1}A\boldsymbol{x}=M^{-1}\boldsymbol{b}.$$

它和原方程组 $A\boldsymbol{x}=\boldsymbol{b}$ 是同解的方程组。我们称它是原方程组经预处理的的方程组，M 称为预处理矩阵。下面我们还要通过不同的 M 引入不同的迭代方法。而 J 迭代法和 G-S 迭代法的预处理矩阵分别是

$$M_J=D,\ M_G=D+L.$$

J 迭代法和 G-S 迭代法的收敛性

定理 4.12 给出了J 迭代法和 G-S 迭代法收敛的充分必要条件。为了再给出一些较容易验证的充分条件，下面讨论对角占优矩阵的性质。

定义 4.3 若 $A=(a_{ij})\in \boldsymbol{R}^{n\times n}$，满足

$$|a_{ii}|>\sum_{j=1,j\neq i}^{n}|a_{ij}|,\quad i=1,2,\cdots,n,$$

称 A 为严格对角占优矩阵。若 A满足

$$|a_{ii}| \geq \sum_{j=1,j\neq i}^{n} |a_{ij}|, \quad i = 1,2,\cdots,n,$$

且其中最少有一个严格不等式成立，称 A 为弱对角占优矩阵。

定义 4.4 设 $A = (a_{ij}) \in R^{n\times n}$，若不能找到置换矩阵 P，使得

$$P^T AP = \begin{pmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ 0 & \widetilde{A}_{22} \end{pmatrix} \tag{4.33}$$

其中 $\widetilde{A}_{11}$ 与 $\widetilde{A}_{22}$ 均为方阵，则称 A 为不可约矩阵。

显然，若 A 是可约的，则可通过行与列的置换变成 (4.33) 的形式，方程组 $Ax = b$ 就可以化为低阶的方程组。

定理 4.14 若 $A = (a_{ij})$ 严格对角占优，则 $a_{ii} \neq 0$，$i = 1,2,\cdots,n$，且 A 非奇异。

定理 4.15 若 $A = (a_{ij})$ 不可约且弱对角占优，则 $a_{ii} \neq 0$，$i = 1,2,\cdots,n$，且 A 非奇异。

J 迭代法和 G-S 迭代法只适用于具有非零对角元的方程组。以上定理给出了 A 满足此性质的条件。

定理 4.16 若 A 是严格对角占优或不可约弱对角占优矩阵，则解 $Ax = b$ 的 J 迭代法和 G-S 迭代法均收敛。

证明：以下只证明若 A 是不可约弱对角占优矩阵，则 G-S 迭代法收敛。其余情形可类似证明。

记 $A = L + D + U$，其中 D 为 A 的对角元素组成的矩阵，L 为 A 的对角线下的元素组成的下三角阵，U 为 A 的对角线上的元素组成的上三角阵。用 B_G 表示 G-S 迭代法的迭代矩阵，只要证明 $\rho(B_G) < 1$。

用反证法，先假定 $B_G = -(D + L)^{-1}U$ 有特征值 λ 满足 $|\lambda| \geq 1$. A 是不可约弱对角占优阵，由定理 4.15 有 $a_{ii} \neq 0$，所以 $\det(D + L) = \det(D) \neq 0$，则由 $\det[\lambda I + (D + L)^{-1}U] = 0$ 可推出

$$\det(D + L + \frac{1}{\lambda}U) = 0 \tag{4.34}$$

而 $A = L + D + U$ 与 $D + L + \frac{1}{\lambda}U$ 的零元素与非零元素位置完全一样，所以 $D + L + \frac{1}{\lambda}U$ 也是不可约的。又因 $|\lambda| \geq 1$，$D + L + \frac{1}{\lambda}U$ 也是弱对角占优矩阵。由定理 4.15，$\det(D + L + \frac{1}{\lambda}U) \neq 0$，这与 (4.34) 矛盾。所以 $\rho(B_G) < 1$。从而 G-S 迭代法收敛。

证毕。

若 A 为对称正定矩阵，则解 $A\boldsymbol{x}=\boldsymbol{b}$ 的 J 迭代法和 G-S 迭代法有下述收敛性定理。

定理 4.17 *设 A 具有正的对角元且为对称矩阵，则 Jacobi 方法收敛的充分必要条件是 A 和 $2D-A$ 同为正定矩阵。*

证明：根据假定， D 是正定矩阵，因此

$$B=I-D^{-1}A=D^{-\frac{1}{2}}(I-D^{-\frac{1}{2}}AD^{-\frac{1}{2}})D^{\frac{1}{2}}.$$

由矩阵 A 的对称性知矩阵 $I-D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ 也对称，因此矩阵 B 与矩阵 $I-D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ 相似，故 B 的特征值全为实数。

必要性：若 J 迭代法收敛，则 $\rho(B)<1$，即 $\rho(I-D^{-\frac{1}{2}}AD^{-\frac{1}{2}})<1$，所以 $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ 的特征值大于 0 小于 2，因此 $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ 是正定矩阵，由此知 A 也是正定矩阵。另一方面，

$$2D-A=D^{\frac{1}{2}}(2I-D^{-\frac{1}{2}}AD^{-\frac{1}{2}})D^{\frac{1}{2}},$$

所以 $2D-A$ 的特征值大于 0 小于 2，也是正定矩阵。

充分性：若矩阵 A 和 $2D-A$ 均正定，从而矩阵 $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ 和 $2I-D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ 正定，这等价于 $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ 的特征值大于 0 小于 2，由此知 $\rho(B)<1$，从而 J 法收敛。 证毕。

以下定理是下一节中定理 4.19 的特例。

定理 4.18 *若 A 为实正定对称矩阵，则 G-S 迭代法收敛。*

综上所述，J 迭代法与 G-S 迭代法的收敛矩阵类互不包含，当系数矩阵严格对角占优时，两种方法都收敛。

例 4.9 *设 $A\boldsymbol{x}=\boldsymbol{b}$ 的系数矩阵*

$$A=\begin{pmatrix}1 & -2 & 2\\ -1 & 1 & -1\\ -2 & -2 & 1\end{pmatrix}.$$

证明 J 迭代法收敛而 G-S 迭代法发散。

证明： 由(4.30)，$B_J=-D^{-1}(L+U)$，所以 B_J 的特征方程为

$$\begin{aligned}\det(\lambda I-B_J)&=\det[\lambda I+D^{-1}(L+U)]\\&=\det(D^{-1})\det[\lambda D+(L+U)]=0,\end{aligned}$$

等价于 $\det[\lambda D+(L+U)]=0$，即

$$\begin{vmatrix} \lambda & -2 & 2 \\ -1 & \lambda & -1 \\ -2 & -2 & \lambda \end{vmatrix}=0.$$

解得特征值为 $\lambda_1=\lambda_2=\lambda_3=0$，因此 $\rho(B_J)=0<1$，由定理 4.12 知 J 法收敛。

由(4.31)，$B_G=-(D+L)^{-1}U$，所以 B_G 的特征方程为

$$\begin{aligned}\det(\lambda I-B_G)&=\det[\lambda I+(D+L)^{-1}U]\\&=\det[(D+L)^{-1}]\det[\lambda(D+L)+U]=0,\end{aligned}$$

等价于 $\det[\lambda(D+L)+U]=0$，即

$$\begin{vmatrix} \lambda & -2 & 2 \\ -\lambda & \lambda & -1 \\ -2\lambda & -2\lambda & \lambda \end{vmatrix}=0.$$

解得特征值为 $\lambda_1=0$, $\lambda_2=2(\sqrt{2}-1)$, $\lambda_3=2(\sqrt{2}+1)$，因此 $\rho(B)=2(\sqrt{2}+1)>1$，由定理 4.12 知 G-S 迭代法发散。

4.3.3 超松弛迭代方法

超松弛迭代方法(简记为 SOR 方法)是为了提高收敛速度而发展起来的方法，是 G-S 迭代法的推广。其矩阵形式为

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}-\alpha D^{-1}[L\boldsymbol{x}^{(k+1)}+(D+U)\boldsymbol{x}^{(k)}-\boldsymbol{b}], \tag{4.35}$$

即

$$\begin{cases} \boldsymbol{x}^{(k+1)}=B_\alpha\boldsymbol{x}^{(k)}+f_\alpha; \\ B_\alpha=(D+\alpha L)^{-1}[(1-\alpha)D-\alpha U]; \\ f_\alpha=\alpha(D+\alpha L)^{-1}\boldsymbol{b}. \end{cases}$$

其中 $D=\mathrm{diag}(a_{11},a_{22},\cdots,a_{nn})$，$L$ 和 U 为 A 的严格上、下三角部分（不包括对角线），参数 α 称为超松弛因子。

(4.35) 的分量形式为

$$x_i^{(k+1)}=x_i^{(k)}+\frac{\alpha}{a_{ii}}\left(b_i-\sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)}-\sum_{j=i}^{n}a_{ij}x_j^{(k)}\right),\quad i=1,2,\cdots,n. \tag{4.36}$$

当 $\alpha=1$ 时，(4.35) 即为 G-S方法，可见 G-S 法是 SOR 方法的特例；$\alpha>1$ 时，算法 (4.35) 称为超松弛方法；$\alpha<1$ 时，算法 (4.35) 称为低松弛方法。α 的选取依据参见下述定理。

定理 4.19 *在矩阵 A 是实对称正定的条件下，超松弛法（4.35）收敛的充分必要条件是 $0<\alpha<2$。*

证明：只证必要性。由于 A 为实对称矩阵，故有

$$\begin{aligned} A &= D+L+L^T, \\ B_\alpha &= (D+\alpha L)^{-1}[(1-\alpha)D-\alpha L^T] \\ &= (I+\alpha D^{-1}L)^{-1}[(1-\alpha)I-\alpha D^{-1}L^T]. \end{aligned}$$

设迭代法（4.35）收敛，则 $\rho(B_\alpha)<1$，而

$$|\det(B_\alpha)| = |\lambda_1\lambda_2\cdots\lambda_n| \le (\rho(B_\alpha))^n < 1,$$

其中 $\lambda_i(i=1,2,\cdots,n)$ 为 B_α 的特征值。由于 L 为严格下三角阵，从而

$$\det(B_\alpha) = \det[(I+\alpha D^{-1}L)^{-1}]\cdot\det[(1-\alpha)I-\alpha D^{-1}L^T] = (1-\alpha)^n.$$

于是必有 $|1-\alpha|<1$，即 $0<\alpha<2$。 证毕。

定理 4.20 *若矩阵 A 是严格对角占优阵或不可约弱对角占优阵，则当 $0<\alpha\le 1$ 时，SOR 法收敛。*

在实际计算中，一般用试算的办法确定松弛因子。最简单的办法是从同一初始向量出发，取不同的松弛因子 α， 迭代相同的次数（注意迭代次数不应太少），然后比较相应的误差，选取使误差之模最小的松弛因子作为最优因子的近似值。

例 4.10 *用 SOR 法解方程组*

$$\begin{cases} 4x_1 - x_2 = 1 \\ -x_1 + 4x_2 - x_3 = 4 \\ -x_2 + 4x_3 = -3 \end{cases}$$

分别取 $\alpha=1.03$, $\alpha=1$ 及 $\alpha=1.1$， 要求当 $\|\boldsymbol{x}^*-\boldsymbol{x}^k\|_\infty \le 5\times 10^{-6}$ 时迭代终止，并对每一个 α 确定迭代次数。其中 $\boldsymbol{x}^*$ 为精确解 $\boldsymbol{x}^* = (\frac{1}{2},1,\frac{1}{2})^T$。

解： 据公式 (4.36), SOR 迭代格式为

$$\begin{cases} x_1^{(k+1)} = (1-\alpha)x_1^{(k)} + \dfrac{\alpha}{4}(1+x_2^{(k)}) \\ x_2^{(k+1)} = (1-\alpha)x_2^{(k)} + \dfrac{\alpha}{4}(4+x_1^{(k+1)}+x_3^{(k)}) \\ x_3^{(k+1)} = (1-\alpha)x_3^{(k)} + \dfrac{\alpha}{4}(-3+x_2^{(k+1)}). \end{cases}$$

取初始向量 $\boldsymbol{x}^{(0)} = (0,0,0)^T$， 当 $\alpha=1.03$ 时，迭代 5 次达到要求，

$$\boldsymbol{x}^{(5)} = (0.50000043, 1.0000001, -0.499999)^T.$$

当 $\alpha=1$ 时，迭代 6 次达到要求，

$$\boldsymbol{x}^{(6)}=(0.50000038,1.0000002,-0.4999995)^T.$$

当 $\alpha=1.1$ 时，迭代 6 次达到要求，

$$\boldsymbol{x}^{(6)}=(0.50000035,0.9999989,-0.5000003)^T.$$

4.4 极小化方法

本节主要介绍共轭梯度方法，它是一种极小化方法，对应于求一个二次函数的极值。该方法出现在 20 世纪 50 年代，特别是 80 至90 年代预处理共轭梯度方法的发展，使这类方法成为解大型稀疏方程组的有效方法。

4.4.1 与方程组等价的变分问题

设 $A\in\boldsymbol{R}^{n\times n}$, $\boldsymbol{b}\in\boldsymbol{R}^n$， A 对称正定，考虑方程组

$$A\boldsymbol{x}=\boldsymbol{b}. \tag{4.37}$$

定义二次函数 $\varphi:\boldsymbol{R}^n\to\boldsymbol{R}$ 为

$$\varphi(\boldsymbol{x})=\frac{1}{2}(A\boldsymbol{x},\boldsymbol{x})-(\boldsymbol{b},\boldsymbol{x}). \tag{4.38}$$

则函数 φ 有如下性质：

(1) 对一切 $\boldsymbol{x}\in\boldsymbol{R}^n$,

$$\mathbf{grad}\varphi(\boldsymbol{x})=A\boldsymbol{x}-\boldsymbol{b}. \tag{4.39}$$

(2) 对一切 $\boldsymbol{x},\boldsymbol{y}\in\boldsymbol{R}^n$, $\alpha\in\boldsymbol{R}$,

$$\begin{aligned}\varphi(\boldsymbol{x}+\alpha\boldsymbol{y})&=\frac{1}{2}(A(\boldsymbol{x}+\alpha\boldsymbol{y}),\boldsymbol{x}+\alpha\boldsymbol{y})-(\boldsymbol{b},\boldsymbol{x}+\alpha\boldsymbol{y})\\&=\varphi(\boldsymbol{x})+\alpha(A\boldsymbol{x}-\boldsymbol{b},\boldsymbol{y})+\frac{\alpha^2}{2}(A\boldsymbol{y},\boldsymbol{y}).\end{aligned} \tag{4.40}$$

(3) 设 $\boldsymbol{x}^*$ 为 (4.37)的解，则对一切 $\boldsymbol{x}\in\boldsymbol{R}^n$，有

$$\varphi(\boldsymbol{x})-\varphi(\boldsymbol{x}^*)=\frac{1}{2}(A(\boldsymbol{x}-\boldsymbol{x}^*),\boldsymbol{x}-\boldsymbol{x}^*). \tag{4.41}$$

证明：因为 $\boldsymbol{x}^*=A^{-1}\boldsymbol{b}$，所以

$$\begin{aligned}\varphi(\boldsymbol{x}^*)&=\frac{1}{2}(A\boldsymbol{x}^*,\boldsymbol{x}^*)-(\boldsymbol{b},\boldsymbol{x}^*)\\&=\frac{1}{2}(AA^{-1}\boldsymbol{b},A^{-1}\boldsymbol{b})-(\boldsymbol{b},A^{-1}\boldsymbol{b})\\&=-\tfrac{1}{2}(\boldsymbol{b},A^{-1}\boldsymbol{b})=-\tfrac{1}{2}(A\boldsymbol{x}^*,\boldsymbol{x}^*).\end{aligned}$$

则

$$\begin{aligned}&\frac{1}{2}(A(\boldsymbol{x}-\boldsymbol{x}^*),\boldsymbol{x}-\boldsymbol{x}^*)\\&=\frac{1}{2}(A\boldsymbol{x},\boldsymbol{x})-(A\boldsymbol{x}^*,\boldsymbol{x})+\frac{1}{2}(A\boldsymbol{x}^*,\boldsymbol{x}^*)\\&=\varphi(\boldsymbol{x})-\varphi(\boldsymbol{x}^*).\end{aligned}$$

定理 4.21 设 A 对称正定，则 $\boldsymbol{x}^*$ 为 (4.37) 的解的充要条件是 $\boldsymbol{x}^*$ 满足

$$\varphi(\boldsymbol{x}^*)=\min_{\boldsymbol{x}\in\boldsymbol{R}^n}\varphi(\boldsymbol{x}).$$

证明：设$\boldsymbol{x}^*$ 为 (4.37) 的解，由 (4.41) 及 A 的正定性，有 $\varphi(\boldsymbol{x})-\varphi(\boldsymbol{x}^*)\geq 0$，即 $\boldsymbol{x}^*$ 使 $\varphi(\boldsymbol{x})$ 最小。

反正，若有 $\overline{\boldsymbol{x}}$ 使 $\varphi(\boldsymbol{x})$ 达到最小，则 $\varphi(\overline{\boldsymbol{x}})\leq\varphi(\boldsymbol{x}^*)$ 又 $\varphi(\overline{\boldsymbol{x}})\geq\varphi(\boldsymbol{x}^*)$，所以 $\varphi(\overline{\boldsymbol{x}})-\varphi(\boldsymbol{x}^*)=0$，故由 (4.41) 可知

$$(A(\overline{\boldsymbol{x}}-\boldsymbol{x}^*),\overline{\boldsymbol{x}}-\boldsymbol{x}^*)=2(\varphi(\boldsymbol{x})-\varphi(\boldsymbol{x}^*))=0.$$

又因 A 正定，所以 $\overline{\boldsymbol{x}}=\boldsymbol{x}^*$。 证毕。

由上述定理，求解线性方程组 $A\boldsymbol{x}=\boldsymbol{b}$ 等价于求解其变分问题，即求 $\boldsymbol{x}\in\boldsymbol{R}^n$，使 $\varphi(\boldsymbol{x})$ 最小。求解的方法一般是构造一个向量序列 $\{\boldsymbol{x}_k\}$，使 $\varphi(\boldsymbol{x}_k)\to\min\varphi(\boldsymbol{x})$。下面我们介绍求二次函数的极小点的方法。

4.4.2 最速下降法与共轭梯度法的定义

设 $\boldsymbol{x}_0$ 是任意给定的一个初始点，从点 $\boldsymbol{x}_0$ 出发沿某一规定的方向 $\boldsymbol{p}_0$，求函数 $\varphi(\boldsymbol{x})$ 在直线

$$\boldsymbol{x}=\boldsymbol{x}_0+t\boldsymbol{p}_0$$

上的极小点，设求得的极小点为 $\boldsymbol{x}_1$。再从点 $\boldsymbol{x}_1$ 出发沿某一规定的方向 $\boldsymbol{p}_1$，求函数 $\varphi(\boldsymbol{x})$ 在直线

$$\boldsymbol{x}=\boldsymbol{x}_1+t\boldsymbol{p}_1$$

上的极小点，设求得的极小点为 $\boldsymbol{x}_2$。如此继续下去，一般地，从点 $\boldsymbol{x}_k$ 出发沿某一规定的方向 $\boldsymbol{p}_k$，求函数 $\varphi(\boldsymbol{x})$ 在直线

$$\boldsymbol{x}=\boldsymbol{x}_k+t\boldsymbol{p}_k$$

上的极小点 $\boldsymbol{x}_{k+1}$。称 $\boldsymbol{p}_k$ 为搜索方向。

命题 4.1 对于已知的 $\boldsymbol{x}_k$, $\boldsymbol{p}_k(\boldsymbol{p}_k\neq 0)$，函数 $\varphi(\boldsymbol{x})$ 在直线

$$\boldsymbol{x}=\boldsymbol{x}_k+t\boldsymbol{p}_k$$

上的极小点 $\boldsymbol{x}_{k+1}$ 为

$$\boldsymbol{x}_{k+1}=\boldsymbol{x}_k+\alpha_k\boldsymbol{p}_k$$

其中 $\alpha_k=-\dfrac{\boldsymbol{r}_k^T\boldsymbol{p}_k}{\boldsymbol{p}_k^TA\boldsymbol{p}_k}$，$\boldsymbol{r}_k=A\boldsymbol{x}_k-\boldsymbol{b}$。

证明：记 $f(t)=\varphi(\boldsymbol{x}_k+t\boldsymbol{p}_k)$，欲确定系数 α_k 使得一元函数 f 当 $t=\alpha_k$ 时为极小。由 (4.40)，

$$f(t)=\varphi(\boldsymbol{x}_k)+t(A\boldsymbol{x}_k-\boldsymbol{b},\boldsymbol{p}_k)+\frac{t^2}{2}(A\boldsymbol{p}_k,\boldsymbol{p}_k).$$

令 $f'(t)=0$ 得

$$t=\alpha_k=-\frac{(A\boldsymbol{x}_k-\boldsymbol{b})^T\boldsymbol{p}_k}{\boldsymbol{p}_k^TA\boldsymbol{p}_k}=-\frac{\boldsymbol{r}_k^T\boldsymbol{p}_k}{\boldsymbol{p}_k^TA\boldsymbol{p}_k}.$$

又由于 $f''=\boldsymbol{p}_k^TA\boldsymbol{p}_k>0$ $(\boldsymbol{p}_k\neq 0)$，因此 $t=\alpha_k$ 时，$f(t)$ 极小。从而

$$\boldsymbol{x}_{k+1}=\boldsymbol{x}_k+\alpha_k\boldsymbol{p}_k.$$

证毕。

注 4.2 在命题 4.1中，余量 $\boldsymbol{r}_k=A\boldsymbol{x}_k-\boldsymbol{b}=\mathbf{grad}\varphi(\boldsymbol{x}_k)$。命题 4.1 所得的迭代公式

$$\begin{cases}\boldsymbol{x}_{k+1}=\boldsymbol{x}_k+\alpha_k\boldsymbol{p}_k\\ \alpha_k=-\dfrac{\boldsymbol{r}_k^T\boldsymbol{p}_k}{\boldsymbol{p}_k^TA\boldsymbol{p}_k}\end{cases}\tag{4.42}$$

具有下降性

$$\varphi(\boldsymbol{x}_{k+1})\leq\varphi(\boldsymbol{x}_k)).$$

(1) 若取

$$\boldsymbol{p}_k=-\boldsymbol{r}_k=-\mathbf{grad}\varphi(\boldsymbol{x}_k),$$

则称迭代法 (4.42) 为最速下降法。此时

$$\begin{cases}\boldsymbol{x}_{k+1}=\boldsymbol{x}_k-\alpha_k\boldsymbol{r}_k\\ \alpha_k=\dfrac{(\boldsymbol{r}_k,\boldsymbol{r}_k)}{(A\boldsymbol{r}_k,\boldsymbol{r}_k)}.\end{cases}$$

当 $\boldsymbol{r}_k\neq 0$ 时，

$$\begin{aligned}\varphi(\boldsymbol{x}_{k+1})-\varphi(\boldsymbol{x}_k)&=\varphi(\boldsymbol{x}_k-\alpha_k\boldsymbol{r}_k)-\varphi(\boldsymbol{x}_k)\\&=-\alpha_k(\boldsymbol{r}_k,\boldsymbol{r}_k)+\frac{1}{2}\alpha_k^2\frac{(\boldsymbol{r}_k,\boldsymbol{r}_k)^2}{(A\boldsymbol{r}_k,\boldsymbol{r}_k)}\\&=-\frac{1}{2}\frac{(\boldsymbol{r}_k,\boldsymbol{r}_k)^2}{(A\boldsymbol{r}_k,\boldsymbol{r}_k)}<0,\end{aligned}$$

因而 $\varphi(\boldsymbol{x}_{k+1}) < \varphi(\boldsymbol{x}_k)$)。可以证明向量序列 $\{\boldsymbol{x}^k\}$ 收敛于方程组 $A\boldsymbol{x} = \boldsymbol{b}$ 的解，并且相邻两次的搜索方向是正交的， 即 $(\boldsymbol{r}_{k+1}, \boldsymbol{r}_k) = 0$。

最速下降法因舍入误差的影响，计算将出现不稳定的现象，很少在实际中直接应用。

(2) 如取搜索方向

$$\boldsymbol{p}_0,\ \boldsymbol{p}_1,\ \boldsymbol{p}_2,\ \cdots$$

为 $\boldsymbol{R}^n$ 中的一个A 共轭向量系，即有性质

$$\boldsymbol{p}_i^T A \boldsymbol{p}_j = 0, \quad i \neq j \tag{4.43}$$

的向量系 $\{\boldsymbol{p}_k\}$，且 $\boldsymbol{p}_k \neq 0$，$k = 0, 1, 2\cdots$，则称迭代法 (4.42) 为共轭梯度法（简称为 CG 法）。

我们用 L_k 表示线性无关的向量系 $\boldsymbol{p}_k$ 张成的子空间，即

$$L_k = span\{\boldsymbol{p}_0, \boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_{k-1}\},$$

令 $S_k = \{\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{u} \mid \boldsymbol{u} \in L_k\}$.

定理 4.22 *从任意一点 $\boldsymbol{x}_0$ 出发，得到的点序列 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots$ 具有性质*

$$\varphi(\boldsymbol{x}_k) = \min \varphi(\boldsymbol{x}_0 + \boldsymbol{u}),\ \boldsymbol{u} \in L_k \tag{4.44}$$

的充分必要条件是 $\boldsymbol{x}_k \in S_k$ 且余量 $\boldsymbol{r}_k$ 和 L_k 正交，即

$$(\boldsymbol{r}_k, \boldsymbol{u}) = 0, \quad \forall \boldsymbol{u} \in L_k. \tag{4.45}$$

证明：必要性，(4.44) 表明, $\boldsymbol{x}_k$ 为二次函数 $\varphi(\boldsymbol{x})$ 在 S_k 上的极小点，因此 $\varphi(\boldsymbol{x})$ 在 $\boldsymbol{x}_k$ 沿任一方向 $\boldsymbol{u} \in L_K$ 的方向导数必须为零，从而有

$$(\text{grad}\varphi(\boldsymbol{x}_k), u) = (\boldsymbol{r}_k, \boldsymbol{u}) = 0, \quad \forall \boldsymbol{u} \in L_k.$$

充分性，任取 $\boldsymbol{x} \in S_k$，令

$$\triangle \boldsymbol{x}_k = \boldsymbol{x} - \boldsymbol{x}_k, \quad \boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{u}, \quad \boldsymbol{x}_k = \boldsymbol{x}_0 + \boldsymbol{u}_k,$$

其中 $\boldsymbol{u},\ \boldsymbol{u}_k \in L_k$，则 $\Delta \boldsymbol{x}_k = \boldsymbol{u} - \boldsymbol{u}_k$，于是

$$\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{x}_k) = (\boldsymbol{r}_k, \Delta \boldsymbol{x}_k) + \frac{1}{2}(\Delta \boldsymbol{x}_k, A\Delta \boldsymbol{x}_k.)$$

由 (4.45),

$$(\boldsymbol{r}_k, \Delta \boldsymbol{x}_k) = (\boldsymbol{r}_k, \boldsymbol{u}) - (\boldsymbol{r}_k, \boldsymbol{u}_k) = 0.$$

又因 A 正定，$(\Delta \boldsymbol{x}_k, A\Delta \boldsymbol{x}_k) \geq 0$，故有 $\varphi(\boldsymbol{x}) \geq \varphi(\boldsymbol{x}_k)$。 证毕。

由定理 4.22 可得到共轭梯度法的重要结论：

定理 4.23 由共轭梯度法得到的点列 $\{\boldsymbol{x}_k\}$ 满足 (4.44).

证明：由定理 4.22 只需证明 $\boldsymbol{x}_k \in S_k$ 且余量 $\boldsymbol{r}_k$ 和 L_k 正交。事实上，

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} + \alpha_{k-1}\boldsymbol{p}_{k-1} = \boldsymbol{x}_{k-2} + \alpha_{k-2}\boldsymbol{p}_{k-2} + \alpha_{k-1}\boldsymbol{p}_{k-1}$$

$$= \cdots = \boldsymbol{x}_0 + \sum_{i=0}^{k-1} \alpha_i \boldsymbol{p}_i \in S_k.$$

因为对任意 $j = 0, 1, \cdots, k-1$，有

$$\begin{aligned}(\boldsymbol{r}_k, \boldsymbol{p}_j) &= \boldsymbol{p}_j^T(A\boldsymbol{x}_k - \boldsymbol{b}) = \boldsymbol{p}_j^T\left(\boldsymbol{r}_0 + \sum_{i=0}^{k-1}\alpha_i A\boldsymbol{p}_i\right) \\ &= \boldsymbol{p}_j^T\boldsymbol{r}_0 + \sum_{i=0}^{k-1}\alpha_i \boldsymbol{p}_j^T A\boldsymbol{p}_i = \boldsymbol{p}_j^T\boldsymbol{r}_0 + \alpha_j \boldsymbol{p}_j^T A\boldsymbol{p}_j \\ &= \boldsymbol{p}_j^T\boldsymbol{r}_0 - \boldsymbol{p}_j^T\boldsymbol{r}_j = \boldsymbol{p}_j^T\boldsymbol{r}_0 - \boldsymbol{p}_j^T(A\boldsymbol{x}_j - \boldsymbol{b}) \\ &= \boldsymbol{p}_j^T\boldsymbol{r}_0 - \boldsymbol{p}_j^T\left(\boldsymbol{r}_0 + \sum_{i=0}^{j-1}\alpha_i A\boldsymbol{p}_i\right) \\ &= \boldsymbol{p}_j^T\boldsymbol{r}_0 - \boldsymbol{p}_j^T\boldsymbol{r}_0 = 0,\end{aligned}$$

所以 $\boldsymbol{r}_k$ 和 L_k 正交。 证毕。

定理 4.24 共轭梯度法至多进行 n 步便得到方程组 (4.37) 的解。

证明：由定理 4.22，4.23 知 $(\boldsymbol{r}_n, \boldsymbol{p}_i) = 0$, $i = 0, 1, \cdots, n-1$，而 $\boldsymbol{p}_0, \boldsymbol{p}_1, \cdots, \boldsymbol{p}_{n-1}$ 为非零的 A 共轭向量系，必为线性无关，因此

$$\boldsymbol{r}_n = A\boldsymbol{x}_n - \boldsymbol{b} = 0,$$

故 $\boldsymbol{x}_n$ 一定是方程组 (4.37) 的解。 证毕。

4.4.3 共轭梯度法的计算公式

下面介绍共轭梯度法一种生成 A 共轭向量系 $\{\boldsymbol{p}_i\}$ 的具体方法。

对于任意初始向量 $\boldsymbol{x}_0$，取第一个搜索方向 $\boldsymbol{p}_0$ 为

$$\boldsymbol{p}_0 = -\boldsymbol{r}_0 = -(A\boldsymbol{x}_0 - \boldsymbol{b}).$$

由 (4.42) 计算 $\boldsymbol{x}_1$:

$$\begin{cases} \boldsymbol{x}_1 = \boldsymbol{x}_0 + \alpha_0\boldsymbol{p}_0 \\ \alpha_0 = -\dfrac{\mathbf{r}_0^T\boldsymbol{p}_0}{\boldsymbol{p}_0^T A\boldsymbol{p}_0}, \end{cases}$$

再计算

$$\boldsymbol{r}_1 = A\boldsymbol{x}_1 - \boldsymbol{b}.$$

而 $\boldsymbol{r}_1 \perp \boldsymbol{p}_0$，因此 $\boldsymbol{r}_1 \perp \boldsymbol{r}_0$，于是我们可在 $\boldsymbol{r}_1, \boldsymbol{r}_0$ 张成的空间中搜索方向 $\boldsymbol{p}_1$. 令

$$\boldsymbol{p}_1 = -\boldsymbol{r}_1 - \beta_0 \boldsymbol{r}_0 = -\boldsymbol{r}_1 + \beta_0 \boldsymbol{p}_0,$$

要使 $\boldsymbol{p}_0$ 与 $\boldsymbol{p}_1$ 为 A 共轭，则

$$\boldsymbol{p}_1^T A \boldsymbol{p}_0 = (-\boldsymbol{r}_1 + \beta_0 \boldsymbol{p}_0)^T A \boldsymbol{p}_0 = 0,$$

从而得到

$$\beta_0 = (\boldsymbol{r}_1, A\boldsymbol{p}_0)/(\boldsymbol{p}_0, A\boldsymbol{p}_0).$$

再由 (4.42) 计算 $\boldsymbol{x}_2$:

$$\begin{cases} \boldsymbol{x}_2 = \boldsymbol{x}_1 + \alpha_1 \boldsymbol{p}_1 \\ \alpha_1 = -\dfrac{\boldsymbol{r}_1^T \boldsymbol{p}_1}{\boldsymbol{p}_1^T A \boldsymbol{p}_1}, \end{cases}$$

再计算

$$\boldsymbol{r}_2 = A\boldsymbol{x}_2 - \boldsymbol{b}.$$

令

$$\boldsymbol{p}_2 = -\boldsymbol{r}_2 + \beta_1 \boldsymbol{p}_1,$$

要使 $\boldsymbol{p}_1$ 与 $\boldsymbol{p}_2$ 为 A 共轭，必须

$$\boldsymbol{p}_2^T A \boldsymbol{p}_1 = (-\boldsymbol{r}_2 + \beta_1 \boldsymbol{p}_1)^T A \boldsymbol{p}_1 = 0$$

从而得到

$$\beta_1 = (\boldsymbol{r}_2, A\boldsymbol{p}_1)/(\boldsymbol{p}_1, A\boldsymbol{p}_1).$$

依次类推，一般地，计算

$$\boldsymbol{r}_{k+1} = A\boldsymbol{x}_{k+1} - \boldsymbol{b}.$$

若 $\boldsymbol{r}_{k+1} = \boldsymbol{0}$，则 $\boldsymbol{x}_{k+1}$ 为方程组 (4.37) 的解，停止计算。否则，令

$$\boldsymbol{p}_{k+1} = -\boldsymbol{r}_{k+1} + \beta_k \boldsymbol{p}_k$$

要使 $\boldsymbol{p}_{k+1}$ 与 $\boldsymbol{p}_k$ 为 A 共轭，必须

$$\beta_k = (\boldsymbol{r}_{k+1}, A\boldsymbol{p}_k)/(\boldsymbol{p}_k, A\boldsymbol{p}_k).$$

又因为

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$$

$$A\boldsymbol{x}_{k+1} = A\boldsymbol{x}_k + \alpha_k A\boldsymbol{p}_k$$

所以

$$\boldsymbol{r}_{k+1}=\boldsymbol{r}_k+\alpha_k A\boldsymbol{p}_k. \tag{4.46}$$

这样我们便得到共轭梯度法的计算公式：

给定初始近似向量 $\boldsymbol{x}_0$，取

$$\boldsymbol{p}_0=-\boldsymbol{r}_0=-(A\boldsymbol{x}_0-\boldsymbol{b}),$$

对于 $k=0,1,2,\cdots$ 计算

$$\begin{cases}\alpha_k=-\dfrac{\boldsymbol{r}_k^T\boldsymbol{p}_k}{\boldsymbol{p}_k^TA\boldsymbol{p}_k},\\ \boldsymbol{x}_{k+1}=\boldsymbol{x}_k+\alpha_k\boldsymbol{p}_k,\\ \boldsymbol{r}_{k+1}=A\boldsymbol{x}_{k+1}-\boldsymbol{b}=\boldsymbol{r}_k+\alpha_kA\boldsymbol{p}_k,\\ \beta_k=(\boldsymbol{r}_{k+1},A\boldsymbol{p}_k)/(\boldsymbol{p}_k,A\boldsymbol{p}_k),\\ \boldsymbol{p}_{k+1}=-\boldsymbol{r}_{k+1}+\beta_k\boldsymbol{p}_k.\end{cases} \tag{4.47}$$

计算过程中，若 $\boldsymbol{r}_{k+1}=\boldsymbol{0}$，则 $\boldsymbol{x}_{k+1}$ 为方程组 (4.37) 的解，停止计算。

例 4.11 *应用共轭梯度法解方程组*

$$\begin{pmatrix}2&0&1\\0&1&0\\1&0&2\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}3\\1\\3\end{pmatrix}.$$

解：显然方程组系数矩阵为对称正定矩阵。取初始近似 $\boldsymbol{x}_0=(0,0,0)^T$，则

$$\boldsymbol{r}_0=A\boldsymbol{x}_0-\boldsymbol{b}=(-3,-1,-3)^T,$$

$$\boldsymbol{p}_0=-\boldsymbol{r}_0=(3,1,3)^T.$$

对于 $k=0$，计算得

$$A\boldsymbol{p}_0=(9,1,9)^T,\quad \boldsymbol{p}_0^TA\boldsymbol{p}_0=55,$$

因此

$$\alpha_0=-\frac{\boldsymbol{r}_0^T\boldsymbol{p}_0}{\boldsymbol{p}_0^TA\boldsymbol{p}_0}=\frac{19}{55},\quad \boldsymbol{x}_1=\boldsymbol{x}_0+\alpha_0\boldsymbol{p}_0=\frac{19}{55}(3,1,3)^T,$$

$$\boldsymbol{r}_1=A\boldsymbol{x}_1-\boldsymbol{b}=\boldsymbol{r}_0+\alpha_0A\boldsymbol{p}_0=\frac{6}{55}(1,-6,1)^T,$$

$$\beta_0=(\boldsymbol{r}_1,A\boldsymbol{p}_0)/(\boldsymbol{p}_0,A\boldsymbol{p}_0)=\frac{72}{55^2},$$

$$\boldsymbol{p}_1=-\boldsymbol{r}_1+\beta_0\boldsymbol{p}_0=\frac{114}{55^2}(-1,18,-1)^T.$$

对于 $k=1$，计算得

$$A\boldsymbol{p}_1=\frac{18\times 19}{55^2}(-1,6,-1)^T,\quad \boldsymbol{p}_1^TA\boldsymbol{p}_1=\frac{6^3\times 19^2}{55^3},$$

因此

$$\alpha_1=-\frac{\boldsymbol{r}_1^T\boldsymbol{p}_1}{\boldsymbol{p}_1^TA\boldsymbol{p}_1}=\frac{55}{3\times 19},\quad \boldsymbol{x}_2=\boldsymbol{x}_1+\alpha_1\boldsymbol{p}_1=(1,1,1)^T,$$

$$\boldsymbol{r}_2=A\boldsymbol{x}_2-\boldsymbol{b}=0.$$

故两次迭代便得到方程组的解 $\boldsymbol{x}_2=(1,1,1)^T$.

4.4.4 共轭梯度法的性质

本节讨论共轭梯度法的若干性质，其中向量系 $\boldsymbol{p}_i$ 按照 4.4.3 节中的方法生成，并且它确实是一个 A 共轭向量系。

定理 4.25 *若 $\boldsymbol{r}_k\neq 0$，则共轭梯度法具有下列性质：*

$$span\{\boldsymbol{r}_0,\boldsymbol{r}_1,\cdots,\boldsymbol{r}_k\}=span\{\boldsymbol{r}_0,A\boldsymbol{r}_0,\cdots,A^k\boldsymbol{r}_0\};\tag{4.48}$$

$$span\{\boldsymbol{p}_0,\boldsymbol{p}_1,\cdots,\boldsymbol{p}_k\}=span\{\boldsymbol{r}_0,A\boldsymbol{r}_0,\cdots,A^k\boldsymbol{r}_0\};\tag{4.49}$$

$$\boldsymbol{p}_k^TA\boldsymbol{p}_i=0,\quad i=0,1,2,\cdots,k-1.\tag{4.50}$$

证明：用归纳法不难证明 (4.48) 和 (4.49) 成立。这里主要用归纳法证明 (4.50) 成立。当 $k=1$ 时，由 $\boldsymbol{p}_1$ 的生成方法知 $\boldsymbol{p}_1^TA\boldsymbol{p}_0=0$。现假设直到 k，(4.50) 成立，下证 $k+1$ 时(4.50) 亦成立。因为

$$\boldsymbol{p}_{k+1}^TA\boldsymbol{p}_i=(-\boldsymbol{r}_{k+1}+\beta_k\boldsymbol{p}_k)^TA\boldsymbol{p}_i=-\boldsymbol{r}_{k+1}^TA\boldsymbol{p}_i+\beta_k\boldsymbol{p}_k^TA\boldsymbol{p}_i,$$

若 $i=k$，则

$$\boldsymbol{p}_{k+1}^TA\boldsymbol{p}_k=-\boldsymbol{r}_{k+1}^TA\boldsymbol{p}_k+\frac{\boldsymbol{r}_{k+1}^TA\boldsymbol{p}_k}{\boldsymbol{p}_K^TA\boldsymbol{p}_k}\boldsymbol{p}_k^TA\boldsymbol{p}_i=0.$$

若 $i<k$，由 (4.49)，$\boldsymbol{p}_i\in span\{\boldsymbol{r}_0,A\boldsymbol{r}_0,\cdots,A^i\boldsymbol{r}_0\}$，所以

$$A\boldsymbol{p}_i\in span\{\boldsymbol{r}_0,A\boldsymbol{r}_0,\cdots,A^{i+1}\boldsymbol{r}_0\}=span\{\boldsymbol{p}_0,\boldsymbol{p}_1,\cdots,\boldsymbol{p}_{i+1}\}=L_{i+2}.$$

由定理 4.23 的证明，$\boldsymbol{r}_{k+1}$ 与 L_{k+1} 正交，所以，$-\boldsymbol{r}_{k+1}^TA\boldsymbol{p}_i=0$。根据归纳假设有 $\boldsymbol{p}_k^TA\boldsymbol{p}_i=0$，故 $\boldsymbol{p}_{k+1}^TA\boldsymbol{p}_i=0$，$i=0,1,\cdots,k$，从而证明了 (4.50)。 证毕。

定理 4.26 *共轭梯度法中的余量互为正交，即*

$$(\boldsymbol{r}_i,\boldsymbol{r}_j)=0,\quad i\neq j\tag{4.51}$$

且余量与搜索方向向量满足：

$$\boldsymbol{r}_k^T\boldsymbol{p}_i=0,\quad i=0,1,2,\cdots,k-1; \tag{4.52}$$

$$\boldsymbol{p}_k^T\boldsymbol{r}_i=-\boldsymbol{r}_k^T\boldsymbol{r}_k,\quad i=1,2,\cdots,k. \tag{4.53}$$

证明：由定理4.22，4.23 知 (4.52) 成立。因为

$$\boldsymbol{p}_k^T\boldsymbol{r}_k=\boldsymbol{p}_k^T(\boldsymbol{r}_{k-1}+\alpha_{k-1}A\boldsymbol{p}_{k-1})=\boldsymbol{p}_k^T\boldsymbol{r}_{k-1}=\cdots=\boldsymbol{p}_k^T\boldsymbol{r}_0 \tag{4.54}$$

$$\boldsymbol{p}_k^T\boldsymbol{r}_k=(-\boldsymbol{r}_k+\beta_{k-1}\boldsymbol{p}_{k-1})^T\boldsymbol{r}_k=-\boldsymbol{r}_k^T\boldsymbol{r}_k+\beta_{k-1}\boldsymbol{p}_{k-1}^T\boldsymbol{r}_k=-\boldsymbol{r}_k^T\boldsymbol{r}_k,$$

所以 (4.53)式成立。

其次，用归纳法证明

$$(\boldsymbol{r}_i,\boldsymbol{r}_j)=0,\quad i\neq j.$$

显然 $(\boldsymbol{r}_1,\boldsymbol{r}_0)=0$. 假设 $(\boldsymbol{r}_k,\boldsymbol{r}_j)=0,\quad j=0,1,2,\cdots,k-1$. 以下证明

$$(\boldsymbol{r}_{k+1},\boldsymbol{r}_j)=0,\quad j=0,1,2,\cdots,k.$$

据定理4.25知，$\boldsymbol{r}_j$ 可以表示成 $\boldsymbol{p}_0,\boldsymbol{p}_1,\cdots\boldsymbol{p}_j$ 的线性组合：

$$\boldsymbol{r}_j=a_0\boldsymbol{p}_0+a_1\boldsymbol{p}_1+\cdots+a_j\boldsymbol{p}_j,$$

因此

$$\boldsymbol{r}_{k+1}^T\boldsymbol{r}_j=a_0\boldsymbol{r}_{k+1}^T\boldsymbol{p}_0+a_1\boldsymbol{r}_{k+1}^T\boldsymbol{p}_1+\cdots+a_j\boldsymbol{r}_{k+1}^T\boldsymbol{p}_j,$$

由（4.52），上式右端为零，因此

$$(\boldsymbol{r}_{k+1},\boldsymbol{r}_j)=0,\quad j=0,1,2,\cdots,k.$$

证毕。

由定理4.26，还可将计算公式 (4.47)中 α_k,β_k 的表达式写成

$$\alpha_k=\frac{\boldsymbol{r}_k^T\boldsymbol{r}_k}{\boldsymbol{p}_k^TA\boldsymbol{p}_k}, \tag{4.55}$$

$$\beta_k=(\boldsymbol{r}_{k+1},\boldsymbol{r}_{k+1})/(\boldsymbol{r}_k,\boldsymbol{r}_k). \tag{4.56}$$

由(4.47)第1式及(4.53) 知 (4.55) 成立。又由 (4.46) 知

$$A\boldsymbol{p}_k=\alpha_k^{-1}(\boldsymbol{r}_{k+1}-\boldsymbol{r}_k)$$

代入(4.47) 第4式, 得

$$\beta_k = (\boldsymbol{r}_{k+1}, A\boldsymbol{p}_k)/(\boldsymbol{p}_k, A\boldsymbol{p}_k) = (\boldsymbol{r}_{k+1}, \alpha_k^{-1}(\boldsymbol{r}_{k+1} - \boldsymbol{r}_k))/(\boldsymbol{p}_k, A\boldsymbol{p}_k).$$

又由 (4.51)及(4.55), 即得 (4.56)。

理论上 CG 法经过有限步迭代便可得到方程组的准确解，因此 CG 法本质上是一种直接方法。在实际问题（特别是大型方程组）的计算中， 由于舍入误差的存在，很难在有限步得到准确解，且其计算公式具有迭代格式的特点，因此被作为一种迭代法使用。应用 CG 法解方程组得到的近似解序列 $\{\boldsymbol{x}_k\}$ 具有下述性质：

定理 4.27 (1) 当 $i < j$ 时，$\boldsymbol{x}_j$ 比 $\boldsymbol{x}_i$ 更接近于方程组的准确解 $\boldsymbol{x}^*$，即

$$\|\boldsymbol{x}^* - \boldsymbol{x}_j\|_2 < \|\boldsymbol{x}^* - \boldsymbol{x}_i\|_2$$

(2) $\|\boldsymbol{x}_i - \boldsymbol{x}^*\|_A \leq \left[\dfrac{\sqrt{K}-1}{\sqrt{K}+1}\right]^i \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_A$，其中 $\|\boldsymbol{x}\|_A = \sqrt{(A\boldsymbol{x}, \boldsymbol{x})}$，$K = \mathrm{cond}_2(A)$。

4.4.5 预处理共轭梯度法

从定理 4.27 的第 (2) 个估计式可知，当 A 病态，即条件数 $K \gg 1$ 时， 求解线性方程组 $A\boldsymbol{x} = \boldsymbol{b}$ 的 CG 法将收敛很慢。此时可考虑与 $A\boldsymbol{x} = \boldsymbol{b}$ 等价的方程组，而新方程组系数矩阵的条件数较小， 这就是预处理法。如果是用 CG 法求解新方程组，进而得到 $A\boldsymbol{x} = \boldsymbol{b}$ 的解，这就是预处理共轭梯度法（简记为 PCG 法）。

若 A 对称正定，那么经预处理后的方程组希望也是对称正定的。设矩阵 $S \in \boldsymbol{R}_n^{n\times n}$ 非奇异，则矩阵 $M = SS^T$ 是对称正定的。线性方程组 $A\boldsymbol{x} = \boldsymbol{b}$ 等价于以下线性方程组

$$S^{-1}A(S^{-1})^T\boldsymbol{u} = S^{-1}\boldsymbol{b}, \quad \boldsymbol{x} = (S^{-1})^T\boldsymbol{u}.$$

令 $G = S^{-1}A(S^{-1})^T$，$\boldsymbol{f} = S^{-1}\boldsymbol{b}$，则以上新方程组为

$$G\boldsymbol{u} = \boldsymbol{f}. \tag{4.57}$$

用 CG 法解 (4.57)，设初始向量为 $\boldsymbol{u}_0$，则初始余量 $\tilde{\boldsymbol{r}}_0 = G\boldsymbol{u}_0 - \boldsymbol{f}$， 初始方向为 $\tilde{\boldsymbol{p}}_0 = -\tilde{\boldsymbol{r}}_0$，记经 CG 法得到的解、余量及共轭方向分别为 $\tilde{\boldsymbol{u}}_k$、 $\tilde{\boldsymbol{r}}_k$ 及 $\tilde{\boldsymbol{p}}_k$。令 $\boldsymbol{x}_k = (S^{-1})^T\tilde{\boldsymbol{u}}_k$，而

$$\tilde{\boldsymbol{r}}_k = G\tilde{\boldsymbol{u}}_k - \boldsymbol{f} = S^{-1}(A(S^{-1})^TS^T\boldsymbol{x}_k - \boldsymbol{b}) = S^{-1}\boldsymbol{r}_k.$$

令 $\boldsymbol{p}_k = (S^{-1})^T\tilde{\boldsymbol{p}}_k$，$\boldsymbol{p}_0 = -(S^{-1})^TS^{-1}\boldsymbol{r}_0$，及 $\boldsymbol{z}_k = (S^{-1})^TS^{-1}\boldsymbol{r}_k = M^{-1}\boldsymbol{r}_k$。将解 (4.57) 的 CG 法迭代公式换回原方程组的变量， 得预处理共轭梯度法（PCG）为：给定初始向量 $\boldsymbol{x}_0$，取

$$\boldsymbol{r}_0 = A\boldsymbol{x}_0 - \boldsymbol{b},\ \boldsymbol{z}_0 = M^{-1}\boldsymbol{r}_0,\ \boldsymbol{p}_0 = -\boldsymbol{z}_0,$$

对于 $k=0,1,2,\cdots$ 计算

$$\begin{cases} \alpha_k = -\dfrac{(\boldsymbol{z}_k, \boldsymbol{r}_k)}{(\boldsymbol{p}_k, A\boldsymbol{p}_k)}, \\ \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k, \\ \boldsymbol{r}_{k+1} = \boldsymbol{r}_k + \alpha_k A\boldsymbol{p}_k, \\ \boldsymbol{z}_{k+1} = M^{-1}\boldsymbol{r}_{k+1}, \\ \beta_k = (\boldsymbol{z}_{k+1}, \boldsymbol{r}_{k+1})/(\boldsymbol{z}_k, \boldsymbol{r}_k), \\ \boldsymbol{p}_{k+1} = -\boldsymbol{z}_{k+1} + \beta_k \boldsymbol{p}_k. \end{cases}$$

上述公式中计算 $\boldsymbol{z}_{k+1} = M^{-1}\boldsymbol{r}_{k+1}$ 时，一般是用解方程组 $M\boldsymbol{z}_{k+1} = \boldsymbol{r}_{k+1}$ 的方式来实现。

如果矩阵 S 使得 $M = SS^T \approx A$，则 $G \approx S^{-1}SS^T(S^{-1})^T = I$，那么 $\mathrm{cond}(G) \approx 1$，这样预处理后的方程组系数矩阵的条件数就得到改善。一般地，可以将矩阵 A 作分裂式 $A = M - N$，其中 M 是对称正定阵，而 N “尽可能小”，再将 M 作 Cholesky 分解 $M = LL^T$，即 $S = L$。

第四章习题

1. 用 Gauss 消去法和 Doolittle 方法求解

$$\begin{pmatrix} 6 & 2 & 1 & -1 \\ 2 & 4 & 1 & 0 \\ 1 & 1 & 4 & -1 \\ -1 & 0 & -1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 6 \\ 1 \\ 5 \\ -5 \end{pmatrix}.$$

2. 用列主元法求解

$$\begin{pmatrix} -0.001 & 1 & 1 \\ 1 & 0.8 & 0 \\ 4 & 5.5 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0.2 \\ 1.4 \\ 7.4 \end{pmatrix}.$$

3. 用改进平方根法解

$$\begin{pmatrix} 16 & 4 & 8 \\ 4 & 5 & 4 \\ 8 & -4 & 22 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -1 \\ 3 \\ 5 \end{pmatrix}.$$

4. 设 A 是对称正定的三对角矩阵，

$$A = \begin{pmatrix} a_1 & c_1 & & & 0 \\ u_2 & a_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & u_{n-1} & a_{n-1} & c_{n-1} \\ 0 & & & u_n & a_n \end{pmatrix},$$

试给出用 Cholesky 方法解 $A\boldsymbol{x} = \boldsymbol{b}$ 的计算公式。

5. 用追赶法解三对角方程组

$$\begin{bmatrix} 4 & -1 & & \\ -2 & 4 & -1 & \\ & -2 & 4 & -1 \\ & & -2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

6. 用广义逆矩阵的方法求矛盾方程组

$$\begin{cases} x_1 - x_3 + x_4 = 4 \\ 2x_2 + 2x_3 + 2x_4 = 1 \\ -x_1 + 4x_2 + 5x_3 + 3x_4 = 2 \end{cases}$$

的全部最小二乘解和极小范数最小二乘解。

7. 用矩阵 QR 分解的方法求矛盾方程组

$$\begin{cases} 3x_2 + x_3 = 0 \\ 4x_2 - 2x_3 = 0 \\ 2x_1 + x_2 + 2x_3 = 0 \\ 2x_1 + 4x_2 + 3x_3 = 1 \end{cases}$$

的最小二乘解。

8. 用矩阵奇异值分解的方法求矛盾方程组

$$\begin{cases} x_1 + x_3 = 0 \\ x_1 + x_2 = 0 \\ 2x_1 + x_2 + x_3 = 1 \end{cases}$$

的全部最小二乘解和极小范数最小二乘解。

9. 设 $A \in \boldsymbol{R}^{m\times n}$ 列满秩，令条件数 $\mathrm{cond}_2(A) = \|A\|_2\|A^+\|_2$，证明 $\mathrm{cond}_2(A) = \sqrt{\|A^TA\|_2\|(A^TA)^{-1}\|_2}$.

10. 设 $A = \begin{pmatrix} 1 & \alpha & \beta \\ 1 & 1 & 1 \\ -\beta & \alpha & 1 \end{pmatrix}$，其中 α,β 是实参数。对方程组 $A\boldsymbol{x} = \boldsymbol{b}$ 来说，在 (α,β) 平面什么范围内 J 法收敛。

11. 分别用 J 法，G-S 法和 $\alpha = 1.46$ 的 SOR 法解方程组

$$\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ & & -1 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

并写出相应的迭代矩阵。

12. 设

$$A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & -1 & 2 \\ 1 & 1 & 1 \\ 1 & 1 & -2 \end{pmatrix}$$

证明：（1）若线性方程组的系数矩阵为 A，则 J 法收敛而 G-S 法不收敛；
（2）若线性方程组的系数矩阵为 B，则 J 法不收敛而 G-S 法收敛。

13. 设 $A = \begin{pmatrix} \alpha & 1 & 3 \\ 1 & \alpha & 2 \\ -3 & 2 & \alpha \end{pmatrix}$，$\alpha$ 为何值时，J 法收敛？$\alpha = 3$ 时怎样？

14. 用迭代公式 $\boldsymbol{x}^{(j+1)}=\boldsymbol{x}^{(j)}-\alpha(A\boldsymbol{x}^{(j)}-\boldsymbol{b})$，求解 $A\boldsymbol{x}=\boldsymbol{b}$.

(1) 若 $A=\begin{pmatrix}3 & 2\\ 1 & 2\end{pmatrix}$，问取什么范围的 α 可使迭代收敛，什么 α 使迭代收敛最快？

(2) 若 $A\in\boldsymbol{R}^{n\times n}$，且 A 对称正定，最大和最小特征值是 λ_1 和 λ_n，求迭代矩阵 $I-\alpha A$ 的谱半径，证明迭代收敛的充要条件是 $0<\alpha<2\lambda_1^{-1}$. 并问什么参数 α 使 $\rho(I-\alpha A)$ 最小？

15. 给定迭代式 $\boldsymbol{x}^{(j+1)}=B\boldsymbol{x}^{(j)}+\boldsymbol{f}$，其中 $B=\begin{pmatrix}\alpha & 4\\ 0 & \alpha\end{pmatrix}$，$0<\alpha<1$，证明在迭代过程中误差的范数 $\|\|_\infty$ 从某次迭代 j_0 开始单调减少（j_0 由充分接近 1 的 α 确定）。

16. 若 $\rho(B)=0$，试证对任意的 $\boldsymbol{x}^{(0)}$，迭代式 $\boldsymbol{x}^{(j+1)}=B\boldsymbol{x}^{(j)}+\boldsymbol{f}$ 最多 n 次迭代就可以得到精确解（提示:考虑 B 及 B^k 的 Jordan 标准形）。

17. 取 $\boldsymbol{x}^{(0)}=0$，用 CG 法求解

$$\begin{pmatrix}6 & 3 & 1\\ 3 & 5 & 1\\ 1 & 1 & 3\end{pmatrix}\begin{pmatrix}x_1\\ x_2\\ x_3\end{pmatrix}=\begin{pmatrix}0\\ -1\\ 2\end{pmatrix}.$$

第五章　最优化方法

优化问题常常出现在工程技术研究及应用中，因而学习最优化方法是理工科课程的重要内容。本章主要介绍线性规划及单纯形算法，非线性优化问题的最优性条件，最速下降法，共轭梯度法，罚函数法，组合优化问题及模拟退火算法和遗传算法。

5.1　线性规划与单纯形法

线性规划是目标函数与约束函数都为线性函数的一类重要的优化问题。它在工程和管理学科中有着广泛的应用，且对该问题的理论和算法研究度都已比较成熟。本小节将介绍线性规划的有关概念和基本性质，及求解线性规划的单纯形法。

5.1.1　线性规划标准型及最优基本可行解

线性规划问题可以写成如下标准型：

$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & Ax = b, \\ & x \geq 0, \end{aligned} \tag{5.1}$$

其中A是$m \times n$矩阵，且$m < n$, $rankA = m$，c为n维列向量，b为m维列向量且假设$b \geq 0$。线性规划问题的理论和算法基本都是基于标准型得到的，实际出现的任何线性规划问题都可以通过变换转为标准型。

(1) 如果原问题是极大化目标函数，即$\max c^T x$。这时只需将目标函数乘以(-1)，则可等价地将其转化为极小化问题, $\min -c^T x$。

(2) 如果右端项b中的第i个元素是负的, 则在第i个约束方程两边同乘以-1，将其改成右端项非负的不等式。

(3) 若约束方程为$\leq$不等式，则可在相应不等式的左端加上非负松弛变量，将原$\leq$不等式变为等式。若约束方程为$\geq$不等式，则可在相应不等式的左端减去非负剩余变量，将原$\geq$不等式变为等式。

(4) 如果某个变量x_i是无非负限制的自由变量，则总可以引入两个非负变量x_i', x_i''使得$x_i = x_i' - x_i''$。

例 5.1 将下面的线性规划问题转化为标准型。

$$\max \quad 2x_1 - 3x_2 + x_3 + 3x_4$$
$$s.t. \quad \begin{cases} 2x_1 - x_2 + 3x_3 + x_4 \geq 3 \\ 3x_1 + 2x_2 + 2x_4 = 7 \\ -x_1 + 4x_2 - 3x_3 - x_4 \leq 6 \\ x_1, x_3, x_4 \geq 0\,, \ x_2 \text{无约束} \end{cases}$$

解：令$x_2 = x_2' - x_2''$, 其中$x_2' \geq 0$, $x_2'' \geq 0$。引入剩余变量$x_5 \geq 0$，引入松弛变量$x_6 \geq 0$，则标准型为

$$\min \quad -2x_1 - 3x_2' + 3x_2'' - x_3 - 3x_4$$
$$s.t. \quad \begin{cases} 2x_1 - x_2' + x_2'' + 3x_3 + x_4 - x_5 = 3 \\ 3x_1 + 2x_2' - 2x_2'' + 2x_4 = 7 \\ -x_1 + 4x_2' - 4x_2'' - 3x_3 - x_4 + x_6 = 6 \\ x_1, x_3, x_4, x_5, x_6, x_2', x_2'' \geq 0 \end{cases}$$

下面我们给出关于线性规划解的相关概念和定理。

设B是矩阵A中$m \times m$阶非奇异子矩阵($|B| \neq 0$)，称矩阵B是线性规化问题的一个基。将A中剩余列向量组成$m \times (n-m)$矩阵，记为N。则A可写成分块矩阵$A = [B, N]$。记$x = \begin{pmatrix} x_B \\ x_N \end{pmatrix}$, 由$Bx_B + Nx_N = b$得

$$x_B = B^{-1}b - B^{-1}Nx_N,$$

再令$x_N = 0$，得到$Ax = b$ 的一个解 $x = \begin{pmatrix} B^{-1}b \\ 0 \end{pmatrix}$.

定义 5.1 (1) $x = \begin{pmatrix} B^{-1}b \\ 0 \end{pmatrix}$是$Ax = b$在基$B$下的基本解，向量$x_B$中的元素称为基变量，$B$中的列向量称为基本列向量。如果基本解中的某些基变量为0，这个基本解称为退化的基本解。

(2) 满足$Ax = b, x \geq 0$的向量x称为可行解。若可行解是基本解，则称之为基本可行解。若基本可行解是退化的基本解，则称之为退化的基本可行解。

定义 5.2 对于任何满足$Ax = b, x \geq 0$使得$c^T x$取得最小值的向量x称为最优可行解。如果最优可行解是基本解，则称为最优基本可行解。

定理 5.1 (1) 线性规划问题(5.1)如果存在可行解，则一定存在基本可行解。

(2) 线性规划问题(5.1)如果存在最优可行解，则一定存在最优基本可行解。

5.1.2 单纯形方法原理

由定理5.1可知，求解线性规划问题归结为寻找最优基本可行解，即从一个基本可行解出发，求得一个使目标函数减少的基本可行解，通过不断改进基本可行解，最终得到最优基本可行解。这就是单纯形方法的基本思想。下面我们分析单纯形法中实现基本可行解之间转换的过程。

最优性判别： 设基本可行解$x^{(k)}$为

$$x^{(k)}=\begin{pmatrix} x_B^{(k)} \\ 0 \end{pmatrix}=\begin{pmatrix} B_k^{-1}b \\ 0 \end{pmatrix}\geq 0$$

其中B_k为基矩阵，$x_B^{(k)}$为基变量。在$x^{(k)}$处的目标函数值为

$$\begin{aligned} f_k=f(x^{(k)}) &=\left(\left(c_B^{(k)}\right)^T,\ \left(c_N^{(k)}\right)^T\right)\begin{pmatrix} B_k^{-1}b \\ 0 \end{pmatrix} \\ &=\left(c_B^{(k)}\right)^T B_k^{-1}b \end{aligned} \tag{5.2}$$

现由$x^{(k)}$出发，得到一可行解x，

$$x=\begin{pmatrix} x_B \\ x_N \end{pmatrix},$$

其中

$$x_B=B_k^{-1}b-B_k^{-1}N_kx_N=x_B^{(k)}-B_kN_kx_N\geq 0,\qquad x_N\geq 0. \tag{5.3}$$

则在x处的目标函数值

$$\begin{aligned} f(x) &=\left(\left(c_B^{(k)}\right)^T\ \left(c_N^{(k)}\right)^T\right)\begin{pmatrix} x_B \\ x_N \end{pmatrix} \\ &=\left(c_B^{(k)}\right)^T\left(B_k^{-1}b-B_k^{-1}N_kx_N\right)+\left(c_N^{(k)}\right)^Tx_N \\ &=\left(c_B^{(k)}\right)^TB_k^{-1}b-\left(\left(c_B^{(k)}\right)^TB_k^{-1}N_k-\left(c_N^{(k)}\right)^T\right)x_N \end{aligned} \tag{5.4}$$

令

$$w_k=B_k^{-T}c_B^{(k)},\ \ z_N^{(k)}=N_k^TB_k^{-T}c_B^{(k)}=N_k^Tw_k \tag{5.5}$$

并称w_k为单纯形乘子。 由(5.4)和(5.2)得：

$$\begin{aligned} f(x) &=f_k-\left(z_N^{(k)}-c_N^{(k)}\right)^Tx_N \\ &=f_k-\sum_{j\in\mathcal{N}_k}(z_j^{(k)}-c_j^{(k)})x_j, \end{aligned} \tag{5.6}$$

其中$\mathcal{N}_k$为非基变量指标集。为检验可行点$x^{(k)}$是否是问题的最优解, 我们分析(5.6), 即根据可行性的要求我们观察当非基变量由零向正方向变动时目标函数值的变化。

(1) 若对所有$j \in \mathcal{N}_k$都有$z_j^{(k)} - c_j^{(k)} \leq 0$,则增大任何非基变量使得从$x^{(k)}$ 向可行域内的移动只能导致目标函数值的增加， 因此$x^{(k)}$ 就是问题的最优解。

(2) 如果有某个$z_j^{(k)} - c_j^{(k)} > 0,\ j \in \mathcal{N}_k$,则通过增大相应的非基变量$x_j$的值可使目标函数值减小, 因而$x^{(k)}$不是问题的最优解。

在线性规划中，通常称$z_j - c_j$为判别数或检验数。且通过以上分析，我们有如下结论

定理 5.2 *对于某个基本可行解，如果所有$z_j - c_j \leq 0$，则这个基本可行解为最优解。*

寻找新的基本可行解： 如果$x^{(k)}$ 不是最优的基本可行解, 则需要确定一个使目标函数值减少的基本可行解， 即确定一组新的基矩阵和非基矩阵。 单纯形方法中新的基矩阵是通过将原非基矩阵的一列(对应入基变量)与原基矩阵的一列(对应出基变量) 交换所得。具体分析如下：

若$z_p^{(k)} - c_p^{(k)} > 0,\ p \in \mathcal{N}_k$, 则增大$x_p$可以使得目标函数值下降，同时可取$x_p(p \in \mathcal{N}_k)$为入基变量, 即把矩阵$A$中相应于变量$x_p$的列$a_p$从非基矩阵$N_k$移入基矩阵$B_k$。为完成此步骤，我们需确定$x_p$的取值。设当$x_p$由 0增加为正值时，$x_N$的其余分量保持0值，由(5.3)得

$$\begin{aligned} x_B &= B_k^{-1}b - B_k^{-1}a_p x_p = \bar{b}^{(k)} - y_p^{(k)} x_p \\ &= \begin{pmatrix} \bar{b}_1^{(k)} \\ \bar{b}_2^{(k)} \\ \vdots \\ \bar{b}_m^{(k)} \end{pmatrix} - \begin{pmatrix} y_{1p}^{(k)} \\ y_{2p}^{(k)} \\ \vdots \\ y_{mp}^{(k)} \end{pmatrix} x_p \end{aligned}$$

其中$\bar{b}^{(k)} = B_k^{-1}b,\ y_p^{(k)} = B_k^{-1}a_p$。 为保持可行性，需要

$$(x_B)_i = \bar{b}_i^{(k)} - y_{ip}^{(k)} x_p \geq 0$$

(1) 如果$y_{ip}^{(k)} \leq 0$，当x_p由 0增加为正值时，可行性总成立。

(2) 如果$y_{ip}^{(k)} > 0$， 当x_p由 0增加为正值时，$(x_B)_i$会减小，则为保证可行性，就必须使

$$x_p \leq \frac{\bar{b}_i^{(k)}}{y_{ip}^{(k)}}$$

因此，为使$x_B \geq 0$, 令

$$x_p = \min\{\frac{\overline{b}_i^{(k)}}{y_{ip}^{(k)}} \mid y_{ip}^{(k)} > 0\} = \frac{\overline{b}_r^{(k)}}{y_{rp}^{(k)}}$$

同时注意到，

$$(x_B)_r = \overline{b}_r^{(k)} - y_{rp}^{(k)} x_p = 0,$$

即基变量对应分量数值变为0.

记矩阵A对应于$(x_B)_r$的列为a_r,将B_k中的列a_r 与N_k中的列a_p交换，就得到新的基矩阵和非基矩阵。同时可以确定新的基本可行解及下降的目标函数值，

$$\begin{cases} x_p^{(k+1)} = \min\{\frac{\overline{b}_i^{(k)}}{y_{ip}^{(k)}} \mid y_{ip}^{(k)} > 0\} = \frac{\overline{b}_r^{(k)}}{y_{rp}^{(k)}} \\ x_j^{(k+1)} = 0, j \neq p, j \in \mathcal{N}_k \\ (x_B)_r^{(k+1)} = 0 \\ (x_B)_i^{(k+1)} = (x_B)_i^{(k)} - x_p^{(k+1)} y_{ip}^{(k)}, i \notin r \end{cases}$$

$$f_{k+1} = f_k - (z_p^{(k)} - c_p^{(k)}) x_p^{(k+1)}$$

对$x^{(k+1)}$重复上述两个过程确定问题的一个最优解或者问题无界的最优解。

算法 5.1 (单纯形方法)

1: 给定一个初始基本可行解，设初始基为$B_1, k = 1$;

2: 求解$x_B^{(k)} = B_k^{-1} b$, 令$x_N^{(k)} = 0$, 计算目标函数值$f = (c_B^{(k)})^T x_B^{(k)}$.

3: 计算单纯形乘子$w_k = B_k^{-T} c_B^{(k)}$，对于所有非基变量，计算判别数

$$z_p^{(k)} - c_p^{(k)} = \max_{j \in N_k} \{z_j^{(k)} - c_j^{(k)}\}$$

若$z_p^{(k)} - c_p^{(k)} \leq 0$，停止计算，现行基本可行解是最优解。否则，$x_p$为入基变量，进行步骤 4。

4: 解$y_p^{(k)} = B_k^{-1} a_p$，若$y_p^{(k)} \leq 0$，停止计算，问题有无界最优解。否则确定出基变量指标

$$\frac{\overline{b}_r^{(k)}}{y_{rp}^{(k)}} = \min\{\frac{\overline{b}_i^{(k)}}{y_{ip}^{(k)}} \mid y_{ip}^{(k)} > 0\}$$

$(x_B)_r$为出基变量。用B_k中的列a_r 与N_k中的列a_p交换，得到新的基矩阵B_{k+1} 和非基矩阵 N_{k+1}，令$k = k + 1$，转步骤 2。

5.1.3 单纯形表格法

上述求解过程可以通过单纯形表来实现。我们首先分析如何构造单纯形表。 将线性规划问题(5.1)等价表示为

$$\begin{aligned}
&\min \quad f \\
&\text{s.t.} \quad x_B \ +B^{-1}Nx_N \ = B^{-1}b \\
&\qquad f \ +0\cdot x_B \ +\left(c_B^TB^{-1}N-c_N^T\right)x_N \ = c_B^TB^{-1}b \\
&\qquad x_B\geq 0,\ \ x_N\geq 0 \quad .
\end{aligned}$$

由上述约束方程系数得到如下 单纯形表, 其包含了单纯形方法所需要的所有数据。

	x_B	x_N	右端项
x_B	I_m	$B^{-1}N$	$B^{-1}b$
f	0	$c_B^TB^{-1}N-c_N^T$	$c_B^TB^{-1}b$

下面我们讨论如何用单纯形表求解标准线性规划问题。由于单纯形表中已经包含了m阶单位矩阵，因此已经给出了一个基本可行解。

(1) 若$c_B^TB^{-1}N-c_N^T\leq 0$, 则现行基本可行解是最优解。

$$x_b=B^{-1}b,\ x_N=0,\ f=c_B^TB^{-1}b.$$

(2) 若存在指标使得$c_B^TB^{-1}N-c_N^T>0$，我们需要进行 主元消去法求解改进的基本可行解。 由

$$z_p-c_p=\max_{j\in N_k}\{z_j-c_j\}$$

决定x_p为进基变量，它所对应的列为 主列。再由

$$\frac{\bar{b}_r}{y_{rp}}=\min\{\frac{\bar{b}_i}{y_{ip}}\mid y_{ip}>0\}$$

决定第r行为 主行。主行和主列交叉的元素y_{rp}为 主元。然后进行主元消去法，使得主列化为单位列向量。

一般地，我们通过从原始变量、引入的松弛变量或人工变量(在下一节中讨论)选取初始基变量，使得相应初始基矩阵为单位阵，然后计算初始的判别数得到初始单纯形表。 进而由上述步骤进行相应计算。

例 5.2 用单纯形方法求解下列线性规划问题:

$$\begin{aligned}
&\min \quad x_1-3x_2-x_3 \\
&s.t. \quad 3x_1-x_2+2x_3\leq 7 \\
&\qquad -2x_1+4x_2\leq 12 \\
&\qquad -4x_1+3x_2+8x_3\leq 10 \\
&\qquad x_1,x_2,x_3\geq 0.
\end{aligned}$$

解：引入松弛变量x_4, x_5, x_6，转化为标准型：

$$\begin{aligned} \min \quad & x_1 - 3x_2 - x_3 \\ s.t. \quad & 3x_1 - x_2 + 2x_3 + x_4 = 7 \\ & -2x_1 + 4x_2 + x_5 = 12 \\ & -4x_1 + 3x_2 + 8x_3 + x_7 = 10 \\ & x_1, x_2, x_3 x_4, x_5, x_6 \geq 0. \end{aligned}$$

由标准型，可选取x_4, x_4, x_5为基变量，建立单纯形表：

	x_1	x_2	x_3	x_4	x_5	x_6	
x_4	3	-1	2	1	0	0	7
x_5	-2	$\boxed{4}$	0	0	1	0	12
x_6	-4	3	8	0	0	1	10
	-1	3	1	0	0	0	0

由于$z_2 - c_2 = \max\{z_i - c_i\} = 3$，因此取第2列作为主列，$x_2$为进基变量。由于$\frac{\bar{b}_2}{y_{22}} = \min\{\frac{12}{4}, \frac{10}{3}\} = 3$, 因此取第2行为主行，且$x_5$为出基变量。主元为$y_{22} = 4$，进行主元消去，将变量$x_2$对应的列变换为单位列向量。新的基变量为$x_4, x_2, x_6$。得到下表。

	x_1	x_2	x_3	x_4	x_5	x_6	
x_4	$\frac{5}{2}$	0	2	1	$\frac{1}{4}$	0	10
x_2	$-\frac{1}{2}$	1	0	0	$\frac{1}{4}$	0	3
x_6	$-\frac{5}{2}$	0	$\boxed{8}$	0	$-\frac{3}{4}$	1	1
	$\frac{1}{2}$	0	1	0	$-\frac{3}{4}$	0	-9

由于$z_3 - c_3 = \max\{z_i - c_i\} = 1$，因此取第3列作为主列，$x_3$为进基变量。由于$\frac{\bar{b}_3}{y_{33}} = \min\{\frac{10}{2}, \frac{1}{8}\} = \frac{1}{8}$, 因此取第3行为主行，且$x_6$为出基变量。主元为$y_{33} = 8$，进行主元消去，将变量$x_3$对应的列变换为单位列向量。新的基变量为$x_4, x_2, x_3$。得到下表。

	x_1	x_2	x_3	x_4	x_5	x_6	
x_4	$\boxed{\frac{8}{25}}$	0	0	1	$\frac{7}{16}$	$-\frac{1}{4}$	$\frac{39}{4}$
x_2	$-\frac{1}{2}$	1	0	0	$\frac{1}{4}$	0	3
x_3	$-\frac{5}{16}$	0	1	0	$-\frac{3}{32}$	$\frac{1}{8}$	$\frac{1}{8}$
	$\frac{13}{16}$	0	0	0	$-\frac{21}{32}$	$-\frac{1}{8}$	$-\frac{73}{8}$

由于$z_1 - c_1 = \max\{z_i - c_i\} = \frac{13}{16}$，因此取第1列作为主列，$x_1$为进基变量。由于$\frac{\bar{b}_1}{y_{11}} = \min\{\frac{39}{4}/\frac{8}{25}\} = \frac{78}{25}$, 因此取第1行为主行，且$x_4$为出基变量。主元为$y_{11} = \frac{8}{25}$，进行主元消去，将变量$x_1$ 对应的列变换为单位列向量。新的基变量为x_1, x_2, x_3。得到下表。

	x_1	x_2	x_3	x_4	x_5	x_6	
x_1	1	0	0	$\frac{8}{25}$	$\frac{7}{50}$	$-\frac{2}{25}$	$\frac{78}{25}$
x_2	0	1	0	$\frac{4}{25}$	$\frac{8}{25}$	$-\frac{1}{25}$	$\frac{114}{25}$
x_3	0	0	1	$\frac{1}{10}$	$-\frac{1}{20}$	$\frac{1}{10}$	$\frac{11}{10}$
	0	0	0	$-\frac{13}{50}$	$-\frac{77}{100}$	$-\frac{3}{50}$	$-\frac{583}{50}$

由于所有$z_i - c_i \leq 0$, 因此达到最优值，由单纯形表可知，所得到的最优值是

$$(x_1, x_2, x_3) = (\frac{78}{25}, \frac{114}{25}, \frac{11}{10}),$$

目标函数最优值

$$f_{\min} = -\frac{583}{50}.$$

5.1.4 两阶段法和大M法

单纯形方法的迭代开始需要一个基本可行解, 由上面例题可以发现，对于约束形式为$Ax \leq b$ 转化成标准型的线性规划问题, 初始基变量可以取为所有的松弛变量，即取一个初始基矩阵为单位阵的基本可行解。但是对于含等式约束$a_i^T x = b_i$和形如$a_i^T x \geq b_i$ 的不等式约束的问题, 通常难以确定一个初始基本可行解。例如，

$$\begin{aligned} \min \quad & 3x_1 - 2x_2 + x_3 \\ s.t. \quad & 2x_1 + x_2 - x_3 \leq 5 \\ & 4x_1 + 3x_2 + x_3 \geq 3 \\ & -x_1 + x_2 + x_3 = 2 \\ & x_1, x_2, x_3 \geq 0. \end{aligned}$$

对第一个约束引入松弛变量x_4, 第二个引入剩余变量x_5, 转化为标准型：

$$\begin{aligned} \min \quad & 3x_1 - 2x_2 + x_3 \\ s.t. \quad & 2x_1 + x_2 - x_3 + x_4 = 5 \\ & 4x_1 + 3x_2 + x_3 - x_5 = 3 \\ & -x_1 + x_2 + x_3 = 2 \\ & x_i \geq 0, \ (i = 1, \cdots 5). \end{aligned} \tag{5.7}$$

从标准型可以发现除x_4可以作为一个基变量，我们难以确定其他两个基变量。

对于这样难以确定初始基本可行解的线性规划问题，我们通常是引入人工变量将问题变换为具有明确初始基本可行解的线性规划问题，再用两阶段方法或大M 方法求解。

两阶段方法： 对于问题(5.1)，设A中不包含m阶单位阵，为使约束方程的系数矩阵中含有m 阶单位阵，引入人工变量x_α，则(5.1)约束变换为

$$\begin{aligned} &Ax + x_\alpha = b, \\ &x \geq 0, x_\alpha \geq 0. \end{aligned} \tag{5.8}$$

构造第一阶段问题：

$$\begin{aligned} \min \quad & e^T x_\alpha \\ s.t. \quad & Ax + x_\alpha = b, \\ & x \geq 0, x_\alpha \geq 0 \end{aligned} \tag{5.9}$$

其中$e = (1, 1, \cdots, 1)^T$。用单纯形方法求解上述问题(5.9)，将人工变量变换成非基变量，求出原问题(5.1) 的一个基本可行解。即由 $\begin{pmatrix} x \\ x_\alpha \end{pmatrix} = \begin{pmatrix} 0 \\ b \end{pmatrix}$ 出发获得问题(5.1) 的最优基本解 $\begin{pmatrix} x \\ x_\alpha \end{pmatrix} = \begin{pmatrix} \overline{x} \\ 0 \end{pmatrix}$，则$x = \overline{x}$为原问题 (5.1) 的一个基本可行解。

两阶段方法的第二阶段： 从第一阶段的最优单纯形表中去掉人工变量对应的列，再由第一阶段获得的基本可行解出发， 按标准型的目标函数系数值修正最后一行的判别数及目标函数值， 用单纯形方法求解新的单纯形表获得原线性规划问题(5.1) 的最优解。

例 5.3 *用两阶段法求解如下问题的最优解：*

$$\begin{aligned} \min \quad & 3x_1 - 2x_2 + x_3 \\ s.t. \quad & 2x_1 + x_2 - x_3 \leq 5 \\ & 4x_1 + 3x_2 + x_3 \geq 3 \\ & -x_1 + x_2 + x_3 = 2 \\ & x_1, x_2, x_3 \geq 0. \end{aligned}$$

解： 第一阶段: 对第二个约束和第三个约束引入人工变量x_6, x_7，由(5.7) 的约束可得第一阶段问题

$$\begin{aligned} \min \quad & x_6 + x_7 \\ s.t. \quad & 2x_1 + x_2 - x_3 + x_4 = 5 \\ & 4x_1 + 3x_2 + x_3 - x_5 + x_6 = 3 \\ & -x_1 + x_2 + x_3 + x_7 = 2 \\ & x_i \geq 0, \ (i = 1, \cdots 7). \end{aligned} \tag{5.10}$$

用单纯形法求解上述标准型：

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	
x_4	2	1	-1	1	0	0	0	5
x_6	4	$\boxed{3}$	1	0	-1	1	0	3
x_7	-1	1	1	0	0	0	1	2
	3	4	2	0	-1	0	0	5

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	
x_4	$\frac{2}{3}$	0	$-\frac{4}{3}$	1	$\frac{1}{3}$	$-\frac{1}{3}$	0	4
x_2	$\frac{4}{3}$	1	$\frac{1}{3}$	0	$-\frac{1}{3}$	$\frac{1}{3}$	0	1
x_7	$-\frac{7}{3}$	0	$\boxed{\frac{2}{3}}$	0	$\frac{1}{3}$	$-\frac{1}{3}$	1	1
	$-\frac{7}{3}$	0	$\frac{2}{3}$	0	$\frac{1}{3}$	$-\frac{4}{3}$	0	1

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	
x_4	-4	0	0	1	1	-1	2	6
x_2	$\frac{5}{2}$	1	0	0	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$
x_3	$-\frac{7}{2}$	0	1	0	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{3}{2}$	$\frac{3}{2}$
	0	0	0	0	0	0	-1	0

由于所有的判别数都小于0, 因此第一阶段问题达到最优解，且人工变量x_6, x_7都是非基变量。于是我们得到(5.7)的初始基本可行解 $(x_1, x_2, x_3, x_4, x_5) = (0, \frac{1}{2}, \frac{3}{2}, 6, 0)$。

第一阶段结束后，去掉人工变量x_6, x_7及对应的列， 以x_2, x_3, x_4为基变量，按(5.7) 的目标函数系数值修正最后一行的判别数及目标函数值，得到第二阶段的单纯形表。迭代过程如下：

	x_1	x_2	x_3	x_4	x_5	
x_4	-4	0	0	1	1	6
x_2	$\frac{5}{2}$	1	0	0	$-\frac{1}{2}$	$\frac{1}{2}$
x_3	$-\frac{7}{2}$	0	1	0	$\boxed{\frac{1}{2}}$	$\frac{3}{2}$
	$-\frac{23}{2}$	0	0	0	$\frac{3}{2}$	$\frac{1}{2}$

	x_1	x_2	x_3	x_4	x_5	
x_4	3	0	-2	1	0	3
x_2	-1	1	1	0	0	2
x_5	-7	0	2	0	1	3
	-1	0	-3	0	0	-4

得到原线性规划的最优解$(x_1, x_2, x_3) = (0, 2, 0)$，最优目标函数值为$f_{\min} = -4$。

大M法： 方法的基本思想是在约束中增加人工变量x_α， 同时在标准形目标函数中加入罚项$Me^T x_\alpha$，其中M是很大的正数。即构造问题：

$$\begin{aligned} \min \quad & c^T x + Me^T x_\alpha \\ s.t. \quad & Ax + x_\alpha = b, \\ & x \geq 0, x_\alpha \geq 0. \end{aligned}$$

进而用单纯形方法求解辅助问题。迫使人工变量离基。

对于例5.3, 我们也可用大M法求解。 构造辅助问题：

$$\begin{aligned} \min \quad & 3x_1 - 2x_2 + x_3 + Mx_6 + Mx_7 \\ s.t. \quad & 2x_1 + x_2 - x_3 + x_4 = 5 \\ & 4x_1 + 3x_2 + x_3 - x_5 + x_6 = 3 \\ & -x_1 + x_2 + x_3 + x_7 = 2 \\ & x_i \geq 0, \ (i = 1, \cdots 7). \end{aligned}$$

具体求解过程留作习题，并观察结果是否与两阶段法一致。

退化情形

在单纯形法确定出基变量时，若存在两个以上相同的比值，就会出现退化解，即出现计算过程循环。虽然这样的情形极少出现，但还是有可能存在。我们可以用摄动法、 字典排序法及勃兰特规则等方法来处理这样的情形。

5.2 非线性规划的最优性条件

最优性条件是指优化问题的最优解所必须满足的条件。最优性条件不仅对于最优化理论的研究具有重要意义, 而且对最优化算法的设计和终止条件的确定起重要作用。我们在本节将介绍最优性的一阶条件和二阶条件。

5.2.1 无约束规划问题的最优性条件

考虑无约束最优化问题：

$$\min_{x \in R^n} f(x) \tag{5.11}$$

其中$f(x)$为R^n的实值函数。

定义 5.3 若存在$x^* \in R^n$使得$f(x) \geq f(x^*)$, $\forall x \in R^n$，则称x^*为问题(5.11)的全局极小点。如果有 $f(x) > f(x^*)$, $\forall x \neq x^* \in R^n$，则称 x^* 为问题(5.11)的严格全局极小点。

定义 5.4 若存在$x^* \in R^n$使得$f(x) \geq f(x^*)$, $\forall x \in N_\delta(x^*)$，则称$x^*$为问题(5.11) 的局部极小点。如果有$f(x) > f(x^*)$, $\forall x \in N_\delta(x^*)/\{x^*\}$，则称$x^*$为问题(5.11)的严格局部极小点。

求全局极小点一般来说相当困难，实际可行的只是求一个局部(或严格局部)极小点。故后面所指极小点通常指局部极小点。下面我们讨论局部极小点的条件。

定理 5.3 (一阶必要条件) 若设 x^* 处是无约束优化问题 (5.11) 的局部极小点，且 $f(x)$ 在 x^* 处连续可微，则 $\nabla f(x^*) = 0$。

证明：用反证法。假设$\nabla f(x^*) \neq 0$。令向量$p = -\nabla f(x^*)$，则有

$$p^T \nabla f(x^*) = -\|\nabla f(x^*)\|^2 < 0.$$

因为$\nabla f(x)$ 在x^*连续，则存在$T > 0$，对所有$t \in [0, T]$有

$$p^T \nabla f(x^* + tp) < 0.$$

又由泰勒定理，对任意$\bar{t} \in (0, T]$，存在$t \in (0, \bar{t})$，我们有

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}p^T \nabla f(x^* + tp) < f(x^*).$$

即由x^*沿方向p使得函数值下降，与x^*是极小值点矛盾。证毕。

满足$\nabla f(x^*) = 0$的点称为驻点，由必要条件可知，极小值点必为驻点，但反之不正确。下面我们给出保证x^*为局部极小点的条件。

定理 5.4 (二阶充分条件)设$f(x)$在x^*处二阶连续可微，如果 $\nabla f(x^*) = 0$且Hesse矩阵$\nabla^2 f(x^*)$正定，则x^*是无约束优化问题(5.11)的严格局部极小点。

证明：由于$\nabla^2 f(x)$在x^*处连续且正定，我们可取$r > 0$使得$\nabla^2 f(x)$在区域$D = \{z | \|z - x^*\| < r\}$内保持正定性。令 $p \neq 0$且$\|p\| \leq r$，则有$x^* + p \in D$，且由$\nabla f(x^*) = 0$，我们有

$$f(x^* + p) = f(x^*) + \frac{1}{2} p^T \nabla^2 f(x^* + tp) p \quad t \in (0, 1).$$

注意到$x^* + tp \in D$，有$p^T \nabla^2 f(x^* + tp) p > 0$, 则$f(x^* + p) > f(x^*)$。证毕。

上述二阶充分条件并非必要的，如$x^* = 0$为函数$f(x) = x^4$的严格局部极小点，但该点处的Hessian值为0。

例 5.4 利用极值条件求解如下问题：

$$\min \quad f(x)=\frac{1}{3}x_2^3+\frac{1}{2}x_1^2-x_2^2-x_1.$$

解： 由于

$$\frac{\partial f}{\partial x_1}=x_1-1, \quad \frac{\partial f}{\partial x_2}={x_2}^2-2x_2$$

令$\nabla f(x)=0$得驻点：

$$x^{(1)}=\begin{pmatrix}1\\0\end{pmatrix}, \quad x^{(2)}=\begin{pmatrix}1\\2\end{pmatrix},$$

函数$f(x)$的Hesse矩阵：

$$\nabla^2 f(x)=\begin{pmatrix}1 & 0\\0 & 2x_2-2\end{pmatrix}$$

由此，在各点的Hesse矩阵依次是

$$\nabla^2 f(x^{(1)})=\begin{pmatrix}1 & 0\\0 & -2\end{pmatrix}, \quad \nabla^2 f(x^{(2)})=\begin{pmatrix}1 & 0\\0 & 2\end{pmatrix}$$

由于$\nabla^2 f(x^{(1)})$不定，$x^{(1)}$不是极小值点， $\nabla^2 f(x^{(2)})$ 为正定阵，所以$x^{(2)}$为局部极小值点。

定理5.3及定理5.4分别刻画了无约束优化问题的必要及充分条件，但这些条件只能研究局部极小点。如果目标函数为凸函数，我们能够给出全局极小点的充要条件。

定理 5.5 设$f(x)$为可微凸函数，则x^*是无约束优化问题(5.11)全局极小点的充分必要条件为 $\nabla f(x^*)=0$。

证明：必要性： 若x^*是无约束优化问题(5.11)全局极小点，则其必为局部极小点，由定理5.3可知$\nabla f(x^*)=0$。

充分性：由于$\nabla f(x^*)=0$，及$f(x)$为可微凸函数，则对任意$x\in R^n$有

$$f(x)\geq f(x^*)+\nabla f(x^*)(x-x^*)=f(x^*).$$

即x^*是全局极小点。 证毕

5.2.2 带约束规划问题的最优性条件

我们考虑如下约束优化问题：

$$\begin{aligned}\min \quad & f(x)\\ s.t. \quad & h_i(x)=0, i\in\mathcal{E}\\ & g_j(x)\geq 0, j\in\mathcal{I},\end{aligned} \tag{5.12}$$

其中$\mathcal{E}=\{1,2,..,m\}$， $\mathcal{I}=\{1,2,...,l\}$。记可行域为

$$\mathcal{F}=\{h_i(x)=0,\ i\in\mathcal{E};\ g_j(x)\geq 0,\ j\in\mathcal{I}\}. \tag{5.13}$$

记指标集$\mathcal{I}(x)=\{j|g_j(x)=0,\ j\in\mathcal{I}\}$， 则$\mathcal{E}\cup\mathcal{I}(x)$表示为在当前可行点处的积极约束集指标集(起作用的约束指标集)。

定义 5.5 设$x^*\in\mathcal{F}$, 如果

$$f(x^*)\leq f(x),\ \forall x\in\mathcal{F},$$

称x^*为问题(5.12)的全局极小点；若

$$f(x^*)<f(x),\ \forall x\in\mathcal{F},\ x\neq x^*,$$

称x^*为问题(5.12)的严格全局极小点。

定义 5.6 设$x^*\in\mathcal{F}$, 如果

$$f(x^*)\leq f(x),\ \forall x\in\mathcal{F}\cap N_\delta(x^*),$$

称x^*为问题(5.12)的局部极小点，其中$N_\delta(x^*)$为以x^*为中心，δ为半径的邻域；若

$$f(x^*)<f(x),\ \forall x\in\mathcal{F}\cap N_\delta(x^*),\ x\neq x^*,$$

称x^*为问题(5.12)的严格局部极小点。

同无约束优化问题一样，对于约束优化问题实际可行的只是求得一个局部(严格局部)极小值点。下面我们给出相应的必要和充分条件。

定理 5.6 (K-K-T必要条件) 设$f(x)$, $h_i(x)(i\in\mathcal{E})$, $g_j(x)(j\in\mathcal{I})$在可行点x^*处一阶连续可微，向量集

$$\{\nabla h_i(x),\nabla g_j(x)|i\in\mathcal{E},j\in\mathcal{I}(x^*)\}$$

线性独立。如果x^*是局部最优解，则存在 w_i $(i\in\mathcal{E})$, v_j $(j\in\mathcal{I})$ 使得

$$\begin{cases}\nabla f(x^*)-\sum\limits_{i=1}^{m}w_i\nabla h_i(x^*)-\sum\limits_{j=1}^{l}v_j\nabla g_j(x^*)=0\\ v_ig_i(x^*)=0\\ v_i\geq 0.\end{cases} \tag{5.14}$$

定义 Lagrange 函数

$$L(x,w,v)=f(x)-\sum_{i=1}^{m}w_ih_i(x)-\sum_{j=1}^{l}v_jg_j(x).$$

则(5.14)中第一式可表示为$\nabla_x L(x^*, w, v) = 0$，因此$w, v$也称为 Lagrange 乘子。由定理5.6可知，求解K-K-T点需求解下列系统：

$$\nabla_x L(x, w, v) = 0 \tag{5.15}$$
$$h_i(x) = 0,\ i \in \mathcal{E} \tag{5.16}$$
$$g_j(x) \geq 0,\ j \in \mathcal{I} \tag{5.17}$$
$$v_j g_j(x) = 0,\ j \in \mathcal{I} \tag{5.18}$$
$$v_j \geq 0,\ j \in \mathcal{I}. \tag{5.19}$$

(5.15)称为梯度条件，(5.16)-(5.17)称为可行条件，(5.18)-(5.19)称为互补松弛条件。

此外，定理中的线性独立条件必不可少，若没有该条件，局部极小点不一定是K-K-T点。

上面讨论的是一阶必要条件，下面讨论充分条件。

定理 5.7 (二阶充分条件) 设 x^* 是优化问题 (5.12) 满足 (5.15) – (5.19) 的K-K-T点，$w_i, v_i \geq 0$ 为对应Lagrange乘子。如果

$$d^T \nabla_{xx}{}^2 L(x^*, w, v) d > 0, \quad \forall 0 \neq d \in G(x^*, w, v), \tag{5.20}$$

其中

$$G(x^*, w, v) = \left\{ d \left| \begin{array}{l} \nabla h_i(x^*)^T d = 0, i \in \mathcal{E} \\ \nabla g_j(x^*)^T d = 0, j \in \mathcal{I}(x^*) \text{且} w_i > 0 \\ \nabla g_j(x^*)^T d \geq 0, j \in \mathcal{I}(x^*) \text{且} w_i = 0 \end{array} \right. \right\}, \tag{5.21}$$

则x^* 是(5.12)的严格局部极小点。

例 5.5 求下列约束优化问题的局部极小解

$$\begin{array}{ll} \min & (x_1 - 1)^2 + x_2 \\ s.t. & -x_1 - x_2 + 2 \geq 0 \\ & x_2 \geq 0. \end{array}$$

解：先求K-K-T点。

$$\nabla f(x) = \begin{pmatrix} 2(x_1 - 1) \\ 1 \end{pmatrix}, \quad \nabla g_1(x) = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \nabla g_2(x) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

由条件(5.15)-(5.19)得：

$$
\begin{gathered}
2(x_1-1)+v_1=0, \\
1+v_1-v_2=0, \\
v_1(-x_1-x_2+2)=0, \\
v_2x_2=0, \\
v_1,v_2\geq 0, \\
-x_1-x_2+2\geq 0, \\
x_2\geq 0.
\end{gathered}
$$

由情形$v_1=0,v_2=0$，$v_1\neq 0,v_2\neq 0$及$v_1\neq 0,v_2=0$所得结果皆不符合要求，舍去。当$v_1=0,v_2\neq 0$时，可得$x_1=1,x_2=0,v_2=1>0$。所以K-K-T点为$x=(1,0)^T$。

下面检验二阶充分条件。

$$
\nabla^2_{xx}L(x,v)=\begin{pmatrix}2&0\\0&0\end{pmatrix}.
$$

由于在点$x=(1,0)^T$积极约束为$g_2(x)=x_2$，所以由(5.21)得

$$
(0,1)\begin{pmatrix}d_1\\d_2\end{pmatrix}=0,
$$

所以有$d=(d_1,0)^T$。再由(5.20), 对于$d_1\neq 0$有

$$
d^T\nabla^2_{xx}L(x,v)\,d=2d_1^2>0
$$

则点$x=(1,0)^T$为局部极小值点，目标函数最优值为$f_{min}=0$。

特别地，对于凸规划我们有如下结果。

定理 5.8 *设在问题(5.12)中，$f(x)$为凸函数，$h_i(x)$ $(i\in\mathcal{E})$为线性函数，$g_j(x)$ $(j\in\mathcal{I})$为凹函数。如果x^*为K-K-T点，则x^*为问题(5.12)的全局极小点。*

5.3 无约束非线性优化算法

随着优化理论和优化技术的发展，已有大量的最优化方法适用于不同问题的求解，从本节开始，我们将学习一些经典的优化方法。

首先，我们介绍迭代算法的一般性结构。设非线性规划问题

$$
\begin{aligned}
\min\quad & f(x)\\
s.t.\quad & x\in\mathcal{F},
\end{aligned}
$$

其中$\mathcal{F} \subseteq R^n$。若$\mathcal{F} = R^n$，则问题为无约束优化问题；如$\mathcal{F}$为(5.13)，则问题即为(5.12)形式约束优化问题。

设$x^{(k)}$为迭代算法的第k次迭代点，以如下形式进行一次迭代

$$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)},$$

我们就得到第$k+1$次的迭代点，同时要求保证$f(x^{(k+1)}) \leq f(x^{(k)})$且迭代点保持可行。这就是求解得非线性最优化问题的基本迭代格式和要求。第4章极小化方法中所提及的迭代格式即为无约束情形下的表示。通常，$d^{(k)} \in R^n$被称为搜索方向，要求沿该方向目标函数能得到下降；$\alpha_k \in R$称为步长因子或沿方向$d^{(k)}$的步长。由迭代格式可以看出，求解非线性优化问题的关键即为获得步长因子α和下降方向d。

5.3.1 线性搜索

选取步长因子α的问题称为线性搜索问题，它是一个一维搜索问题，其基本要求在于从$x^{(k)}$出发沿给定的搜索方向$d^{(k)}$确定α_k使得

$$f(x^{(k)} + \alpha_k d^{(k)}) < f(x^{(k)}). \tag{5.22}$$

对于这样的问题，我们有两种解决途径：

(1) 精确线搜索。选取α_k使得

$$f\left(x^{(k)} + \alpha_k d^{(k)}\right) = \min_{\alpha \geq 0} f\left(x^{(k)} + \alpha d^{(k)}\right), \tag{5.23}$$

所得的α_k称为精确步长因子。一般地，我们可通过直接搜索或插值法来数值来求解(5.23)，如0.618法、Fibonacci法、割线法、三点二次插值法等。但是这种方法往往计算量很大，特别当迭代点远离最优解时，效率较低。

(2) 不精确线搜索。只要求选取α_k满足(5.22)使得目标函数每迭代一步都有充分下降即可，这样可以节省工作量。常用的方法有后退法、Armijo-Goldstein法及Wolfe-Powell法。

在本章节算法中，我们将采用精确线搜索获取步长因子。下面给出精确线搜索下目标函数值的下降量估计。

定理 5.9 *设$d^{(k)}$是下降方向，α_k是精确线搜索的步长因子。若存在常数$M > 0$使得所有$\alpha > 0$有$\|\nabla^2 f(x^{(k)} + \alpha d^{(k)})\| \leq M, \forall k$，则*

$$f(x^{(k)}) - f(x^{(k)} + \alpha_k d^{(k)}) \geq \frac{1}{2M} \|\nabla f(x^{(k)})\|^2 cos^2\theta_k,$$

其中θ_k为$d^{(k)}$与$-\nabla f(x^{(k)})$之间的夹角。

证明：由Taylor展开及假设条件可知对任意α有

$$\begin{aligned} f\left(x^{(k)}+\alpha d^{(k)}\right) &= f\left(x^{(k)}\right)+\alpha\nabla f\left(x^{(k)}\right)^T d^{(k)}+\tfrac{1}{2}\left(d^{(k)}\right)^T\nabla^2 f\left(x^{(k)}+\theta\alpha d^{(k)}\right)d^{(k)} \\ &\le f\left(x^{(k)}\right)+\alpha\nabla f\left(x^{(k)}\right)^T d^{(k)}+\tfrac{1}{2}\alpha^2 M\left\|d^{(k)}\right\|^2 \end{aligned}$$

取 $\overline{\alpha}=-\frac{\nabla f\left(x^{(k)}\right)^T d^{(k)}}{M\left\|d^{(k)}\right\|^2}$，因为$\alpha_k$是精确线搜索步长因子，则有

$$\begin{aligned} f\left(x^{(k)}\right)-f\left(x^{(k)}+\alpha_k d^{(k)}\right) &\ge f\left(x^{(k)}\right)-f\left(x^{(k)}+\overline{\alpha}d^{(k)}\right) \\ &\ge -\overline{\alpha}\nabla f\left(x^{(k)}\right)^T d^{(k)}-\tfrac{1}{2}\overline{\alpha}^2 M\left\|d^{(k)}\right\|^2 \\ &= \tfrac{1}{2}\frac{\left(\nabla f\left(x^{(k)}\right)^T d^{(k)}\right)^2}{M\left\|d^{(k)}\right\|^2} \\ &= \tfrac{1}{2M}\left\|\nabla f\left(x^{(k)}\right)\right\|^2\frac{\left(\nabla f\left(x^{(k)}\right)^T d^{(k)}\right)^2}{\left\|\nabla f\left(x^{(k)}\right)\right\|^2\left\|d^{(k)}\right\|^2} \\ &= \tfrac{1}{2M}\left\|\nabla f\left(x^{(k)}\right)\right\|^2\cos^2\theta_k. \end{aligned}$$

证毕。

不同的搜索方向将形成不同的最优化方法，本节将继续将对无约束优化问题(5.11)的搜索方向的选取，即各类经典算法的相关知识进行详细介绍。

5.3.2 最速下降法

最速下降法是无约束最优化最简单的方法，由法国科学家Cauchy于1847年提出，继而由Curry等人进一步研究得到的方法。

设目标函数$f(x)$在$x^{(k)}$附近连续可微，令$\nabla f(x^{(k)})\neq 0$。将$f(x)$在$x^{(k)}$处Taylor展开

$$f(x)=f(x^{(k)})+\left(\nabla f\left(x^{(k)}\right)\right)^T\left(x-x^{(k)}\right)+o\left(\left\|x-x^{(k)}\right\|\right)$$

令$x-x^{(k)}=\alpha d^{(k)}$，有

$$f(x)=f(x^{(k)})+\alpha\left(\nabla f\left(x^{(k)}\right)\right)^T d^{(k)}+o\left(\left\|\alpha d^{(k)}\right\|\right)$$

若$\left(\nabla f\left(x^{(k)}\right)\right)^T d^{(k)}<0$，则$d^{(k)}$为下降方向。且由Cauchy-Schwartz不等式

$$\left|\left(\nabla f\left(x^{(k)}\right)\right)^T d^{(k)}\right|\le\left\|\nabla f\left(x^{(k)}\right)\right\|\left\|d^{(k)}\right\|$$

有当且仅当$d^{(k)}=-\nabla f\left(x^{(k)}\right)$时，$-\left(\nabla f\left(x^{(k)}\right)\right)^T d^{(k)}$最大。因而以$-\nabla f\left(x^{(k)}\right)$为下降方向的方法称为最速下降法。其迭代格式为

$$x^{(k+1)}=x^{(k)}-\alpha_k\nabla f(x^{(k)}) \tag{5.24}$$

其中α_k由精确线搜索策略选取。具体计算步骤如下：

算法 5.2 (最速下降法)

1: 给定初点$x^{(1)} \in R^n$，允许误差$\varepsilon > 0$，置$k = 1$.

2: 计算搜索方向$d^{(k)} = -\nabla f(x^{(k)})$, 如果$\|\nabla f(x^{(k)})\| \leq \varepsilon$，停止计算.

3: 由精确线搜索计算步长α_k, 使得

$$f(x^{(k)} + \alpha_k d^{(k)}) = \min_{\alpha \geq 0} f(x^{(k)} + \alpha d^{(k)}).$$

4: 令$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$，令$k = k + 1$, 转步骤 2.

例 5.6 用最速下降法求解

$$\min \quad f(x) = x_1^2 + 3x_2^2,$$

设初始点$x^{(1)} = (2, 1)^T$,迭代一次.

解： 由$\nabla f(x) = \begin{pmatrix} 2x_1 \\ 6x_2 \end{pmatrix}$ 有

$$d^{(1)} = -\nabla f(x^{(1)}) = \begin{pmatrix} -4 \\ -6 \end{pmatrix} \quad x^{(1)} + \alpha d^{(1)} = \begin{pmatrix} 2 - 4\alpha \\ 1 - 6\alpha \end{pmatrix}.$$

令

$$\phi(\alpha) = f(x^{(1)} + \alpha d^{(1)}) = (2 - 4\alpha)^2 + 3\,(1 - 6\alpha)^2.$$

求解 $\min\limits_{\alpha} \phi(\alpha)$，即

$$\phi'(\alpha) = -8\,(2 - 4\alpha) - 36\,(1 - 6\alpha) = 0.$$

解得：$\alpha = \frac{13}{62}$，则得到

$$x^{(2)} = x^{(1)} + \alpha_1 d^{(1)} = \begin{pmatrix} \frac{36}{31} \\ -\frac{8}{31} \end{pmatrix}.$$

下面给出最速下降法的总体收敛性。

定理 5.10 设函数$f(x)$二次连续可微，且$\|\nabla^2 f(x)\| \leq M$, 其中$M > 0$为一常数。则对任何给定的初始点$x^{(1)}$， 由最速下降算法5.2产生的序列$\{x^{(k)}\}$满足对某个k有$\nabla f(x^{(k)}) = 0$, 或$\lim\limits_{x \to \infty} f\left(x^{(k)}\right) = -\infty$，或 $\lim\limits_{x \to \infty} \nabla f(x^{(k)}) = 0$。

证明：考虑无限迭代情形。由$d^{(k)} = -\nabla f(x^{(k)})$可知$\cos\theta_k = 1$。再由定理5.9得

$$f\left(x^{(k)}\right) - f\left(x^{(k+1)}\right) \geq \frac{1}{2M}\left\|\nabla f\left(x^{(k)}\right)\right\|^2.$$

进一步有

$$\begin{aligned} f\left(x^{(1)}\right)-f\left(x^{(k)}\right) &\geq \sum_{i=1}^{k-1}\left(f\left(x^{(i)}\right)-f\left(x^{(i-1)}\right)\right) \\ &\geq \frac{1}{2M}\sum_{i=1}^{k-1}\left\|\nabla f\left(x^{(i)}\right)\right\|^2, \end{aligned}$$

两边取极限，有 $\lim\limits_{x\to\infty} f\left(x^{(k)}\right)=-\infty$，或 $\lim\limits_{x\to\infty}\nabla f(x^{(k)})=0$成立。证毕。

最速下降法具有计算工作量小, 存储量小等优点。但是, 最速下降方向仅是函数的局部性质, 对整体求解过程而言, 这个方法下降非常缓慢。其原因是精确线性搜索使得 $\nabla f\left(x^{(k+1)}\right)^T d^{(k)}=0$，即最速下降法中前后两次迭代的搜索方向是正交的。 所以在靠近极小点附近时的路径呈现锯齿形 (图 5.1)，下降十分缓慢。

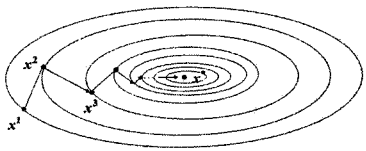


图 5.1: 最速下降法的锯齿现象

可以证明最速下降法的收敛速度是线性的。

定理 5.11 设$f(x)$二阶连续可微，由最速下降算法5.2产生的序列$\{x^{(k)}\}$收敛于最优解x^*。如果存在$m, M>0$使得对任意属于x^*领域的x有

$$m\|d\|^2 \leq d^T\nabla^2 f(x^{(k)})d \leq M\|d\|^2,\ \ \forall d\in R^n,$$

则有$\{x^{(k)}\}$线性收敛于x^*。

5.3.3 牛顿法

为加快收敛速度，我们考虑利用目标函数$f(x)$在迭代点$x^{(k)}$ 处的二阶Taylor展开式作为逼近函数，并用这个二次函数的极小点序列去逼近目标函数的极小值点。这就是牛顿法的基本思想。

设目标函数$f(x)$二次连续可微，将$f(x)$在当前迭代点$x^{(k)}$处作Taylor展开，并取二阶近似得

$$f(x)\approx f\left(x^{(k)}\right)+\nabla f\left(x^{(k)}\right)^T\left(x-x^{(k)}\right)+\frac{1}{2}\left(x-x^{(k)}\right)^T\nabla^2 f\left(x^{(k)}\right)\left(x-x^{(k)}\right) \quad (5.25)$$

极小化上式右端有

$$\nabla f\left(x^{(k)}\right)+\nabla^2 f\left(x^{(k)}\right)\left(x-x^{(k)}\right)=0.$$

若Hesse矩阵$\nabla^2 f(x^{(k)})$正定，则得到(5.25) 右端二次函数的极小点，并以它作为无约束问题最优值点的第$k+1$次近似，即

$$x^{(k+1)} = x^{(k)} - (\nabla^2 f(x^{(k)}))^{-1}\nabla f(x^{(k)}) \tag{5.26}$$

这就是牛顿法迭代公式，其中

$$d^{(k)} = -(\nabla^2 f(x^{(k)}))^{-1}\nabla f(x^{(k)}), \quad \alpha_k = 1.$$

算法 5.3 (牛顿法)

1: 给定初点$x^{(1)} \in R^n$，允许误差$\varepsilon > 0$，置$k = 1$.
2: 计算$\nabla f(x^{(k)})$, 如果$\|\nabla f(x^{(k)})\| \le \varepsilon$，停止计算。否则，转步骤 3
3: 构造牛顿方向$d^{(k)} = -(\nabla^2 f(x^{(k)}))^{-1}\nabla f(x^{(k)})$.
4: 令$x^{(k+1)} = x^{(k)} + d^{(k)}$，令$k = k+1$, 转步骤 2.

例 5.7 *用牛顿法求解例5.6，仍取初始点*$x^{(1)} = (2,1)^T$。

解：$\nabla f(x) = \begin{pmatrix} 2x_1 \\ 6x_2 \end{pmatrix}$，$\nabla^2 f(x) = \begin{pmatrix} 2 & 0 \\ 0 & 6 \end{pmatrix}$，则

$$d^{(1)} = -\begin{pmatrix} 2 & 0 \\ 0 & 6 \end{pmatrix}^{-1}\begin{pmatrix} 4 \\ 6 \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \end{pmatrix}.$$

所以

$$x^{(2)} = x^{(1)} + d^{(1)} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} + \begin{pmatrix} -2 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

且

$$\nabla f(x^{(2)}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \|\nabla f(x^{(2)})\| = 0.$$

故$x^{(2)}$为问题的最优解。

上述例子说明，牛顿法要比最速下降法收敛快。尤其对于二次凸函数，牛顿法只需一次迭代就可得到全局极小点。我们把算法用于二次凸函数时，经过有限次迭代必达到极小值点的性质称为二次终止性。对于一般非二次函数。牛顿法不能保证有限次迭代获得最优解，但是若初始点$x^{(1)}$充分靠近极小点，牛顿法在一定条件下能够达到二阶收敛速度。

定理 5.12 *设*$f(x)$*二阶连续可微，*x^**为无约束问题(5.11)的局部较小点，*$\nabla^2 f(x^*)$*正定。如果Hesse矩阵*$\nabla^2 f(x^{(k)})$*满足Lipschitz条件，即存在*$L > 0$*使得*$a \le i, j \le n$*有*

$$|(\nabla^2 f(x)_{ij} - (\nabla^2 f(x)_{ij}| \le L\|x - y\|, \quad \forall x, y \in R^n.$$

当初始点$x^{(1)}$*充分靠近*x^**，由牛顿法得到的*$\{x^{(k)}\}$*收敛到*x^**，且具有二阶收敛速度。*

值得注意的时，当初始点远离极小点时，牛顿方向可能不是下降方向，导致目标函数值可能上升。

例 5.8 用牛顿法求解：

$$min \ \ f(x) = 2(x_1 - x_2^2)^2 + (1 + x_2)^2.$$

取初始点$x^{(1)} = (0,0)^T$.

解：由于

$$\nabla f(x) = \begin{pmatrix} 4(x_1 - x_2^2) \\ -8(x_1 - x_2^2)x_2 + 2(1 + x_2) \end{pmatrix}, \quad \nabla^2 f(x) = \begin{pmatrix} 4 & -8x_2 \\ -8x_2 & 24x_2^2 - 8x_1 + 2 \end{pmatrix}$$

所以

$$\nabla f\left(x^{(1)}\right) = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \ \nabla f(x^{(1)}) = 2, \ \ \nabla^2 f\left(x^{(1)}\right) = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}.$$

构造牛顿方向：

$$d^{(1)} = -\left(\nabla^2 f\left(x^{(1)}\right)\right)^{-1} \nabla f\left(x^{(1)}\right) = -\begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}.$$

则

$$x^{(2)} = x^{(1)} + d^{(1)} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}.$$

但是注意到$f(x^{(1)}) = 1, f(x^{(2)}) = 2 > f(x^{(1)})$. 目标函数值上升。

为克服牛顿法这样的缺陷，人们提出了阻尼牛顿法，即沿牛顿方向进行一次一维搜索。具体计算步骤如下：

算法 5.4 (阻尼牛顿法)

1: 给定初点$x^{(1)} \in R^n$，允许误差$\varepsilon > 0$，置$k = 1$.

2: 计算$\nabla f(x^{(k)})$， 如果$\|\nabla f(x^{(k)})\| \leq \varepsilon$，停止计算。否则，转步骤 3

3: 构造牛顿方向$d^{(k)} = -(\nabla^2 f(x^{(k)}))^{-1} \nabla f(x^{(k)})$。由精确线搜索计算步长$\alpha_k$， 使得

$$f(x^{(k)} + \alpha_k d^{(k)}) = \min_{\alpha \geq 0} f(x^{(k)} + \alpha d^{(k)}).$$

4: 令$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$，令$k = k + 1$， 转步骤2.

例 5.9 用阻尼牛顿法求解例题5.8， 初始点$x^{(1)} = (0,0)^T$。

解：由

$$x^{(1)}+\alpha d^{(1)}=\begin{pmatrix}0\\-\alpha\end{pmatrix},$$

求

$$\min_{\alpha\geq 0} f(x^{(1)}+\alpha d^{(1)})=\min_{\alpha\geq 0} 2\alpha^4+(1-\alpha)^2.$$

得到$\alpha_1=\frac{1}{2}$，则

$$x^{(2)}=x^{(1)}+\alpha d^{(1)}=\begin{pmatrix}0\\-\frac{1}{2}\end{pmatrix}$$

且

$$\nabla f\left(x^{(2)}\right)=\begin{pmatrix}-1\\0\end{pmatrix},$$

第二次迭代：

$$\nabla^2 f\left(x^{(2)}\right)=\begin{pmatrix}4&4\\4&8\end{pmatrix}.$$

构造牛顿方向

$$d^{(2)}=-\left(\nabla^2 f\left(x^{(2)}\right)\right)^{-1}\nabla f\left(x^{(2)}\right)=\begin{pmatrix}\frac{1}{2}\\-\frac{1}{4}\end{pmatrix}.$$

有

$$x^{(2)}+\alpha d^{(2)}=\begin{pmatrix}\frac{1}{2}\alpha\\-\frac{1}{2}-\frac{1}{4}\alpha\end{pmatrix}.$$

求解

$$\min_{\alpha\geq 0} f(x^{(2)}+\alpha d^{(2)})=\min_{\alpha\geq 0}\frac{1}{128}\left[8(2-\alpha)^2+(2-\alpha)^4\right],$$

得$\alpha=2$. 则

$$x^{(3)}=x^{(2)}+\alpha d^{(2)}=\begin{pmatrix}1\\-1\end{pmatrix}.$$

因为

$$\nabla f\left(x^{(3)}\right)=\begin{pmatrix}0\\0\end{pmatrix},\quad \|\nabla f\left(x^{(3)}\right)\|=0,$$

所以极小值点为$x^{(3)}=(1,-1)^T$.

由于阻尼牛顿法有一维搜索，因此每次迭代目标函数值一般有所下降。且阻尼牛顿法在适当条件下具有全局收敛性。

定理 5.13 设$f(x)$二阶连续可微，对任意初始点$x^{(1)}$,存在常数$m>0$使得$f(x)$在水平集$L(x^{(1)})=\{x|f(x)\leq f(x^{(1)})\}$上有

$$y^T\nabla^2 f(x)y\geq m\|y\|^2,\quad \forall y\in R^n, x\in L(x^{(1)}),$$

则在精确线搜索下，阻尼牛顿法产生的序列$\{x^{(k)}\}$满足：

(1) 当$\{x^{(k)}\}$为有限序列时，其最后一个点为唯一极小点.

(2) 当$\{x^{(k)}\}$为无限序列时，其收敛到唯一极小点.

此外，牛顿法还有其他一些缺点，例如当初始点远离局部极小点时，Hesse矩阵可能不正定，则牛顿方向不是下降方向，从而导致算法失效。为克服上述缺点，人们已提出不少修正方法。解决Hesse矩阵非正定的基本思想为构造一对称正定矩阵$G(x^{(k)})$取代当前点的Hesse矩阵$\nabla^2 f(x^{(k)})$。以

$$d^{(k)} = -(G(x^{(k)}))^{-1}\nabla f(x^{(k)})$$

作为下降搜索方向。具体步骤及其他修正牛顿法可参阅相关文献。

5.3.4 共轭梯度法

在第4章中，我们介绍了求解线性方程组的共轭梯度法，事实上，这也是求解正定二次函数的共轭梯度方法。在这一节中，我们将该方法推广应用到一般的非线性函数。

注意到二次函数的Hesse矩阵即为常数矩阵A，且Hesse矩阵A出现在计算α_k和β_k的公式中。而对一般非线性函数而言，每次迭代都需重新计算当前点的Hesse矩阵$\nabla^2 f(x^{(k)})$，计算量偏大。因此为方便将方法推广，我们需对α_k和β_k的计算做必要的修改。

(1) 对于非二次函数，精确步长因子无法获得显式公式表达，α_k将由精确线搜索过程确定，即计算一维优化问题$\min\limits_{\alpha\geq 0} f(x^{(k)}+\alpha d^{(k)})$.

(2) 我们必须去除β_k公式中的Hesse矩阵的计算。由第4.4节中(4.56)及$\nabla\varphi(x) = Ax-b = -r_k$，我们有

$$\begin{aligned}\beta_k &= (r_{k+1}, r_{k+1})/(r_k, r_k)\\ &= (\nabla\varphi(x^{(k+1)}), \nabla\varphi(x^{(k+1)}))/(\nabla\varphi(x^{(k)}), \nabla\varphi(x^{(k)})).\end{aligned} \tag{5.27}$$

由此可知，β_k的计算可以只用目标函数值和梯度值完成公式修正。进而，将公式(5.27)中的二次函数$\varphi(x)$取代为一般非线性函数$f(x)$，则得到一般函数情形下的 $Fletcher-Reeves$公式，

$$\beta_k = \frac{\|\nabla f(x^{(k+1)})\|^2}{\|\nabla f(x^{(k)})\|^2} \tag{5.28}$$

我们给出求解一般非线性函数的FR共轭梯度法计算步骤：

由于共轭梯度法中含有最速下降法步骤，因此在较弱的条件下可以保证方法的总体收敛性。

算法 5.5 (FR共轭梯度法)

1: 给定初点$x^{(1)} \in R^n$，允许误差$\varepsilon > 0$，令$d^{(1)} = -\nabla f(x^{(1)})$。置$k = 1$.

2: 如果$\|\nabla f(x^{(k)})\| \le \varepsilon$，停止计算。否则，转步骤 3.

3: 由精确线搜索计算步长α_k, 使得

$$f(x^{(k)} + \alpha_k d^{(k)}) = \min_{\alpha \ge 0} f(x^{(k)} + \alpha d^{(k)}).$$

令$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$.

4: 计算$\beta_k = \frac{\|\nabla f(x^{(k+1)})\|^2}{\|\nabla f(x^{(k)})\|^2}$, $d^{(k+1)} = -\nabla f(x^{(k+1)} + \beta_k d^{(k)}$.

5: 令$k = k + 1$, 转步骤2.

定理 5.14 设$f(x)$在有界水平集$L = \{x | f(x) \le f(x^{(1)})\}$上连续可微且下有界，则由算法5.5产生的序列$\{x^{(k)}\}$至少有一个聚点是驻点，即

(1) 当$\{x^{(k)}\}$是有穷点列时，其最后一个点是$f(x)$的驻点.

(2) 当$\{x^{(k)}\}$是无穷点列时，其必有聚点且每个聚点都是$f(x)$的驻点.

证明：(1) 当$\{x^{(k)}\}$是有穷点列时，由算法的终止性条件可知, 其最后一个点x^*满足$\nabla f(x^*) = 0$, 故x^*是$f(x)$的驻点.

(2) 当$\{x^{(k)}\}$是无穷点列时，对所有$k, \nabla f(x^{(k)}) \ne 0$, 由于$d^{(k)} = -\nabla f(x^{(k)} + \beta_{k-1} d^{(k-1)}$，故

$$(\nabla f(x^{(k)})^T g^{(k)} = -\|\nabla f(x^{(k)}\| + \beta_{k-1} (\nabla f(x^{(k)})^T d^{(k-1)} = -\|\nabla f(x^{(k)}\| < 0.$$

从而$d^{(k)}$是下降方向。由于$\{f(x^{(k)})\}$是单调下降且有下界的序列，故 $\lim\limits_{x \to \infty} f\left(x^{(k)}\right) = f^*$ 存在。由 $\{x^{(k)}\}$ 属于水平集 L 及 L 有界得 $\{x^{(k)}\}$ 是有界点列，故存在收敛子列 $\{x^{(k)}\}_{K_1} \to x^*$，这里 K_1 是子序列的指标集。由于 $\{x^{(k)}\}_{K_1} \subset \{x^{(k)}\}$, 故$\{f(x^{(k)})\}_{K_1} \subset \{f(x^{(k)})\}$，从而由 f 的连续性可知，对于 $k \in K_1$，有

$$f(x^*) = f\left(\lim_{\substack{k \to \infty \\ k \in K_1}} x^{(k)}\right) = \lim_{\substack{k \to \infty \\ k \in K_1}} f\left(x^{(k)}\right) = f^*.$$

类似地, $\{x_{k+1}\}$也是有界点列, 故存在$\{x_{k+1}\}_{K_2} \to (\overline{x})^*$, 这里$K_2$ 是$\{x_{k+1}\}$ 的子序列的指标集.故对于$k + 1 \in K_2$,

$$f(\overline{x}^*) = f\left(\lim_{\substack{k \to \infty \\ k+1 \in K_2}} x^{(k+1)}\right) = \lim_{\substack{k \to \infty \\ k+1 \in K_2}} f\left(x^{(k+1)}\right) = f^*.$$

于是$f(\overline{x}^*) = f(x^*)$.

下证$\nabla f(x^*)=0$。用反证法。假设$\nabla f(x^*)\neq 0$,则存在一个方向d^*使得对充分小的α有

$$f(x^*+\alpha d^*)<f(x^*). \tag{5.29}$$

由于

$$f(x^{(k+1)})=f(x^*+\alpha_k d^*)<f(x^*+\alpha d^*),\forall\alpha,$$

故对$k+1\in K_2$, 令$k\to\infty$, 应用(5.29)得

$$f(\overline{x}^*)=f(x^*+\alpha d^*)<f(x^*+\alpha d^*)<f(x^*),$$

这与$f(\overline{x}^*)=f(x^*)$矛盾。则证明$\nabla f(x^*)=0$, 则x^*为驻点。 证毕。

应用定理5.14的条件可以证明采用精确线搜索的FR共轭梯度法产生的迭代点序列$\{x^{(k)}\}$ 线性收敛到极小点x^*。

例 5.10 *用FR共轭梯度法求解如下问题*

$$\min\ f(x)=x_1^2+2x_2^2-4x_1-2x_1x_2,$$

初始点为$x^{(1)}=(1,1)^T$.

解：由

$$\nabla f(x)=\begin{pmatrix}2x_1-4-2x_2\\-2x_1+4x_2\end{pmatrix},$$

则

$$\nabla f(x^{(1)})=\begin{pmatrix}-4\\2\end{pmatrix},\quad d^{(1)}=-\nabla f\left(x^{(1)}\right)=\begin{pmatrix}4\\-2\end{pmatrix},$$

$$x^{(1)}+\alpha d^{(1)}=\begin{pmatrix}1+4\alpha\\1-2\alpha\end{pmatrix}.$$

代入目标函数，求解

$$\min\ 40\alpha^2-20\alpha-3,$$

得：$\alpha_1=\frac{1}{4}$. 则有

$$x^{(2)}=x^{(1)}+\alpha_1 d^{(1)}=\begin{pmatrix}1\\1\end{pmatrix}+\frac{1}{4}\begin{pmatrix}4\\-2\end{pmatrix}=\begin{pmatrix}2\\\frac{1}{2}\end{pmatrix}.$$

又

$$\nabla f(x^{(2)})=\begin{pmatrix}-1\\-2\end{pmatrix},\quad \left\|\nabla f(x^{(2)})\right\|=\sqrt{5},$$

由公式 (5.28) 得

$$\beta_1 = \frac{\left\|\nabla f(x^{(2)})\right\|^2}{\left\|\nabla f(x^{(1)})\right\|^2} = \frac{1}{4}.$$

则

$$d^{(2)} = -\nabla f(x^{(2)}) + \beta_1 d^{(1)} = -\begin{pmatrix} -1 \\ -2 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 4 \\ -2 \end{pmatrix} = \begin{pmatrix} 2 \\ \frac{3}{2} \end{pmatrix},$$

$$x^{(2)} + \alpha d^{(2)} = \begin{pmatrix} 2+2\alpha \\ \frac{1}{2}(1+3\alpha) \end{pmatrix}.$$

带入目标函数，求解

$$\min \frac{5}{2}\alpha^2 - 5\alpha - \frac{11}{2},$$

得 $\alpha_2 = 1$. 且

$$x^{(3)} = x^{(2)} + \alpha_2 d^{(2)} = \begin{pmatrix} 2 \\ \frac{1}{2} \end{pmatrix} + 1\begin{pmatrix} 2 \\ \frac{3}{2} \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \end{pmatrix},$$

$$\nabla f(x^{(3)}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \left\|\nabla f(x^{(3)})\right\| = 0.$$

则极小值点为

$$x^* = x^{(3)} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}.$$

除公式(5.28)外，还有如下常用的β_k公式：

Polak-Ribiere-Polyak公式(PRP公式)

$$\beta_k = \frac{\nabla f(x^{(k+1)})^T[\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})]}{\left\|\nabla f(x^{(k)})\right\|^2} \tag{5.30}$$

Dixon-Myers公式(DM公式)

$$\beta_k = \frac{\left\|\nabla f(x^{(k+1)})\right\|^2}{(d^{(k)})^T\nabla f(x^{(k)})} \tag{5.31}$$

Dai-Yuan公式(DY公式)

$$\beta_k = \frac{\left\|\nabla f(x^{(k+1)})\right\|^2}{(d^{(k)})^T[\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})]} \tag{5.32}$$

将上述三个公式分别带入算法中，可分别获得相应的梯度算法。

需要说明的是，对于正定二次函数的无约束最优化问题，如果采用精确线搜索，以上关于β_k的共轭梯度方法完全等价。但是对于非二次函数的无约束优化问题，他们产生的搜索方向是不同的，且n步之后构造的搜索方向不再是共轭的，从而降低了收敛速度。克服这样缺点的方法是采取在开始技巧，即每n步以后周期性采用x_n为初始点，最速下降方向作为新的搜索方向重新迭代。

5.4 罚函数法

我们考虑如下约束优化问题：

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & h_i(x) = 0, i \in \mathcal{E} \\ & g_j(x) \geq 0, j \in \mathcal{I}, \end{aligned} \tag{5.33}$$

其中$\mathcal{E} = \{1, 2, .., m\}$，$\mathcal{I} = \{1, 2, ..., l\}$。本节中我们将介绍求解约束最优化问题的罚函数方法，它是求解约束问题的重要方法之一。

罚函数方法的基本思想是借助罚函数将约束问题转化为无约束优化问题，进而通过求解一系列无约束最优化问题来获取原约束问题的解。迭代过程中，罚函数法通过对不可行点施加惩罚，迫使迭代点向可行域靠近。一旦迭代点成为可行点，则这个可行点就是原问题的最优解。

采用不同的罚函数，就形成不同的罚方法。同时，我们是将求解约束问题转化为求解一系列无约束问题，从而获得原问题的逼近最优解。因此该方法也称为序列无约束极小化方法。

5.4.1 外点罚函数法

一般地，罚函数是由目标函数及约束函数所构成的一辅助函数

$$F(x, \sigma) = f(x) + \sigma P(x),$$

其中σ为一很大的正常数，称为罚因子，$\sigma P(x)$称为罚项，且$P(x)$需满足

(1) $P(x)$为连续函数.

(2) 对所有$x \in R^n$, $P(x) \geq 0$.

(3) $P(x) = 0$当且仅当x为问题可行点.

对于问题(5.33)，我们定义函数

$$P(x)=\sum_{i=1}^{m}\varphi(h_i(x))+\sum_{j=1}^{l}\psi(g_j(x)) \tag{5.34}$$

其中$\varphi(x),\psi(x)$满足

$$\begin{cases}\varphi(x)=0,x=0\\ \varphi(x)>0,x\neq 0,\end{cases}\quad\begin{cases}\psi(x)=0,x\geq 0\\ \psi(x)>0,x<0.\end{cases}$$

例如，我们取函数$\varphi(x),\psi(x)$为

$$\begin{aligned}\varphi(x)&=|h_i(x)|^{\alpha}\\ \psi(x)&=[\max\{0,\ -g_j(x)\}]^{\beta}\end{aligned}$$

其中$\alpha\geq 1,\beta\geq 1$。通常我们设定$\alpha=\beta=2$。

应用(5.34)，我们取一个趋向无穷大的严格递增正数序列$\{\sigma_k\}$，对每一个k，得到求解约束问题(5.33)的无约束问题，即求解

$$\min F(x,\sigma)=f(x)+\sigma_k P(x). \tag{5.35}$$

方法的优点在于初始点的选择没有可行性限制，且随着σ增大，迭代从可行域外部逼近约束问题的最优解。我们把这种利用罚函数生成一系列外点逼近该约束问题最优解的方法称为 外点罚函数法。具体计算步骤如下：

算法 5.6 (外点罚函数法)

1: 给定初点$x^{(1)}\in R^n$，初始罚因子σ_1，放大系数$\gamma>1$，允许误差$\varepsilon>0$，置$k=1$.

2: 以$x^{(k)}$为初始点，求解无约束问题(5.35)得最优解$x^{(k+1)}$.

3: 如果$\sigma_k P(x^{(k+1)})<\varepsilon$，则停止计算，$x^{(k+1)}$为约束问题(5.33)的近似最优解；否则，增大罚因子$\sigma_{k+1}=\gamma\sigma_k$，令$k=k+1$，转步骤2.

为给出外点罚函数法收敛性，我们先给出方法的性质。

性质 5.1 设$0<\sigma_k<\sigma_{k+1}$，$x^{(k)}$和$x^{(k+1)}$分别是罚因子分别为σ_k和σ_{k+1}时的无约束问题(5.35) 的最优解，则

(1) $\{F(x^{(k)},\sigma_k)\}$为非减序列，即 $F(x^{(k+1)},\sigma_{k+1})\geq F(x^{(k)},\sigma_k)$.

(2) $\{P(x^{(k)})\}$为非增序列，即$P(x^{(k+1)})\leq P(x^{(k)})$.

(3) $\{f(x^{(k)})\}$为非减序列，即$f(x^{(k+1)})\geq f(x^{(k)})$.

(4) 设x^*为约束问题(5.33)的最优解，则$f(x^*)\geq F(x^{(k)},\sigma_k)\geq f(x^{(k)})$.

证明：(1) 由$x^{(k)}$为$F(x,\sigma_k)=f(x)+\sigma_k P(x)$的极小点及$\sigma_k<\sigma_{k+1}$，有

$$\begin{aligned}F(x^{(k+1)},\sigma_{k+1}) &= f(x^{(k+1)})+\sigma_{k+1}P(x^{(k+1)})\\ &\geq f(x^{(k+1)})+\sigma_k P(x^{(k+1)})\\ &\geq f(x^{(k)})+\sigma_k P(x^{(k)})\\ &= F(x^{(k)},\sigma_k)\end{aligned}$$

(2) 由于$x^{(k)}$和$x^{(k+1)}$分别是$F(x,\sigma_k),F(x,\sigma_{k+1})$的极小点，则有

$$f(x^{(k+1)})+\sigma_k P(x^{(k+1)})\geq f(x^{(k)})+\sigma_k P(x^{(k)}), \tag{5.36}$$

$$f(x^{(k)})+\sigma_{k+1}P(x^{(k)})\geq f(x^{(k+1)})+\sigma_{k+1}P(x^{(k+1)}), \tag{5.37}$$

将(5.36)和(5.37)相加，得

$$\sigma_k P(x^{(k+1)})+\sigma_{k+1}P(x^{(k)})\geq \sigma_k P(x^{(k)})+\sigma_{k+1}P(x^{(k+1)}).$$

即$(\sigma_{k+1}-\sigma_k)\,P(x^{(k)})\geq(\sigma_{k+1}-\sigma_k)\,P(x^{(k+1)})$.由于$\sigma_k<\sigma_{k+1}$, 则有$P(x^{(k+1)})\leq P(x^{(k)})$.

(3) 由(5.36)及$P(x^{(k+1)})\leq P(x^{(k)})$得

$$f(x^{(k+1)})-f(x^{(k)})\geq\sigma_k\left(P(x^{(k)})-P(x^{(k+1)})\right)\geq 0,$$

即 $f(x^{(k+1)})\geq f(x^{(k)})$.

(4) 因为x^*是约束问题 (5.33) 的最优解，则$P(x^*)=0$且$f(x^*)=F(x^*,\sigma_k)$。又因为$x^{(k)}$为$F(x,\sigma_k)$的极小点且有$\sigma_k P(x^{(k)})\geq 0$，得

$$f(x^*)=F(x^*,\sigma_k)\geq F(x^{(k)},\sigma_k)=f(x^{(k)})+\sigma_k P(x^{(k)})\geq f(x^{(k)}).$$

证毕。

下面给出外点罚函数的收敛性结论。

定理 5.15 *设约束最优化问题 (5.33) 存在最优解，其中 $f(x),h_i,g_j$ 为实值连续函数，$\{\sigma_k\}$ 为严格递增正数列且趋向于无穷大的数列，序列 $\{x^{(k)}\}$ 由算法 5.6 产生，则序列 $\{x^{(k)}\}$ 的任何极限点是问题 (5.33) 的解。*

证明：设x^*为约束优化问题(5.33)的最优解，$\bar{x}$为$\{x^{(k)}\}$的一个极限点。由性质5.1(4)可知，$f(x^{(k)})\leq F(x^{(k)},\sigma_k)\leq f(x^*)$, 因为$f(x)$为连续函数，则有

$$f(\bar{x})\leq f(x^*). \tag{5.38}$$

又由性质 5.1 (1)和(3)有 $\{F(x^{(k)},\sigma_k)\}$ 和 $\{f(x^{(k)})\}$ 是单调递增且有上届的数列，则 $\lim\limits_{k\to\infty}F(x^{(k)},\sigma_k)$ 和 $\lim\limits_{k\to\infty}f(x^{(k)})$ 存在。另一方面，

$$P\left(x^{(k)}\right)=\frac{F(x^{(k)},\sigma_k)-f\left(x^{(k)}\right)}{\sigma_k}.$$

令 $k \to \infty$, 两端同时取极限，由 $\sigma_k \to \infty$，我们有 $\lim\limits_{k\to\infty} P\left(x^{(k)}\right) = 0$. 再由 $h_i(x)$ 和 $g_j(x)$ 的连续性，得

$$P(\overline{x}) = \lim_{k\to\infty} P\left(x^{(k)}\right) = 0.$$

这说明 $\overline{x}$ 为约束问题的可行点。

此外，由 x^* 为约束问题 (5.33) 的极小点及 $\overline{x}$ 为可行点，得

$$f(x^*) \le f(\overline{x}). \tag{5.39}$$

因此由 (5.38) 和 (5.39) 得 $f(x^*) = f(\overline{x})$. 即 $\overline{x}$ 为约束问题 (5.33) 的极小点。证毕。

例 5.11 用外点罚函数求解

$$\begin{aligned} &\min\ (x_1-3)^2 + (x_2-2)^2 \\ &s.t\ \ 4 - x_1 - x_2 \ge 0. \end{aligned}$$

解：构造罚函数

$$F(x,\sigma) = (x_1-3)^2 + (x_2-2)^2 + \sigma[\max\{0,\ x_1+x_2-4\}]^2.$$

简化得

$$F(x,\sigma) = \begin{cases} (x_1-3)^2 + (x_2-2)^2, & x_1+x_2-4 \le 0 \\ (x_1-3)^2 + (x_2-2)^2 + \sigma(x_1+x_2-4)^2, & x_1+x_2-4 > 0 \end{cases}$$

通过求解无约束问题 $\min\limits_{x\in R^n} F(x,\sigma)$ 可求得原问题的近似解。我们用解析方法求解无约束问题， 则有

$$\frac{\partial F(x,\sigma)}{\partial x_1} = \begin{cases} 2(x_1-3), & x_1+x_2 \le 4 \\ 2(x_1-3) + 2\sigma(x_1+x_2-4), & x_1+x_2 > 4, \end{cases}$$

$$\frac{\partial F(x,\sigma)}{\partial x_2} = \begin{cases} 2(x_2-2), & x_1+x_2 \le 4 \\ 2(x_2-2) + 2\sigma(x_1+x_2-4), & x_1+x_2 > 4. \end{cases}$$

令 $\frac{\partial F(x,\sigma)}{\partial x_1} = \frac{\partial F(x,\sigma)}{\partial x_2} = 0$, 得

$$x(\sigma) = \begin{pmatrix} \frac{5\sigma+3}{2\sigma+1} \\ \frac{3\sigma+2}{2\sigma+1} \end{pmatrix}.$$

同时求得

$$\nabla^2 F(x,\sigma) = \begin{pmatrix} 2(\sigma+1) & 2\sigma \\ 2\sigma & 2(\sigma+1) \end{pmatrix},$$

由于$\sigma > 0$，所以$\nabla^2 F$为正定阵。因此$x(\sigma)$为$F(x,\sigma)$的极小点。

令$\sigma \to +\infty$，则有

$$x^* = \lim_{\sigma\to\infty} x(\sigma) = \left(\frac{5}{2},\ \frac{3}{2}\right)^T,$$

且x^*为可行点，所以原问题的最优解为

$$x^* = \begin{pmatrix} \frac{5}{2} \\ \frac{3}{2} \end{pmatrix}$$

最优值分别为 $f(x^*) = \frac{1}{2}$.

外点罚函数形式简单，但是由于罚参数趋向于无穷时，可能导致罚函数的Hesse矩阵条件数变化，导致方法的数值性能变坏。因此σ需选取适中，或者我们可采用广义乘子法。

5.4.2 内点罚函数法

内点罚函数法是一类保持严格可行性的方法，它总是从可行点出发，并保持在可行域内部进行搜索。因而这类方法只适用于只有不等式约束的非线性最优化问题：

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & g_i(x) \geq 0, i \in \mathcal{I}, \end{aligned} \tag{5.40}$$

其中$\mathcal{I} = \{1, 2, ..., l\}$.

内点罚函数法的基本思想为在目标函数上引入一个关于约束的障碍项，当迭代点由可行域的内部接近可行域的边界时，障碍项将趋于无穷大来迫使迭代点返回可行域的内部，从而保持迭代点的严格可行性。进而将求解约束问题转为求解一系列容易的子问题，从而获得原问题的最优逼近解。方法也称为内点障碍函数法。

记问题(5.40)的可行域$\mathcal{F}$，可行域的内部

$$int\ \mathcal{F} = \{x \in R^n | g_i > 0, i \in \mathcal{I}\}.$$

对于问题(5.40)，障碍函数$B(x)$一般需满足：

(1) 在$int\ \mathcal{F}$中连续.

(2) 当x趋于$int\ \mathcal{F}$的边界，$B(x) \to \infty$.

两种常用的障碍函数形式为

(1) 倒数障碍函数

$$B(x) = \sum_{i=1}^{l} \frac{1}{g_i(x)}$$

(2) 对数障碍函数

$$B(x) = -\sum_{i=1}^{l} \ln g_i(x)$$

由障碍函数，我们可以定义罚函数

$$F(x,\mu)=f(x)+\mu B(x),$$

其中 μ 是很小的数。则当 x 趋于边界时，$F(x,\mu)\to+\infty$，否则，当 μ 很小时，$F(x,\mu)$ 的数值近似于 $f(x)$。因此，我们可以通过求解

$$\begin{aligned}&\min\quad F(x,\mu)\\&s.t.\quad x\in int\ \mathcal{F}\end{aligned}\tag{5.41}$$

得到原约束问题(5.40)的近似解。对于近似问题(5.41), 从形式上看还是约束问题，但是由于障碍函数$B(x)$的阻拦是自动实现的，所以在计算时，我们仍可把它当作无约束问题处理。但另一方面，如果μ取值太小，将给求解问题(5.41)带来很大的困难。因此，我们仍采取序列无约束极小化技巧处理计算。即取一个严格单调递减且趋向于0的罚因子数列$\{\mu_k\}$，对每一个k，从内点出发，求解

$$\begin{aligned}&\min\quad F(x,\mu_k)\\&s.t.\quad x\in int\ \mathcal{F}.\end{aligned}\tag{5.42}$$

具体计算步骤如下：

算法 5.7 (内点罚函数法)

1: 给定初点$x^{(1)}\in int\ \mathcal{F}$，初始罚因子$\mu_1$, 缩小系数$\gamma<1$, 允许误差$\varepsilon>0$，置$k=1$.
2: 以$x^{(k)}$为初始点，求解无约束问题(5.42)得最优解$x^{(k+1)}$.
3: 如果$\mu_k B(x^{(k+1)})<\varepsilon$，则停止计算，$x^{(k+1)}$为约束问题(5.40)的近似最优解；否则，减小罚因子$\mu_{k+1}=\gamma\mu_k$，令$k=k+1$, 转步骤2.

内点罚函数有如下收敛性结果。

定理 5.16 *设约束最优化问题 (5.40) 的可行域内部$int\ \mathcal{F}$ 非空且存在最优解，其中$f(x),g_i$ 为实值连续函数，$\{\mu_k\}$ 为严格递减正数列且趋向于零的数列，序列 $\{x^{(k)}\}$ 由算法 5.7 产生，则序列 $\{x^{(k)}\}$ 的任何极限点是问题 (5.40) 的解。*

例 5.12 *用对数障碍罚函数法求解下列问题*

$$\begin{aligned}&\min\ \tfrac{1}{2}(x_1+1)^2+x_2\\&s.t\quad x_1-1\geq 0\\&\qquad x_2\ \ \geq 0.\end{aligned}$$

解：构造罚函数

$$F(x,\mu)=\frac{1}{2}(x_1+1)^2+x_2-\mu\ln(x_1-1)-\mu\ln(x_2),$$

其中μ为很小的正数。令

$$\frac{\partial F(x,\mu)}{\partial x_1} = (x_1+1) - \frac{\mu}{x_1-1} = 0,$$

$$\frac{\partial F(x,\mu)}{\partial x_2} = 1 - \frac{\mu}{x_2} = 0.$$

解得

$$x(\mu) = \begin{pmatrix} \sqrt{1+\mu} \\ \mu \end{pmatrix}.$$

同时，当前点的Hesse矩阵

$$\nabla^2 F(x,\mu) = \begin{pmatrix} 1+\frac{\mu}{(\sqrt{1+\mu}-1)^2} & 0 \\ 0 & \frac{1}{\mu} \end{pmatrix}$$

为正定阵，所以$x(\mu)$为$F(x,\mu)$的极小点。当$\mu \to 0$时，$x(\mu) \to x^* = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. 目标函数最优解：$f^* = 1$.

需要注意的是内点罚函数中必须知道一个初始内点$x^{(1)} \in int\ \mathcal{F}$, 如果这个初始内点不能直观找出，就必须寻求寻找初始内点的算法。同时，类似于外点罚函数，随着μ_k 趋向于零，罚函数的Hesse矩阵条件数可能变得病态，从而给计算带来不便。

5.4.3 广义乘子法

广义乘子法的基本思为把外点罚函数与Lagrange函数结合起来，构造出新罚函数，使得在罚因子适当大的情况下，借助于Lagrange乘子就能逐步达到原约束问题的最优解。

我们首先考虑等式约束问题

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & h_i(x) = 0, i \in \mathcal{E}, \end{aligned} \tag{5.43}$$

其中$\mathcal{E} = \{1,2,...,m\}$.

记

$$h(x) = \begin{pmatrix} h_1(x) \\ \vdots \\ h_m(x) \end{pmatrix}, \quad v = \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix}.$$

定义乘子罚函数：

$$\begin{aligned} \phi(x,v,\sigma) &= f(x) - \sum_{i=1}^{m} v_i h_i(x) + \frac{\sigma}{2}\sum_{i=1}^{m} h_i^2(x) \\ &= f(x) - v^T h(x) + \frac{\sigma}{2} h(x)^T h(x). \end{aligned} \tag{5.44}$$

进而我们可以通过求解

$$\min \phi(x, v, \sigma) \tag{5.45}$$

获得原问题(5.43)的局部最优解。

定理 5.17 *设x^*为等式约束问题(5.43)的一个局部最优解，且满足二阶充分条件，即存在乘子$v^* = (v_1^*, \cdots, v_m^*)$使得*

$$\nabla f(x^*) - Av^* = 0,$$
$$h_i(x^*) = 0, i = 1, \cdots, m$$

且对每一个满足$d^T \nabla h_i(x^*) = 0 (i = 1, \cdots, m)$的非零向量$d$，有

$$d^T \nabla_x^2 L(x^*, v^*) d > 0$$

其中$A = (\nabla h_1(x^*), \cdots, \nabla h_l(x^*))$，$L(x, v) = f(x) - v^T h(x)$. 则存在$\overline{\sigma} \geq 0$使得所有$\sigma > \overline{\sigma}$, x^*为$\phi(x, v^*, \sigma)$的严格局部极小解。反之，若存在点x^*使得

$$h_i(x^*) = 0, i = 1, \cdots, m,$$

及对于某个v^*, x^*为$\min \phi(x, v^*, \sigma)$的极小点且满足二阶充分条件，则$x^*$为等式约束问题(5.43)的一个严格局部最优解。

根据定理5.17，我们可知如果最优乘子v^*已知，那么只要取充分大的罚因子，且不需趋向无穷大，就能通过极小化求出问题(5.43)的解。但v^*的值并不能事先知道，这就需要构造迭代过程来不断修正乘子v的值使其趋向于最优乘子v^*。下面我们给出具体构造过程：在第k次迭代中，对于Lagrange乘子向量的估计值$v^{(k)}$，罚因子σ，我们通过计算得到的$\phi(x, v^{(k)}, \sigma)$的极小点$x^{(k)}$，且有

$$\nabla_x \varphi(x^{(k)}, v^{(k)}, \sigma) = \nabla f(x^{(k)}) - \sum_{i=1}^{m} (v_i^{(k)} - \sigma h_i(x^{(k)})) \nabla h_i(x^{(k)}) = 0. \tag{5.46}$$

若$x^{(k)}$也是问题(5.43)的最优解，则存在v^*使得

$$\nabla f(x^{(k)}) - \sum_{j=1}^{l} v_i^* \nabla h_i(x^{(k)}) = 0. \tag{5.47}$$

比较(5.46)和(5.47)得到

$$v_i^* = v_i^{(k)} - \sigma h_i(x^{(k)}).$$

由此我们可以给出乘子v的修正公式

$$v_i^{(k+1)} = v_i^{(k)} - \sigma h_i(x^{(k)}), i = 1, \cdots, m. \tag{5.48}$$

等式约束问题的乘子法计算步骤如下：

算法 5.8 (等式约束问题的广义乘子罚函数法)

1: 给定初点$x^{(1)}$，初始乘子向量$v^{(1)}$, 初始罚因子σ_1, 放大系数$\gamma > 1$, 常数$\beta \in (0,1)$, 允许误差$\varepsilon > 0$，置$k = 1$.

2: 以$x^{(k)}$为初始点，固定$v = v^{(k)}$求解无约束问题(5.45)得最优解$x^{(k+1)}$.

3: 如果$\|h(x^{(k+1)})\| < \varepsilon$，则停止计算，$x^{(k+1)}$为约束问题(5.43)的近似最优解. 否则，进行步骤4.

4: 若$\frac{\|h(x^{(k+1)})\|}{\|h(x^{(k)})\|} \geq \beta$, 令$\sigma_{k+1} = \gamma\sigma_k$. 否则转步骤5; 否则，进行步骤5.

5: 用(5.48)更新乘子,令$k = k+1$, 转步骤2.

例 5.13 *用广义乘子法求解*

$$\begin{aligned} &\min 2x_1^2 + x_2^2 - 2x_1x_2 \\ &s.t \quad x_1 + x_2 - 1 = 0, \end{aligned}$$

取$\sigma = 2$，初始乘子$v^{(1)} = 1$。

解：定义增广Lagrange函数如下

$$\varphi(x, v, \sigma) = 2x_1^2 + x_2^2 - 2x_1x_2 - v(x_1 + x_2 - 1) + \frac{\sigma}{2}(x_1 + x_2 - 1)^2.$$

用解析法解

$$\min \ \varphi(x, 1, 2),$$

即求$\frac{\partial\varphi}{\partial x_1} = \frac{\partial\varphi}{\partial x_2} = 0$，得

$$x^{(1)} = \begin{pmatrix} \frac{1}{2} \\ \frac{3}{4} \end{pmatrix}.$$

按公式修正乘子

$$v^{(2)} = v^{(1)} - \sigma h(x^{(1)}) = 1 - 2 \cdot \frac{1}{4} = \frac{1}{2}.$$

再设定乘子值为$v^{(2)}$，求解

$$\min \ \varphi\left(x, \frac{1}{2}, 2\right)$$

依次迭代。一般地，在第k次迭代时,设定$v^{(k)}$, 由$\sigma = 2$求解

$$\min \ \varphi(x, v^{(k)}, 2)$$

得

$$x^{(k)} = \begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \end{pmatrix} = \begin{pmatrix} \frac{1}{6}(v^k + 2) \\ \frac{1}{4}(v^k + 2) \end{pmatrix}.$$

再由修正公式(5.48)及$x^{(k)}$的取值，得

$$v^{(k+1)} = \frac{1}{6}v^{(k)} + \frac{1}{3}.$$

令$k \to \infty$, 得

$$v^{(k)} \to \frac{2}{5},\ x^{(k)} \to \begin{pmatrix} \frac{2}{5} \\ \frac{3}{5} \end{pmatrix}.$$

下面我们讨论不等式约束问题的广义乘子法。考虑不等式约束问题

$$\begin{aligned} &\min \quad f(x) \\ &s.t. \quad g_i(x) \geq 0, i \in \mathcal{I}, \end{aligned} \tag{5.49}$$

其中$\mathcal{I} = \{1, 2, ..., l\}$. 为应用等式约束问题的广义乘子方法，我们引入变量$y_i \geq 0$, 将不等式约束问题(5.49) 化为如下等式约束问题

$$\begin{aligned} &\min \quad f(x) \\ &s.t. \quad g_i(x) - y_i = 0, i \in \mathcal{I}, \end{aligned} \tag{5.50}$$

由(5.50)定义增广Lagrange函数

$$\varphi(x, y, v, \sigma) = f(x) - \sum_{i=1}^{l} v_i(g_i(x) - y_i) + \frac{\sigma}{2} \sum_{i=1}^{l} (g_i(x) - y_i)^2. \tag{5.51}$$

将$\varphi(x, y, v, \sigma)$关于y极小化，且由$y_i \geq 0$，我们有

$$\begin{cases} y_i = \frac{1}{\sigma}(\sigma g_i(x) - w_i) & \sigma g_i(x) - v_i \geq 0 \\ y_i = 0 & \sigma g_i(x) - v_i < 0 \end{cases}$$

即

$$y_i = \frac{1}{\sigma} \max\{0, \sigma g_i(x) - v_i\}.$$

带入到(5.51)得

$$\varphi(x, v, \sigma) = f(x) + \frac{1}{2\sigma} \sum_{i=1}^{l} \{[\max(0, v_i - \sigma g_i(x))]^2 - v_i^2\}. \tag{5.52}$$

则我们可以通过极小化问题(5.52)获得原问题的逼近最优解。同时类似等式约束问题的方法，我们可以获得乘子得更新公式为

$$v_i^{(k+1)} = \max(0, v_i^{(k)} - \sigma g_i(x^{(k)})),\ i = 1, \cdots, l. \tag{5.53}$$

例 5.14 *用广义乘子法求解下列问题*

$$\begin{aligned} &\min \quad x_1^2 + x_2^2 \\ &s.t. \quad x_1 - 1 \geq 0. \end{aligned}$$

算法 5.9 (不等式约束问题的广义乘子罚函数法)

1: 给定初点$x^{(1)}$，初始乘子向量$v^{(1)}$, 初始罚因子σ_1, 放大系数$\gamma>1$, 常数$\beta\in(0,1)$, 允许误差$\varepsilon>0$，置$k=1$.

2: 以$x^{(k)}$为初始点，固定$v=v^{(k)}$求解无约束问题(5.52)得最优解$x^{(k+1)}$.

3: 如果$\|\max\{0,-g(x^{(k+1)})\}\|<\varepsilon$，则停止计算，$x^{(k+1)}$为约束问题(5.43)的近似最优解. 否则，进行步骤4.

4: 若$\frac{\|\max\{0,-g(x^{(k+1)})\}\|}{\|\max\{0,-g(x^{(k)})\}\|}\geq\beta$, 令$\sigma_{k+1}=\gamma\sigma_k$; 否则转步骤5; 否则，进行步骤5.

5: 用(5.53)更新乘子,令$k=k+1$, 转步骤2.

解：定义增广Lagrange函数如下

$$\begin{aligned}\varphi(x,v,\sigma) &= x_1^2+x_2^2+\frac{1}{2\sigma}\left[\left(\max\left\{0,v-\sigma\left(x_1-1\right)\right\}\right)^2-v^2\right]\\ &= \begin{cases} x_1^2+x_2^2+\frac{1}{2\sigma}\left\{\left[v-\sigma\left(x_1-1\right)\right]^2-v^2\right\}, & x_1-1\leq\frac{v}{\sigma}\\ x_1^2+x_2^2-\frac{v^2}{2\sigma}, & x_1-1>\frac{v}{\sigma}.\end{cases}\end{aligned}$$

设第k次迭代取乘子v^k,σ, 用解析法求 $\min\ \varphi\left(x,v^{(k)},\sigma\right)$, 由

$$\frac{\partial\varphi}{\partial x_1}=\begin{cases}2x_1-\left[v-\sigma\left(x_1-1\right)\right] & x_1-1\leq\frac{v}{\sigma}\\ 2x_1 & x_1-1>\frac{v}{\sigma},\end{cases}$$

$$\frac{\partial\varphi}{\partial x_2}=2x_2,$$

得

$$\begin{cases}2x_1-\left[v^{(k)}-\sigma\left(x_1-1\right)\right]=0\\ 2x_1=0\end{cases}$$

解得

$$x^{(k)}=\begin{pmatrix}x_1^{(k)}\\ x_2^{(k)}\end{pmatrix}=\begin{pmatrix}\frac{v^{(k)}+\sigma}{2+\sigma}\\ 0\end{pmatrix}.$$

由(5.53),

$$v^{(k+1)}=\max\left(0,v^{(k)}-\sigma\left(x_1^{(k)}-1\right)\right)=\frac{2\left(v^{(k)}+\sigma\right)}{2+\sigma}.$$

当$v^{(k)}<2$时，可知$\{v^k\}$为单调递增有上届的数列，必有极限。则当$k\to\infty$,

$$v^{(k)}\to 2,\quad x^{(k)}\to x^*=\begin{pmatrix}1\\ 0\end{pmatrix}.$$

所以 x^* 为极小值点，目标函数值$f^*=1$.

若在上述例题中固定罚因子$\sigma=2$，则有

$$x^{(k)}=\begin{pmatrix}x_1^{(k)}\\ x_2^{(k)}\end{pmatrix}=\begin{pmatrix}\frac{v^{(k)}+2}{4}\\ 0\end{pmatrix},$$

$$v^{(k+1)}=\frac{(v^{(k)}+2)}{2}.$$

显然对于上述序列，我们依然有

$$v^{(k)}\to 2,\quad x^{(k)}\to x^*=\begin{pmatrix}1\\0\end{pmatrix}.$$

所以说，在广义乘子法中，我们不必令σ趋向无穷大就能求得约束问题的最优解。

对于一般约束问题

$$\begin{aligned}&\min\quad f(x)\\&s.t.\quad h_i(x)=0,i\in\mathcal{E}\\&\qquad g_j(x)\geq 0,j\in\mathcal{I},\end{aligned}\tag{5.54}$$

其中$\mathcal{E}=\{1,2,..,m\}$，$\mathcal{I}=\{1,2,...,l\}$. 我们综合以上两种情形得到增广Lagrange函数

$$\begin{aligned}\varphi(x,w,v,\sigma)\ &=f(x)-\sum_{i=1}^{m}w_ih_i(x)+\frac{\sigma}{2}\sum_{i=1}^{m}h_i^2(x)\\&\quad+\frac{1}{2\sigma}\sum_{j=1}^{l}\{[\max(0,v_j-\sigma g_j(x))]^2-v_j^2\}.\end{aligned}\tag{5.55}$$

同时也可得到乘子w, v的更新

$$\begin{cases}w_i^{(k+1)}=w_i^{(k)}-\sigma h_i(x^{(k)}),\quad i=1,\cdots,m\\v_j^{(k+1)}=\max(0,v_j^{(k)}-\sigma g_j(x^{(k)})),\quad j=1,\cdots,l.\end{cases}\tag{5.56}$$

记$c(x)=\|h(x)\|+\|\max\{0,-g(x)\|$，则计算步骤为

算法 5.10 (一般约束问题的广义乘子罚函数法)

1: 给定初点$x^{(1)}$，初始乘子向量$v^{(1)}$, 初始罚因子σ_1, 放大系数$\gamma>1$, 常数$\beta\in(0,1)$, 允许误差$\varepsilon>0$，置$k=1$.

2: 以$x^{(k)}$为初始点，固定$v=v^{(k)}$求解无约束问题(5.55)得最优解$x^{(k+1)}$.

3: 如果$c(x^{(k+1)})<\varepsilon$，则停止计算，$x^{(k+1)}$为约束问题(5.43)的近似最优解. 否则，进行步骤4.

4: 若$\frac{c(x^{(k+1)})}{c(x^{(k)})}\geq\beta$, 令$\sigma_{k+1}=\gamma\sigma_k$; 否则转步骤5; 否则，进行步骤5.

5: 用(5.56)更新乘子,令$k=k+1$, 转步骤2.

5.5 组合优化问题

当最优化问题中的可行域是有限个元素组成的集合时， 该最优化问题称为组合

优化问题. 通常组合优化问题可表示为

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) \geq 0, \\ & x \in D = \{x_1, x_2, \cdots, x_n\}. \end{aligned}$$

现实生活中大量问题是组合优化问题， 典型的组合优化问题有旅行商问题， 背包问题等。

旅行商问题（traveling salesman problem, TSP）

设有 n 个城市 $1, 2, \cdots, n$， 城市 i 与城市 j 间的距离为 d_{ij}. 一售货商要去这些城市推销货物， 他希望从一城市出发后走遍所有的城市且旅途中每个城市只经过一次， 最后回到起点. 选择一条路径使得售货商所走路线总长度最短， 这就是旅行商问题。

引进决策变量 x_{ij}，若商人从城市 i 出来后紧接着到城市 j， 则 $x_{ij} = 1$，否则 $x_{ij} = 0$ $(i, j = 1, 2, \cdots, n)$。 那么 TSP 的数学模型可表示为

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} x_{ij} \\ \text{s.t.} \quad & \sum_{j=1}^{n} x_{ij} = 1, \quad i = 1, 2, \cdots, n, && (5.57) \\ & \sum_{i=1}^{n} x_{ij} = 1, \quad j = 1, 2, \cdots, n, && (5.58) \\ & \sum_{i,j \in S} x_{ij} \leq |S| - 1, \quad S \text{ 为 } \{1, 2, \cdots, n\} \text{ 的非空真子集}, && (5.59) \\ & x_{ij} \in \{0, 1\}, \quad i, j = 1, 2, \cdots, n, i \neq j. \end{aligned}$$

其中 $|S|$ 表示集合 S 中元素的个数，式 (5.57) 表示商人从城市 i 出来恰好一次，式 (5.58) 表示商人恰好进入城市 j 一次，式 (5.59) 表示商人在任何一个城市真子集中不形成回路。 因为 $D = \{0, 1\}^{n \times (n-1)}$，可见旅行商问题是组合优化问题。 当 $d_{ij} = d_{ji}$ 时，为对称距离 TSP， 否则为非对称距离 TSP。

背包问题（knapsack problem）

设有一个容量为 b 的背包， n 个容积分别为 w_i， 价值分别为 c_i $(i = 1, 2, \cdots, n)$ 的物品， 选择那些物品放入背包中以使装入的物品总价值最大， 这就是背包问题。

引入决策变量 x_i，若第 i 个物品被放入包中，则 $x_i = 1$，否则 $x_i = 0$ $(i = 1, 2, \cdots, n)$。那么背包问题的数学模型为

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} c_i x_i \\ \text{s.t.} \quad & \sum_{i=1}^{n} w_i x_i \le b, \\ & x_i \in \{0,1\}, \quad i = 1, 2, \cdots, n. \end{aligned}$$

因为 $D = \{0,1\}^n$，可见背包问题是组合优化问题。

计算复杂性

由于组合优化问题中的可行域是有限集，所以从理论上来讲，可以将这有限个可行解枚举出来，一一地计算出他们对应的目标值，然后通过比较大小找出最优解。对于小规模的组合优化问题，用这种方法很容易求出最优解。但对于大规模或稍大规模的组合优化问题，这种求解方法（枚举法）就不一定可行了。

如考虑非对称距离 TSP，用城市序号的一个排列来表示其可行解。固定一个城市为起终点，那么就有 $(n-1)!$ 个可行解。若把列出一种方案作为一次基本操作，则需 $(n-1)!$ 次基本操作。用每秒执行一千万次操作的计算机来运算，当 $n = 18$ 时，至少需要 1.13 年才能完成这些操作找出最优解。当 $n = 19$ 时，至少需要 20.3 年才能完成这些操作找出最优解。当 $n = 20$ 时，至少需要 385.7 年才能完成这些操作找出最优解。当 $n = 21$ 时，至少需要 7714.6 年才能完成这些操作找出最优解。这是不可实现的。

如何才能给一个问题设计一个有效的算法呢？这就要了解计算的复杂性。一个算法是针对一个问题来设计的，但对计算机来说，算法是以计算机语言实现的，它只能对问题的参数给定了具体数值后的实例进行求解。所以先要讨论“问题”和“实例”等基本概念。

定义 5.7 问题中的参数赋予了具体值的例子称为问题的实例。一个数在计算机中存储时占据的位数称为这个数的规模，用 $l(x)$ 表示数 x 的规模。一个实例中所有参数数值的规模之和称为这个实例的规模，用 $l(I)$ 表示实例 I 的规模。

例 5.15 一个正整数 $x \in [2^s, 2^{s+1})$ 的二进制展开为

$$x = a_s 2^s + a_{s-1} 2^{s-1} + \cdots + a_1 2 + a_0,$$

$$(a_s = 1, a_i \in \{0,1\}, i = 0, 1, \cdots, s-1),$$

那么 x 在计算机中占据 $s+1$ 位空间。所以一个整数 x (在研究计算复杂性时，为

方便起见，可只限定考虑整数) 的规模为（包含一个符号位和一个数据分隔位）

$$l(x) = \lceil \log_2(|x|+1) \rceil + 2$$

其中 $\lceil x \rceil$ 表示不小于 x 的最小整数。

例 5.16 TSP 的任何一个实例由城市数 n 和城市间的距离 $D = \{d_{ij} \mid 1 \le i,j \le n, i \ne j\}$ 确定。那么 TSP 的任何一个实例 I 的规模为

$$l(I) = \lceil \log_2(n+1) \rceil + 2 + \sum_{i=1}^{n} \sum_{j=1, j\ne i}^{n} \{\lceil \log_2(|d_{ij}|+1) \rceil + 2\}$$

因为

$$\log_2 x < \lceil \log_2(x+1) \rceil \le 1 + \log_2 x,$$

所以

$$2n(n-1) + 2 + \log_2 |P| < l(I) \le 3n(n-1) + 3 + \log_2 |P| \tag{5.60}$$

其中，$P = n \prod_{d_{ij}\ne 0, i\ne j} d_{ij}$。

记问题的实例为 I，实例规模为 $l(I)$，算法 A 在求解 I 时的计算量（算法求解中的加、减、乘、除、比较、读和写磁盘等基本运算的总次数）为 $C_A(I)$。当存在多项式函数 $g(x)$ 和一个常数 α，使得

$$C_A(I) \le \alpha g(l(I)), \tag{5.61}$$

则记 $C_A(I) = O(g(l(I)))$。

定义 5.8 对给定的问题和求解算法 A，若存在多项式函数 $g(x)$ 使得 $C_A(I) = O(g(l(I)))$ 对问题的所有实例 I 成立，则称算法 A 为解决对应问题的多项式时间算法。否则称算法 A 为指数时间算法。

注 5.1 单纯形算法不是求解线性规划问题的多项式时间算法，枚举算法不是求解 TSP 的多项式时间算法。

以上考虑了计算复杂性的一个方面，就是算法的复杂性。计算复杂性的另一个方面是问题的复杂性。

定义 5.9 对给定的问题，若存在多项式时间算法，则该问题称为多项式时间可解问题，或简称为多项式问题。所有多项式问题的集合记为 P。

Khachian 构造了椭球算法并证明其是求解线性规划问题的多项式时间算法。因此，线性规划问题是多项式问题。

如果一个问题的每一个实例只有“是” 或 “否” 两种答案， 则这个问题为判定问题。 称有肯定答案的实例为“是”实例， 否定答案的实例为“否”实例。

TSP 对应的判定问题： 用 n 个城市的一个排列 $(i_1, i_2, \cdots, i_n)$ 表示商人从城市 i_1 出发依次通过 $i_2, \cdots, i_n$, 最后返回 i_1 这样一个路径。 判定问题为： 给定 z， 是否存在 n 个城市的一个排列 $W = (i_1, i_2, \cdots, i_n)$, 使得

$$f(W) = \sum_{j=1}^{n} d_{i_j i_{j+1}} \leq z? \tag{5.62}$$

其中 $i_{n+1} = i_1$。 满足 $f(W) \leq z$ 的一个排列 W 称为对应判定问题的一个可行解。

定义 5.10 *若存在一个多项式函数 $g(x)$ 和一个验证算法 A, 对一类判定问题的任何一个“是”的判定实例 I 都存在一个字符串 S 是 I 的“是”回答, 其规模满足 $l(S) = O(g(l(I)))$, 且算法 A 验证 S 为实例 I 的“是”答案的计算量为 $O(g(l(I)))$, 则称这个判定问题是非确定多项式的, 简记为 NP。*

根据 (5.60)， TSP实例的规模 $l(I)$ 至少是 $2n(n-1) + 2 + \log_2 |P|$。 TSP 实例的解是 $(1, 2, \cdots, n)$ 的一个排列 $W = (i_1, i_2, \cdots, i_n)$， 因此该字符串的规模不超过 $\sum_{i=1}^{n}(\lceil \log_2(i+1) \rceil) + 2n < 3n + n\log_2 n = n(3 + \log_2 n)$。 而验证 W 是否满足 (5.62) 有 n 个加法和一个比较， 计算量为 $n+1$。 取多项式 $g(x) = x$， 由定义可知， TSP 属于 NP。

称判定问题 Q_1 多项式归约为 Q_2， 如果存在 Q_1 的算法和多项式函数 $g(x)$, 算法求解 Q_1 任何实例 I 的过程中， 将 Q_2 的算法作为子程序多次调用。 如果将一次调用 Q_2 算法看成一个单位 （1次基本计算量）， 则这个 Q_1 算法的计算量为 $O(g(l(I)))$。

定义 5.11 *如果判定问题 $Q \in$NP 且 NP 中的任何一个问题都可在多项时间内归约为 Q, 则称 Q 为 NP完全 (简记为 NP-C)。 若 NP 中的任何一个问题都可在多项式时间归约为判定问题 Q， 则称 Q 为 NP难 (简记为 NP-hard)。*

TSP 与背包问题均是 NP-C。 一般地认为， P∩NP-C= $\emptyset$, P≠NP。 所述四类问题的关系可见图 5.2。

目前人们普遍认为 NP-C 或 NP-hard 问题不存在求最优解的多项式时间算法， 转而寻找一些启发式算法来求解， 但启发式算法不能保证可以求得最优解， 甚至不能预估求到的近似解与最优解间的误差。 然而启发式算法能在可接受的计算费用 (指计算时间， 占用空间等) 内寻得最好解。 以下我们将要介绍的模拟退火算法与遗传算法就是属于启发式算法， 也称为智能优化算法。 由于智能优化算法是依概率导向性随机搜索算法， 所以这些算法均要编写相应的程序借助计算机实现计算过程。 其中随机数的产生是必要的手段， 如需要从区间 $[0, 1]$ 中随机产生一数。 以 C 语言

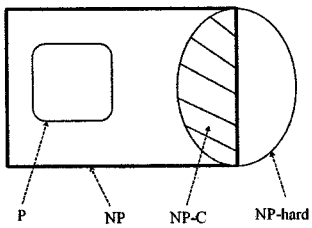


图 5.2: 四类问题的关系图

为例，从区间 $[a,b]$ 中随机产生一数的步骤为： $u=\mathrm{rand}()$; $u:=u/\mathrm{RAND\_MAX}$; 返回 $a+u(b-a)$。 其中符号 ":=" 表示将后者的值赋予前者。

5.6 模拟退火算法

模拟退火算法 (SA) 是一种导向性随机搜索的启发式算法， 它是受加热金属的退火规律所启发而提出的一种求解组合优化问题的逼近算法。 这个规律就是， 在某个温度下, 金属分子停留在能量小的状态的概率比停留在能量大的状态的概率要大。

SA 在求复杂优化问题时已显示出其非常好的有效性。 自 Kirkpatrick 等于1983将 Metropolis 在1953年提出的模拟退火思想应用到组合优化问题以来， 受到大家的普遍关注。

5.6.1 受热金属物体分子状态分布

我们先来分析金属物体分子加热后的状态。 在温度 t 下, 金属物体的分子呈现出不同的状态, 停留在状态 r 满足 Boltzmann 概率分布

$$P\{\overline{E}=E(r)\}=\frac{1}{Z(t)}\exp\left(-\frac{E(r)}{kt}\right),$$

其中, $E(r)$ 表示分子在状态 r 的能量, $k>0$ 为 Boltzmann 常数, 而 $\overline{E}$ 表示分子能量的一个随机变量。 设分子状态空间 U 是有限的， 那么 $Z(t)$ 应为

$$Z(t)=\sum_{s\in U}\exp\left(-\frac{E(s)}{kt}\right).$$

分子运动规律

根据 Boltzmann 概率分布，受热金属物体分子将依概率呈一定的规律性运动。

1. 温度 t 很高时，金属物体的分子停留在任何状态的概率近似相等。 由于

$$P\{\overline{E}=E(r)\}=\frac{\exp\left(-\frac{E(r)}{kt}\right)}{\sum\limits_{s\in U}\exp\left(-\frac{E(s)}{kt}\right)}\to\frac{1}{|U|},\quad t\to\infty,$$

其中 $|U|$ 是状态空间 U 中状态的个数。 所以温度 t 很高时，金属物体的分子停留在任何状态的概率几乎是一样的。

2. 在同一温度 t 下，金属物体的分子停留在能量低的状态比在能量高的状态的概率大。 对于 $E_1<E_2$，因为

$$P\{\overline{E}=E_1\}-P\{\overline{E}=E_2\}=\frac{1}{Z(t)}\exp\left(-\frac{E_1}{kt}\right)\left[1-\exp\left(-\frac{E_2-E_1}{k_Bt}\right)\right],$$

有

$$P\{\overline{E}=E_1\}-P\{\overline{E}=E_2\}>0,\quad\forall t>0.$$

也就是说在同一温度 t 下，金属物体的分子停留在能量低的状态比在能量高的状态的概率大。

3. 分子在能量最低状态的概率关于温度 t 下降。 设 $r_{\min}$ 是在温度 t 时分子能量最低的状态，而 U_0 是具有最低能量的状态集合。 因为

$$\begin{aligned}&\frac{\partial P\{\overline{E}=E(r_{\min})\}}{\partial t}\\&=-\left[\sum_{s\in U}\exp\left(-\frac{E(s)-E(r_{\min})}{kt}\right)\right]^{-2}\\&\quad\times\sum_{s\in U\setminus U_0}\exp\left(-\frac{E(s)-E(r_{\min})}{kt}\right)\frac{E(s)-E(r_{\min})}{kt^2}\\&<0.\end{aligned}$$

所以 $P\{\overline{E}=E(r_{\min})\}$ 是 t 的递减函数。

4. 分子停留在最低能量状态的概率随温度降低趋于 1。 当 $t\to0$，

$$\begin{aligned}P\{\overline{E}=E(r_{\min})\}&=\frac{\exp\left(-\frac{E(r_{\min})}{kt}\right)}{\sum\limits_{s\in U}\exp\left(-\frac{E(s)}{kt}\right)}\\&=\frac{1}{\sum\limits_{s\in U_0}1+\sum\limits_{s\in U\setminus U_0}\exp\left(-\frac{E(s)-E(r_{\min})}{kt}\right)}\\&\to\frac{1}{|U_0|},\end{aligned}$$

那么

$$P\{\cup_{r\in U_0}\{\overline{E}=E(r)\}\}\to 1,\quad t\to 0.$$

故分子停留在最低能量状态的概率趋于 1。

5. 分子在非能量最低状态的概率随温度降低趋于 0。这从 Boltzmann 概率分布易知。

可知，温度越低，分子在能量越低的状态的概率值越高。在极限状况，只有在能量最低的状态概率不为零。

例 5.17 设金属分子的状态概率分布为

$$P\{\overline{E}=E(r)\}=\frac{1}{Z(t)}\exp\left(-\frac{(r-2)^2+1}{2t}\right)$$

其中状态取为 $r=2,3,4,5,6$, 而

$$Z(t)=\sum_{r=2}^{6}\exp\left(-\frac{(r-2)^2+1}{2t}\right).$$

分子停留在各状态的概率随时间的变化情况可由图 5.3 形象地描述出，它们符合上述对 Boltzmann 概率分布的分析。

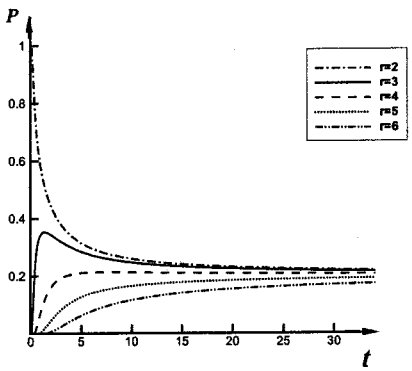


图 5.3: Boltzmann 函数曲线

5.6.2 基本模拟退火算法

在前一小节我们分析了对金属物体加热降温过程中分子所处状态的规律，随着温度从一定高度逐渐降低趋于零度，分子停留在能量最低状态的概率趋于 1。将优化问题的可行解对应于分子的状态，优化问题中的目标函数 $f(x)$ 对应于分子的状态能量，那么通过构造模拟的降温过程，可将金属物体分子状态概率分布规律应用于求解优化问题，这一算法称为模拟退火算法，其基本步骤如下：

算法 5.11 (基本模拟退火算法)

1: 初始化可行解和温度.
2: 根据 Boltzmann 概率退火.
3: 重复第 2 步直到稳定状态（内循环）.
4: 降温.
5: 重复第 2 步至第 4 步直到 满足终止条件或直到给定的步数（外循环）.
6: 输出最好的解作为最优解.

5.6.3 模拟退火算法实现技术

本小节主要介绍基本模拟退火算法各步骤的具体操作方法。

1. *初始化过程*: 要预先确定一个初始可行解，从这个解出发搜索下一个可行解。此外还要确定一个初始温度，以这个温度开始实施逐步降温过程。一般地，初始可行解 s_0 可根据问题随机产生。而初始温度 t_0 理论上要求应保证平稳分布中产生任意可行解的概率相等，即 $\exp(-\Delta f_{ij}/t_0) \approx 1$, 其中 $\Delta f_{ij} = f(s_j) - f(s_i)$。如可取 $t_0 = K\Delta_0$, K 为充分大的数, 而

$$\Delta_0 = \max\{f(s_i) \mid s_i \in D\} - \min\{f(s_i) \mid s_i \in D\}.$$

初始温度也可以如下启发式地产生：

2. 退火: 退火过程就是在一给定温度下，由一个状态变到另一个状态，每一个状态到达的次数服从一个概率分布，即基于 Metropolis 接受准则的过程，该过程达到平稳时停止。在状态 s_i时, 产生的状态 s_j 被接受的概率为

$$A_{ij}(t) = \begin{cases} 1, & \text{如果 } f(s_i) \geq f(s_j), \\ \exp\left(-\frac{\Delta f_{ij}}{t}\right), & \text{如果 } f(s_i) < f(s_j). \end{cases}$$

注意到，算法是从一个可行解转移到另一个可行解，而另一个可行解一般是从当前可行解的邻域中产生。邻域的构造与解的表达形式有关，应简洁明了易于操作，邻域中每个邻居都是可行解，解空间中任何两状态可达。

算法 5.12 (初始温度算法)

1: 给定一个常数 T, 温度 t_0, 接近于 1 的常数 c_0, $R_0 = 0$.

2: 在这个温度退火 L 步 （退火过程以下再介绍）. 记接受状态的个数为 L', 计算 $R_k = L'/L$.

3: 如果 $|R_k - c_0| < \varepsilon$, 停止. 否则, 如果 $R_{k-1}, R_k < c_0$, 则 $k := k+1$, $t_0 := t_0 + T$, 返回步骤 2; 如果 $R_{k-1}, R_k \geq c_0$, 则 $k := k+1$, $t_0 := t_0 - T$, 返回步骤 2; 如果 $R_{k-1} \geq c_0, R_k \leq c_0$, 则 $k := k+1$, $t_0 := t_0 + T/2$, 返回步骤 2; 如果 $R_{k-1} \leq c_0, R_k \geq c_0$, 则 $k := k+1$, $t_0 := t_0 - T/2$, 返回步骤 2.

算法 5.13 (退火算法)

在温度 t, 有了可行解 s_i时, 对另一个可行解 s_j, 如果 $\Delta f_{ij} = f(s_j) - f(s_i) \leq 0$, 则 $s_i := s_j$. 否则, 如果 $\exp(-\Delta f_{ij}/t) > \text{random}(0,1)$, 则 $s_i := s_j$.

例 5.18 对 TSP 问题， 用城市的一个排序表示一个可行解。 解的邻域可用不同的操作算子定义， 如 互换操作：即随机交换解码中两个不同的字符位置。 逆序操作：即将解码中两不同的随机位置间的字符串逆序。 插入操作：即随机选择某个位置的字符插入到串中的另一个随机位置。

如果邻域中有不是可行解的邻居， 可用罚值法，将其视为可行解， 目标值为一个充分大的数。 但该法的缺陷是扩大了搜索区域， 从而使计算时间增加。

3. 降温: 一种降温方法为

$$t_{k+1} = d(t_k),$$

其中 $d(t_k) = \alpha t_k$。 另一种降温方法为

$$t_k = \frac{M-k}{M} t_0$$

其中 M 为温度下降的总次数。

4. 内循环终止准则： 常用的有 (1) 固定步数：即在每一温度迭代相同的步数。 (2) 由接受和拒绝的比率控制迭代步数：给定一个迭代步数上限 U 和一个接受次数指标 r， 在温度 t 实施退火过程，当接受次数等于 r 时， 不再迭代，否则一直迭代到步数上限 U。 或者给定一个接受指标 R 和迭代步数下限 L，在温度 t 实施退火过程，迭代到步数 L 时，开始计算接受次数与总次数的比率，一旦比率超过 R，不再迭代，否则一直迭代到步数上限 U。 同样可以用拒绝次数控制终止准则。

5. 外循环终止准则： 常用的有 (1) 设置终止温度的阈值（比较小的正数）$\varepsilon > 0$，当温度下降到 $t_k < \varepsilon$ 时，算法停止。 (2) 设置循环总数。 迭代次数达到指定数目时，算法停止。 (3) 基于不改进规则。 若连续若干步搜索到的最优解不再改进，算法停止。 (4) 设置接受概率。 给定指标 $\chi > 0$ 是一个比较小的数，在温度 t，除

局部最优解外，其他状态的接受概率均小于 χ，算法停止。

5.7 遗传算法

遗传算法 (GA) 是一种解优化问题的导向随机搜索方法，它模拟生物在自然进化中的选择和遗传（即适者生存）规律而提出来的全局优化算法。遗传算法的思想和基本概念最早由美国密执根大学的 Holland 教授于20世纪70年代提出来的。80年代 De Jong 和 Goldberg 等学者进一步地完善了遗传算法的理论和方法。遗传算法不仅在求解组合优化问题时显示出优越性，而且也普遍被应用于求解连续型优化问题。

遗传算法在解优化问题中的构成要素:

染色体: 解的编码（字符串，向量等）
基因: 解的编码中每一分量的特征（如各分量的值）
适应性: 适应函数值
群体: 选定的一组解（其中解的个数为群体的规模 *pop_size*）
种群: 根据适应函数值选取的一组解（其中种群的规模与群体的规模可相同也可不相同）
交叉: 通过交叉原则产生一组新解的过程
变异: 编码的某一些分量发生变化的过程

5.7.1 基本遗传算法

基本遗传算法可描述为如下步骤:

算法 5.14 (基本遗传算法)

1: 随机初始化 *pop_size* 个染色体.
2: 用交叉算法更新染色体.
3: 用变异算法更新染色体.
4: 计算所有染色体的目标值.
5: 根据目标值计算每个染色体的适应度.
6: 通过轮盘赌的方法选择染色体.
7: 重复第 2 至第 6 步直到终止条件满足.
8: 输出最好的染色体作为最优解.

为利于遗传算法的计算，首先要对解进行编码，编码后的解称为染色体。对

于约束优化问题，遗传算法是在染色体中进行操作，而把操作结果解码后去检验其可行性。

5.7.2 遗传算法实现技术

上小节我们叙述了遗传算法的基本步骤，该算法中涉及到的运算技术实现问题由本小节来阐述。

1. *编码与解码*: GA 的关键问题之一是把解编码为染色体，也要能把染色体解码为解。常用的编码方法有

(1) 二进制码：就是 $0-1$ 编码。采用 $0-1$ 码可以精确地表示整数。用 $0-1$ 编码精确表示 a 到 b 所有整数，只需编码长度 n 满足 $\frac{b-a}{2^n}<1$, 即 $n>\log_2(b-a)$。如满足 $0\le x\le 31$ 的整数 x 只需 5 位数的二进制码，如

$$10000\to 16 \quad 11111\to 31 \quad 01001\to 9 \quad 00010\to 2$$

(2) 根据问题确定的编码: 如 TSP，可用城市编号的一个序列来表示可行解，对 8 个城市的 TSP，用数字 $1\sim 8$ 分别表示 8 个城市，那么这 8 个数字的任意一个排列就是一个可行解的编码，如 27658134。

(3) 实数码: 对于连续的实数变量，可以采用实数编码。实数编码可以用实数本身作为实数码，也可以在解空间与码空间中作一个对应关系。如, 设 (x_1,x_2,x_3) 是以下解空间中的向量

$$\begin{cases} x_1+x_2^2+x_3^3=1 \\ x_1\ge 0,\ x_2\ge 0,\ x_3\ge 0. \end{cases}$$

我们可以用以下码空间中的染色体 (v_1,v_2,v_3) 来对解编码

$$v_1\ge 0, \quad v_2\ge 0, \quad v_3\ge 0,\ v_1,v_2,v_3\text{不同时为零.}$$

那么编码与解码的过程可以体现在以下关系上

$$x_1=\frac{v_1}{v_1+v_2+v_3},\ x_2=\sqrt{\frac{v_2}{v_1+v_2+v_3}},\ x_3=\sqrt[3]{\frac{v_3}{v_1+v_2+v_3}}.$$

连续的实数变量在一定精度下也可以采用二进制编码。对给定的区间 $[a,b]$, 设二进制编码的长为 n, 则变量

$$x=a+a_1\frac{b-a}{2}+a_2\frac{b-a}{2^2}+\cdots+a_n\frac{b-a}{2^n}$$

与二进制码 $a_1a_2\cdots a_n$ 相对应。二进制码与实际变量的误差为 $\frac{b-a}{2^n}$。

2. *初始化群体*: 假设群体的规模为 pop_size， 随机产生 pop_size 个染色体作为初始群体。 初始化的方法根据编码方式的不同而设计。

对二进制编码，随机产生 n 个整数 0 或 1 组成长度为 n 的二进制编码染色体。

对根据问题确定的编码，如用城市序号的排列编码 TSP 的可行解，可用如下方式随机产生: 先有 $u[i]=i,\ i=1,2,\cdots,n$; 令 $r=(int)\text{random}[1,n]$, 那么 $x[1]:=u[r]$; 更新 $u[i]:=u[i]\ (i=1,2,\cdots,r)$ 与 $u[i]:=u[i+1]\ (i=r,r+1,\cdots,n-1)$; 令 $r=(int)\text{random}[1,n-1]$, 那么 $x[2]:=u[r]$; 依此类推，可得到一个染色体 $x=x[1]x[2]\cdots x[n]$。

对于实数编码， 可预估一个含有最优解的区域，如立方体 $\prod_{i=1}^{n}[a_i,b_i]$。 在这个区域中随机产生一个点，

$$x=(x_1,x_2,\cdots,x_n),\quad x_i=\text{random}[a_i,b_i],\ i=1,2,\cdots,n.$$

然后检验这个解的可行性。 如是可行的，则接受为染色体； 如不可行， 则重新在该区域中随机产生一个点直到是可行点为止。 重复刚才的过程 pop_size 次得到 pop_size 个初始染色体。

3. *群体的规模*: 群体的规模可以设定为个体编码长度数的一个线性倍数； 群体的规模可以是一个给定数； 群体的规模也可以是变化的，当连续多代没能改变解的性能， 则可扩大群体的规模， 若解的改进非常好， 则可以减少群体的规模。

4. *评价函数*: 评价适应函数 $Eval(V)$ 是根据每个染色体 V 的目标值而赋予的数值，体现其与其他染色体的相对重要程度， 可用它来决定该染色体被选为种群的概率。 设可行解 x 编码后的染色体记为 V。 构造评价函数常用的方法有:

(1) 简单地基于目标函数。 设 $f(x)\geq 0$ 为目标函数， 则

$$Eval(V)=f(x),\qquad \max \text{ 优化问题}$$
$$Eval(V)=M-f(x),\quad \min \text{ 优化问题},\ M>\max_{x} f(x).$$

(2) 基于非线性加速函数。 取

$$Eval(V)=\begin{cases}\dfrac{1}{f_{max}-f(x)}, & f(x)<f_{max}\\ M, & f(x)=f_{max}\end{cases}$$

其中 $M>0$ 是一个充分大的数, f_{max} 是当前的最优目标值。

(3) 基于线性加速适应函数:

$$Eval(V)=\alpha f(x)+\beta$$

其中 α, β 满足

$$\begin{cases} \alpha = \dfrac{\sum\limits_{i=1}^{pop\_size} f(x_i)}{pop\_size} + \beta = \dfrac{\sum\limits_{i=1}^{pop\_size} f(x_i)}{pop\_size} \\ \alpha \max\limits_{1 \le i \le pop\_size} \{f(x_i)\} + \beta = M \dfrac{\sum\limits_{i=1}^{pop\_size} f(x_i)}{pop\_size} \end{cases}$$

(4) 基于序的评价函数: 设 $a \in (0,1)$, $b > 0$ 取

$$Eval(V_i) = b(1-a)^{i-1}, \qquad i = 1, 2, \cdots, pop\_size.$$

注意: $i = 1$ 代表最好的个体, $i = pop\_size$ 代表最坏的个体。

5. 选择过程: 根据评价函数选取一批染色体作为种群, 选取的方法一般采用轮盘赌方式, 评价值大的染色体被选为种群的可能性就大, 在该过程中一个染色体可以允许多次被选上, 所以有的参考书也把选择过程称为复制过程。选择过程可用以下算法描述:

算法 5.15 (轮盘赌的选择过程)

1: 计算所有染色体 V_i 的累积概率 q_i,

$$q_0 = 0, \quad q_i = \sum_{j=1}^{i} Eval(V_j), \quad i = 1, 2, \cdots, pop\_size.$$

2: 在 $(0, q_{pop\_size}]$ 中产生一个随机数 r.

3: 若 $q_{i-1} < r \le q_i$, 则选择染色体 V_i.

4: 重复第 2 至第 3 步 pop_size 次以获得 pop_size 个染色体.

6. 交叉运算: 交叉运算是遗传算法的核心步骤。种群中的染色体以一定的概率 P_c (常称为交叉概率) 被选来作交叉运算产生新的染色体, 被选中作交叉运算的染色体称为父代, 可见大约有 $P_c \cdot pop\_size$ 个被选到的父代, 将被选到的父代两两配对。即从种群中第 $i = 1$ 个染色体到第 pop_size 个染色体逐个地如下操作: 从 $[0,1]$ 中随机产生 r, 如 $r < P_c$, 则染色体 V_i 被选为父代。记父代为 $V_1', V_2', V_3', \cdots$, 将他们配成对 (V_1', V_2'), (V_3', V_4'), (V_5', V_6'), $\cdots$。两个父代通过交叉运算产生的新染色体称为其后代, 在原种群中用后代替换父代形成新的种群或群体。

(1) 用二进制编码时常用的交叉运算方法有:

单交叉位法. 如设两个父代为 11010011 与 01110100, 随机选中交叉位为第 3 位, 那么两个父代的前 3 位保持不变, 第 4 位以后的基因相交换, 产生两个后代

$$\left.\begin{matrix} 110|10011 \\ 011|10100 \end{matrix}\right\} \Longrightarrow \left\{\begin{matrix} 110|10100 \\ 011|10011 \end{matrix}\right.$$

多交叉位法. 对码长较大的染色体，交叉位可以设定 2 个或 2 个以上，这些交叉位将父代分割成若干段，各段间隔进行交换后，产生后代。如设两个父代为 1010110100110010 与 1100011101001011，如随机确定了第 6 和第 10 位是交叉位，那么将父代的第 6 和第 10 位之间的第 7 至第 10 位交换，其它位保持不变，产生两个后代

$$\left.\begin{array}{l}101011|0100|110010\\110001|1101|001011\end{array}\right\}\Longrightarrow\left\{\begin{array}{l}101011|1101|110010\\110001|0100|001011\end{array}\right.$$

如确定第 2，5，9，13 位是交叉位，那么将父代的第 1，2 位保持不变，第 3，4，5 位相交换，第 6 至第 9 位保持不变，第 10 位至第 13 位相交换，第 14，15，16 位保持不变，产生两个后代

$$\left.\begin{array}{l}10|101|1010|0110|010\\11|000|1110|1001|011\end{array}\right\}\Longrightarrow\left\{\begin{array}{l}10|000|1010|1001|010\\11|101|1110|0110|011\end{array}\right.$$

(2) 对于根据问题确定的编码，交叉运算可用如下之方法:

常规不变位法. 随机确定一个交叉位，父代交叉位前的基因分别继承给两个后代，两后代交叉位之后的基因分别按对方父代基因顺序选取不重基因。如确定第 4 位是交叉位，那么

$$\left.\begin{array}{ll}\text{父代A} & 2864|7513\\\text{父代B} & 5467|8312\end{array}\right\}\Longrightarrow\left\{\begin{array}{ll}\text{后代A} & 2864|5731\\\text{后代B} & 5467|2813\end{array}\right.$$

后代 A 继承了父代 A 在交叉位 4 之前的基因 2864，然后以父代 B (54678312) 依序选取不重复的基因 5，7，3，1。后代 B 同样的方式产生。

随机不变位法. 随机产生一个与染色体相等维数的二进制编码，其中 1 所在位置对应父代相应位置的基因不变，其它基因位按前述方法处理。如随机产生 8 位基因的二进制编码 01100101，那么

$$\left.\begin{array}{ll}\text{父代A} & 2\underline{8}6\underline{4}7\underline{5}\underline{1}\underline{3}\\\text{父代B} & 5\underline{4}6\underline{7}8\underline{3}\underline{1}\underline{2}\end{array}\right\}\Longrightarrow\left\{\begin{array}{ll}\text{后代A} & 4\underline{8}6\underline{7}1\underline{5}\underline{2}\underline{3}\\\text{后代B} & 8\underline{4}6\underline{7}5\underline{3}\underline{1}\underline{2}\end{array}\right.$$

后代 A 在第 2，3，6，8 位的基因继承父代 A 的基因分别为 8，6，5，3，而第 1，4，5，7位的基因依前述方法以父代 B (54678312) 依序选取不重复的基因。后代 B 同样的方式产生。

(3) 对于实数码，可按以下方法实现交叉运算：随机产生一个数 $c\in(0,1)$，由父代 V_1, V_2 产生后代 X, Y 为

$$X=c\cdot V_1+(1-c)\cdot V_2,\quad Y=(1-c)\cdot V_1+c\cdot V_2.$$

7. 变异运算: 变异运算是遗传算法的另一显著特征，种群中的染色体以一定的概率 P_m （常称为变异概率） 被选来作变异运算产生新的染色体，可见大约有

$P_m \cdot pop\_size$ 个被选到的父代。即从种群中第 $i=1$ 个染色体到第 pop_size 个染色体逐个地如下操作: 从 $[0,1]$ 中随机产生 r, 如 $r<P_m$, 则染色体 V_i 被选为父代。将被选到的父代通过变异运算产生其后代, 在原种群中用后代替换父代形成新的种群或群体。常用的变异运算方法有:

(1)单点、多点变异法. 对二进制编码随机选一个或多个变异位, 将父代在变异位的基因由 1 变为 0 或由 0 变为 1, 产生后代, 如

$$11\underline{0}11011 \Longrightarrow 11\underline{1}11011 \quad \text{(变异位 3)}$$

$$11\underline{0}11\underline{0}11 \Longrightarrow 11\underline{1}11\underline{0}01 \quad \text{(变异位 3, 7)}$$

(2) 2-opt 法. 随机产生两个变异位, 然后将两变异位间的基因按逆序排, 其它位的基因保持不变, 产生后代。如若变异位为 2 和 6, 那么将第 2 与第 6 位间的基因倒排获得后代, 如

$$1\underline{0110}100 \Longrightarrow 1\underline{1011}000 \quad \text{(二进制编码)}$$

$$2\underline{5831}746 \Longrightarrow 2\underline{7138}546 \quad \text{(TSP的城市序号编码)}$$

(3) 对实数码, 可如下进行变异操作。设 V 为父代, 取 $M>0$ 适当大, 随机选个变异方向 $\boldsymbol{d}$, 如 $V+M\cdot\boldsymbol{d}$ 不可行, 置 M 为 0 到 M 间的随机数直到可行为止; 若该过程在规定的代数还不能找到可行解, 则置 $M=0$。由 V 变异产生的后代为

$$X = V + M\cdot\boldsymbol{d}.$$

关于遗传算法的收敛性, 有以下结论。

定理 5.18 (收敛定理) (1) 若变异概率 $0<p_m<1$, 交叉概率 $0\le p_c\le 1$, 则基本遗传算法不收敛到全局最优解。

(2) 如果改进基本遗传算法按交叉、变异、种群选取之后更新当前最优染色体(解)的进化循环过程, 则收敛于全局最优。

(3) 如果改进基本遗传算法按交叉、变异后就更新当前最优染色体 之后进行种群选取的进化循环过程, 则收敛于全局最优。

从收敛性定理可看出, 遗传算法求解最优化问题时, 每次迭代后需要更新当前最优染色体, 才能保证算法停止时, 输出的最好解近似为最优解。

第五章习题

1. 用单纯形方法求解

(1)
$$\begin{aligned} \min \quad & -3x_1 - x_2 \\ s.t. \quad & 3x_1 + 3x_2 + x_3 = 30 \\ & 4x_1 - 4x_2 + x_4 = 16 \\ & 2x_1 - x_2 \le 12 \\ & x_i \ge 0, i = 1, \cdots 4 \end{aligned}$$

(2)
$$\begin{aligned} \min \quad & 3x_1 - 5x_2 - 2x_3 - x_4 \\ s.t. \quad & x_1 + x_2 + x_3 \le 4 \\ & 4x_1 - x_2 + x_3 + 2x_4 \le 6 \\ & -x_1 + x_2 + 2x_3 + 3x_4 \le 12 \\ & x_i \ge 0, i = 1, \cdots 4 \end{aligned}$$

2. 用两阶段法求解下列问题

$$\begin{aligned} \min \quad & x_1 - 3x_2 + x_3 \\ s.t. \quad & 2x_1 - x_2 + x_3 = 8 \\ & 2x_1 + x_2 \ge 2 \\ & x_1 + 2x_2 \le 10 \\ & x_i \ge 0, i = 1, 2, 3 \end{aligned}$$

3. 用大M法求解例5.3.

4. 求解下列问题的稳定点，并判断是否是问题的局部极小解。

$$\begin{aligned} \min \quad & -3x_1 + x_2 - 4x_3^2 \\ s.t. \quad & -x_1 + x_2 + x_3^2 = 0 \\ & 1 - x_2 - x_3 \ge 0 \end{aligned}$$

5. 用最速下降法求解

$$\min \ (x_1 - 1)^2 + x_2^2,$$

初始点$x^{(1)} = (1, 1)^T$.

6. 用牛顿法求解

$$\min \ (x_1 - 1)^4 + x_2^2,$$

初始点$x^{(1)} = (0, 1)^T$, 迭代两次。

7. 用FR共轭梯度法求解例5.8中的问题，初始点仍为$(0, 0)^T$。

8. 用外点罚函数法和倒数障碍罚函数法求解如下问题

$$\begin{aligned} \min \quad & \tfrac{1}{12}(x_1 + 1)^2 + x_2 \\ s.t. \quad & x_1 + 1 \ge 0 \\ & x_2 \ge 0 \end{aligned}$$

9. 用广义乘子法求解

$$\begin{aligned} \min \quad & x_1^2 + 2x_2^2 \\ s.t. \quad & x_1 + x_2 \geq 1 \end{aligned}$$

10. 什么是 P 问题和 NP 问题？

11. 受热金属物体分子运动有什么规律，对求解组合优化问题有什么启发？

12. 什么是 Metropolis 接受准则？

13. 遗传算法中交叉运算如何进行，变异运算如何进行？

14. 轮盘赌如何实施？

15. 编写模拟退火算法求解 TSP 的程序。

16. 编写遗传算法求解 TSP 的程序。

第六章　函数逼近与数据拟合

函数逼近问题指构造一个简单函数$P(x)$来近似表示实际中的复杂函数$f(x)$，并限制其误差在某种度量意义下最小。而构造的$P(x))$函数类可以有不同的选择，即使函数类选定，用此类函数表示$f(x)$的方式仍有多种选择。同样地，近似程度（误差）的度量也可以有各种不同的含义。因此，函数逼近问题的提法具有多样的形式，其内容十分丰富。如，当取$P(x)$ 为多项式函数，并要求在某些点上$P(x)$和$f(x)$有相同的值，甚至是某阶导数值相同，就衍生出了多项式插值问题。当用小波函数表示$f(x)$ 时，对应了小波的多尺度逼近。而当函数用某类简单函数，表示的方式采用多个简单线性的叠加时，则构成了人工神经网络的基本模块。本章主要介绍几种多项式插值、最小二乘方法、人工神经网络中的BP算法以及小波变换的方法。

6.1　多项式插值

在科学研究中，经常需要研究两个变量之间的函数关系$y = f(x)$，并且这种关系通常是通过实验和观察得到$f(x)$的有限个点，

$$f(x_i) = y_i, i = 0, 1, ..., n. \tag{6.1}$$

针对这些有限个点，需要构造一个函数$y = P(x)$,使得

$$P(x_i) = y_i, i = 0, 1, ..., n. \tag{6.2}$$

$P(x)$作为$f(x)$的近似值。称$f(x)$为被插值函数，点$x_0, x_1, ...x_n$为插值节点，$P(x)$为插值函数，(6.2)称为插值条件。 在处理插值问题中，首先根据需要选择$P(x)$为哪一类函数，如果选择为代数多项式，则问题就是多项式插值问题。类似地，可选择$P(x)$为三角多项式、有理函数、样条函数等，从而衍生出不同的插值问题。下面主要介绍多项式插值。

6.1.1　Lagrange插值

多项式插值即构造一个次数不超过n次的多项式

$$P_n(x) = a_0 + a_1 + ... + a_n x^n. \tag{6.3}$$

使其满足插值条件(6.2)，即

$$\begin{cases} a_0 + a_1x_0 + ... + a_nx_0^n = f(x_0) \\ a_0 + a_1x_1 + ... + a_nx_1^n = f(x_1) \\ \cdots\ \cdots\ \cdots \\ a_0 + a_1x_n + ... + a_nx_n^n = f(x_n) \end{cases} \tag{6.4}$$

则插值多项式的存在唯一性由此多项式中系数$a_0, a_1, ..., a_n$构成的线性方程组的解的存在唯一性决定。当系数矩阵A的行列式不等于0 时，此线性方程组有唯一解。可以看到，$|A|$是一个Vandermonde行列式

$$|A| = \begin{vmatrix} 1 & x_0 & x_0^2 & ... & x_0^n \\ 1 & x_1 & x_1^2 & ... & x_1^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & ... & x_n^n \end{vmatrix}$$

当$i \neq j$时，$x_i \neq x_j$时，$|A| \neq 0$,此时方程组有唯一解，则存在唯一的插值多项式满足插值条件(6.2). 但此方法的计算量大，因此，需要寻找更有效的插值多项式方法。

针对上述的多项式插值，可以构造次数不超过n次的多项式$l_i(x)$,使其满足

$$l_i(x_j) = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

则可以看出，$x_j(i \neq j)$都是$l_i(x)$的根，且$l_i(x_i) = 1$.所以

$$l_i(x) = \frac{(x - x_0)(x - x_1)...(x - x_{i-1})(x - x_{i+1})...(x - x_n)}{(x_i - x_0)(x_i - x_1)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n)}. \tag{6.5}$$

令

$$L_n(x) = y_0l_0(x) + y_1l_1(x) + ... + y_nl_n(x). \tag{6.6}$$

容易验证，$L_n(x)$即为问题(6.2)的解。称$L_n(x)$为$f(x)$的n次Largrange插值多项式，$l_i(x)(i = 1, 2, ...n)$为n次的插值基函数。 为了记号简单，记

$$\omega(x) = \prod_{i=0}^{n}(x - x_i).$$

则 $\omega'(x_i) = \prod_{j=0, j\neq i}^{n}(x_i - x_j)$. 则$l_i(x)$可表示成

$$l_i(x) = \frac{\omega(x)}{(x - x_i)\omega'(x_i)}, \tag{6.7}$$

从而Largange插值函数表示为:

$$L_n(x)=\sum_{i=0}^{n}\frac{\omega(x)}{(x-x_i)\omega'(x_i)}f(x_i). \tag{6.8}$$

插值多项式$L_n(x)$是被插函数$f(x)$的一种近似，在节点处是精确的，但在节点之外则存在误差, 称$f(x)-L_n(x)$为插值余项，也成为截断误差，记为$R_n(x)$。

定理 6.1 *设$f(x),f'(x),...,f^{(n)}$在区间$[a,b]$上连续，$f^{(n+1)}$区间(a,b)内存在，节点$a\le x_0<x-1<...<x_n\le b$, $L_n(x)$是满足插值条件的插值多项式。则对任意的$x\in[a,b]$, 插值余项为*

$$R_n(x)=f(x)-L_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}\omega(x). \tag{6.9}$$

6.1.2 差商与Newton插值

利用上节所提到的Lagrange插值多项式进行函数逼近时，如果需要增减多项式的次数，即增减插值节点的个数时，则需要重新计算所有的$l_i(x)$，这显然造成了前面计算工作量的浪费，增加了计算量。为了可以利用前面所计算的结果,构造通过$n+1$个点的n次插值多项式

$$P_n(x)=a_0+a_1(x-x_0)+a_2(x-x_0)(x-x_1)+...+a_n(x-x_0)(x-x_1)...(x-x_{n-1}). \tag{6.10}$$

其中$a_0,a_1,...a_n$为待定系数。

依次将节点$x_0,x_1,...,x_n$代入插值条件$P_n(x_i)=f(x_i),i=0,1,...n$ 中，可得：当$x=x_0$时，$P_n(x_0)=a_0=f(x_0)$. 当$x=x_1$时，$P_n(x_1)=a_0+a_1(x_1-x_0)=f(x_1)$，则

$$a_1=\frac{f(x_1)-f(x_0)}{x_1-x_0}.$$

当$x=x_2$时，$P_n(x_2)=a_0+a_1(x_1-x_0)+a_2(x_2-x_0)(x_2-x_1)=f(x_2)$，则

$$a_2=\frac{\frac{f(x_2)-f(x_1)}{x_2-x_1}-\frac{f(x_1)-f(x_0)}{x_1-x_0}}{x_2-x_0}.$$

依次递推，可得到$a_3,a_4,...,a_n$，为了给出系数a_n的一般形式，下面引入差商的概念。

定义 6.1 *设已知函数$f(x)$在$n+1$个互异节点$x_0,x_1,...,x_n$上的函数值为 $f(x_0)$, $f(x_1)$, ..., $f(x_n)$, 称*

$$\frac{f(x_j)-f(x_i)}{x_j-x_i},$$

为函数$f(x)$在节点x_i,x_j处的一阶差商，并记为$f[x_i,x_j]$,即

$$f[x_i,x_j]=\frac{f(x_j)-f(x_i)}{x_j-x_i}.$$

表 6.1: 差商表

x_0	$f(x_i)$	一阶差商	二阶差商	三阶差商	四阶插商
x_0	$f(x_0)$	$f[x_0,x_1]$	$f[x_0,x_1,x_2]$	$f[x_0,x_1,x_2,x_3]$	$f[x_0,x_1,x_2,x_3,x_4]$
x_1	$f(x_1)$	$f[x_1,x_2]$	$f[x_1,x_2,x_3]$	$f[x_1,x_2,x_3,x_4]$	
x_2	$f(x_2)$	$f[x_2,x_3]$	$f[x_2,x_3,x_4]$		
x_3	$f(x_3)$	$f[x_3,x_4]$			
x_4	$f(x_4)$				

称一阶差商的差商为二阶差商，即

$$f[x_i,x_j,x_k]=\frac{f[x_j,x_k]-f[x_i,x_j]}{x_k-x_i}.$$

一般地，$f(x)$在$x_0,x_1,...,x_k$处的k阶差商定义为$k-1$阶差商的差商，即

$$f[x_0,x_1,...x_k]=\frac{f[x_1,x_2,...,x_k]-f[x_0,x_1,...,x_{k-1}]}{x_k-x_0}. \tag{6.11}$$

差商可用表格(表 6.1)方便地计算。

差商具有以下的性质：

性质 6.1 k阶差商可以表示为函数值$f(x_0),f(x_1),...,f(x_k)$的线性组合

$$f[x_0,x_1,...x_k]=\sum_{j=0}^{k}\frac{f(x_j)}{\omega_k'(x_j)}. \tag{6.12}$$

其中$\omega(x)=\prod_{j=0}^{k}(x-x_j)$.

性质 6.2 差商具有对称性，即在k阶差商$f[x_0,x_1,...x_k]$中，任意改变节点x_i,x_j的次序，其值不变。

性质 6.3 n次多项式$f(x)$的k阶差商$f[x,x_0,x_1,...x_{k-1}]$,当$k\le n$时，是$n-k$次多项式，当$k>n$时，其值恒等于零。

性质 6.4

$$f[x_0,x_1,...,x_k]=\frac{f^{(k)}(\xi)}{k!}.$$

其中ξ介于$x_0,x_1,...,x_n$的最小值与最大值之间。

上述差商的定义都是建立在节点互不相等的情形下，但有时我们需要用到节点相重时的差商，这时可定义

$$f[x_0,x_0]=\lim_{\Delta x\to 0}f[x,x+\Delta x]=f'(x).$$

一般地，有

$$f[x,x,x_0,...x_n]=\frac{d}{dx}f[x,x_0,...,x_n].$$

一个极端的情况是当所有的n个节点都与x_0重合时，则

$$f[\underbrace{x_0,x_0,...x_0}_{n+1}]=\frac{1}{n!}f^{(n)}(x_0).$$

有了差商定义后，我们来讨论Newton插值公式。

由插商定义,得到：

$$\begin{aligned}
&f(x)=f(x_0)+(x-x_0)f[x,x_0]\\
&f[x,x_0]=f[x_0,x_1]+(x-x_1)f[x,x_0,x_1]\\
&\qquad\cdots\cdots\\
&f[x,x_0,...,x_{n-1}]=f[x_0,x_1,...,x_n]+(x-x_n)f[x,x_0,...,x_n].
\end{aligned}$$

将以上各式右后向前依次代入，可得

$$\begin{aligned}
f(x)=&f(x_0)+(x-x_0)f[x+0,x_1]\\
&+(x-x_0)(x-x_1)f[x_0,x_1]+...\\
&+(x-x_0)(x-x_1)...(x-x_{n-1})f[x_0,x_1,...,x_n]\\
&+(x-x_0)(x-x_1)...(x-x_n)f[x,x_0,x_1,...,x_n].
\end{aligned}$$

记$N_n(x)$为

$$\begin{aligned}
N_n(x)=&f(x_0)+(x-x_0)f[x_0,x_1]\\
&+(x-x_0)(x-x_1)f[x_0,x_1]+...\\
&+(x-x_0)(x-x_1)...(x-x_{n-1})f[x_0,x_1,...,x_n].
\end{aligned}\tag{6.13}$$

以及$\tilde{R}_n(x)$为

$$\tilde{R}_n(x)=(x-x_0)(x-x_1)...(x-x_n)f[x,x_0,x_1,...,x_n],\tag{6.14}$$

则

$$f(x)=N_n(x)+\tilde{R}_n(x).\tag{6.15}$$

称$N_n(x)$为n次Newton插值多项式，$\tilde{R}_n(x)$为相应的插值余项。

容易看到，$N_n(x)$ 是 $f(x)$ 的 n 次多项式，而Lagrange 插值 $L_n(x)$ ((6.6)式) 也是 n 次多项式，两者之间有什么关系呢？

下面来证明，$N_n(x)$与$L_n(x)$是同一插值多项式。 设$p_n(x)$是$f(x)$的n次Lagrange插值多项式，则由(6.13)式有

$$p_n(x)=p_n(x_0)+(x-x_0)p_n[x_0,x_1]$$

$$+(x-x_0)(x-x_1)p_n[x_0,x_1]+\cdots$$
$$+(x-x_0)(x-x_1)\cdots(x-x_{n-1})p_n[x_0,x_1,\cdots,x_n]$$
$$+(x-x_0)(x-x_1)\cdots(x-x_n)p_n[x,x_0,x_1,\cdots,x_n].$$

$p_n(x)$为n次多项式，则其$n+1$阶差商为0。又由插值条件$p_n(x)$与$f(x)$ 在所有插值节点处值相同，其各阶差商也相同，所以得到

$$\begin{aligned}p_n(x) &= f(x_0)+(x-x_0)f[x_0,x_1]\\ &\quad +(x-x_0)(x-x_1)f[x_0,x_1]+\cdots\\ &\quad +(x-x_0)(x-x_1)\cdots(x-x_{n-1})f[x_0,x_1,\cdots,x_n]\\ &= N_n(x)\end{aligned}$$

上述证明说明了Lagrange插值多项式$L_n(x)$和Newton插值多项式$N_n(x)$是同一多项式，所不同的是Lagrange插值多项式用$l_i(x)$ 作为插值基函数，而Newton插值多项式是用$1,(x-x_0),(x-x_0)(x-x_1),\cdots,(x-x_0)(x-x_1),\cdots(x-x_{n-1})$的线性组合表示的。当然，我们还可以用其他的多项式空间的基函数来进行多项式插值。但相对于Lagrange插值公式，Newton插值公式的优点在于，当插值节点增加时一个(如$\bar{x}$)时，只需要增加一项，且有递推公式

$$N_{k+1}(x)=N_k(x)+f[x_0,x_1,\cdots,x_k,\bar{x}](x-x_0)(x-x_1),\cdots(x-x_k)(x-\bar{x}).$$

对于函数$f(x)$，插值节点$x_0,x_1,\cdots,x_n$的n次插值多项式，分别用Lagrange插值方法和Newto 插值方法，则$f(x)$可分别表示为

$$f(x)=L_n(x)+R_n(x)=N_n(x)+\tilde{R}_n(x).$$

结合公式(6.9)，则证明了差商的第4个性质。

例 6.1 *已知$f(x)=\sqrt{(}x)$在点$x=2,2.1,2.2$的值，求二次Newton插值多项式，若增加一个点，再求三次Newton插值多项式.*

解：作差商表(表 6.2). 则

$$N_2(x)=1.414214+0.34924(x-2)-0.04110(x-2)(x-2.1).$$

$$\begin{aligned}N_3(x) =&1.414214+0.34924(x-2)-0.04110(x-2)(x-2.1)\\ &+0.009167(x-2)(x-2.1)(x-2.2)\\ =&N_2(x)+0.009167(x-2)(x-2.1)(x-2.2).\end{aligned}$$

可见，当增加一个插值节点时，即增加一次插值多项式次数时，Newton顿插值多项式，只要增加一项即可.

表 6.2: 差商表

x_i	$f(x_i)$	一阶差商	二阶差商	三阶差商
2	1.414214	0.34924	-0.04110	0.009167
2.1	1.449138	0.34102	-0.03835	
2.2	1.483240	0.33335		
2.3	1.516575			

6.1.3 差分及等距节点的插值公式

Newton插值方法增强了插值多项式的灵活性，但在计算差商时，需要多次的除法计算。当插值节点为等距时，插值多项式可进一步简化，同时避免了除法运算。为此，先引入下面的差分的概念。

定义 6.2 *已知函数$f(x)$在等距节点$x_0, x_1 = x_0 + h, ..., x_n = x_0 + nh$处的值$f(x_i) = f_i, (i = 0, 1, ..., n)$, 其中$h > 0$为相邻两节点间的距离，称为步长。定义*

$$\Delta f_i = f_{i+1} - f_i,$$

为$f(x)$在x_i处以h为步长的一阶向前差分，简称为一阶差分。并称

$$\Delta^m f_i = \Delta^{m-1} f_{i+1} - \Delta^{m-1} f_i, \quad m = 2, 3, ...$$

为m阶差分。

特别地，规定零阶差分为$\Delta^0 f_i = f_i, k = 0, 1, ..., n.$

差分与差商具有如下的重要关系：

$$f[x_i, x_{i+1}, ..., x_{i+m}] = \frac{\Delta^m f_i}{m! h^m}. \tag{6.16}$$

特别地，当$i = 0$时，有

$$f[x_0, x_1, ..., x_m] = \frac{\Delta^m f_0}{m! h^m}. \tag{6.17}$$

另外，还可用数学归纳法证明高阶差分和函数值之间的关系为

$$\Delta^m f_i = \sum_{j=0}^{m} (-1)^j C_m^j f_{i+m-j}, \tag{6.18}$$

其中$C_m^j = \frac{m!}{j!(m-j)!}$为二项式系数.

差分的计算可仿照差商，通过构造差分表计算(表 6.3)

表 6.3: 差分表

x_i	f_i	Δf_i	$\Delta^2 f_i$	$\Delta^3 f_i$	$\Delta^4 f_i$
x_0	f_0	Δf_0	$\Delta^2 f_0$	$\Delta^3 f_0$	$\Delta^4 f_0$
x_1	f_1	Δf_1	$\Delta^2 f_1$	$\Delta^3 f_1$	
x_2	f_2	Δf_2	$\Delta^2 f_2$		
x_3	f_3	Δf_3			
x_4	f_4				

上面所讨论的向前差分，同样可以定义向后差分以及中心差分，在此就不一一介绍了。考虑在等距节点处的Newton插值，插值多项式(6.13)中的各阶差商分别用差分代替，得到各种形式的等距节点插值公式。

1. Newton前插公式

已知节点为$x_i = x_0 + ih$以及$f_i = f(x_i), i = 0, 1, ..., n$. 如果要用$n$次Newton插值多项式计算$x = x_0 + th, 0 < t \leq \frac{1}{2}$处的$f(x)$的近似值，利用差分和差商的关系，可得

$$N_n(x) = f_0 + \frac{\Delta f_0}{h}(x - x_0) + \frac{\Delta^2 f_0}{2h^2}(x - x_0)(x - x_1) + ... + \frac{\Delta^n f_0}{n!h^n}(x - x_0)(x - x_1)...(x - x_{n-1}) \tag{6.19}$$

将$x = x_0 + th$代入，得到

$$\begin{aligned} & N_n(x) \\ &= N_n(x_0 + th) \\ &= f_0 + t\Delta f_0 + \frac{t(t-1)}{2!}\Delta^2 f_0 + ... + \frac{t(t-1)...(t-n+1)}{n!}\Delta^n f_0 + R_n(x) \end{aligned} \tag{6.20}$$

其中

$$R_n(x) = R_n(x_0 + th) = \frac{t(t-1)...(t-n)}{(n+1)!}h^{n+1}f^{n+1}(\xi), \quad x_0 < \xi < x_0 + nh.$$

2. Newton 后插公式

如果要计算函数值的点x在最后一个节点x_n附近的近似值，则Newton插值多项式调整为

$$N_n(x) = f_n + \frac{\Delta f_n}{h}(x - x_n) + \frac{\Delta^2 f_n}{2h^2}(x - x_n)(x - x_{n-1}) + ... + \frac{\Delta^n f_n}{n!h^n}(x - x_n)(x - x_{n-1})...(x - x_1) \tag{6.21}$$

同样，将$x = x_n + th, \frac{-1}{2} \leq t < 0$带入上式，则得到Newton后插公式

$$N_n(x)$$

$$= N_n(x_n + th)$$
$$= f_n + t\Delta f_n + \frac{t(t+1)}{2!}\Delta^2 f_n + ... + \frac{t(t+1)...(t+n-1)}{n!}\Delta^n f_n + R_n(x) \qquad (6.22)$$

其中

$$R_n(x) = R_n(x_n + th) = \frac{t(t+1)...(t+n)}{(n+1)!}h^{n+1}f^{n+1}(\xi), \quad x_n - th < \xi < x_n.$$

例 6.2 已知$\tan x$的函数表如下，近似计算$\tan 1.325$.

x_i	1.3	1.31	1.32	1.33
$f(x_i)$	3.6021	3.7471	3.9033	4.0723

解：采用Newton向后插值公式，作差分表

x_i	f_i	Δf_i	$\Delta^2 f_i$	$\Delta^3 f_i$
1.3	3.6021	0.1450	0.0112	0.0016
1.31	3.7471	0.1562	0.0128	
1.32	3.9033	0.1690		
1.33	4.0723			

则

$$P_x(x) = 4.0723 + 0.1690t + \frac{0.0128}{2!}(t+1) + \frac{0.0016}{3!}t(t+1)(t+2).$$

将$t = \frac{1.325-1.33}{0.01}$代入上式，得

$$\tan 1.325 \approx N_3(1.325) = 3.9869.$$

6.1.4 Hermite插值

在实际应用中，为了更好的逼近被插值函数，不仅要求插值多项式函数在节点处与被插函数相同，还要求在节点的导数值也相同。这种插值称为Hermite插值。

设$x_0, x_1, ..., x_n$是$n+1$个互补相同的插值节点，构造一个$2n+1$次的插值多项式$H_{2n+1}(x)$,使其满

$$\begin{cases} H_{2n+1}(x_i) = f(x_i) \\ H'_{2n+1}(x_i) = f'(x_i) \quad i = 0, 1, \cdots, n \end{cases} \qquad (6.23)$$

则称$H_{2n+1}(x)$为Hermient插值多项式。

与前面构造Lagrange插值多项式类似，首先来构造如下的$2n+1$次的插值基函数$h_i(x), \bar{h}_i(x), i = 0, 1, ..., n$,满足条件

$$h_i(x_j) = \delta_{ij}, \qquad h'_i(x_j) = 0 \qquad j = 0, 1, ..., n \qquad (6.24)$$

$$\bar{h}_i(x_j)=0, \qquad \bar{h}_i'(x_j)=\delta_{ij}, \qquad j=0,1,...,n \tag{6.25}$$

由条件(6.24)及(6.25)，以及Lagrange插值基函数$l_i(x)$的性质可知$l_i^2(x)$是$2n$次多项式，且有

$$l_i^2(x_j)=\delta_{ij}, \qquad \left(l_i^2(x_j)\right)'=0, \quad j\neq i$$

从而可将$h_i(x)$表示为

$$h_i(x)=(1+c_i(x-x_i))l_i^2(x), \tag{6.26}$$

其中c_i为待定常数。对上式两端求导，得到

$$h_i'(x)=c_i l_i^2(x)+2l_i(x)l_i'(x)(1+c_i(x-x_i)). \tag{6.27}$$

将(6.24)式中关于h_i的约束代入，可得

$$c_i=-2l_i'(x_i).$$

从而得到

$$h_i(x)=(1-2l_i'(x-x_i))l_i^2(x). \tag{6.28}$$

同理，$\bar{h}_i(x)$可表示为

$$\bar{h}_i(x)=b_i(x-x_i)l_i^2(x).$$

将(6.25)式中关于$\bar{h}_i$的约束代入，计算得到$b_i=1$。从而得到

$$\bar{h}_i(x)=(x-x_i)l_i^2(x). \tag{6.29}$$

于是，类似Lagrange插值公式，得到如下的满足插值条件(6.23)的Hermite插值公式

$$H_{2n+1}(x)=\sum_{i=0}^{n}h_i(x)f(x_i)+\sum_{i=0}^{n}\bar{h}_i(x)f'(x_i). \tag{6.30}$$

由代数学基本定理，易知上述插值多项式是唯一的。类似定理(6.1)，同样可给出Hermite插值余项。

定理 6.2 *设$f(x),f'(x),...,f^{(2n+1)}$在区间$[a,b]$上连续，$f^{(2n+1)}$区间$(a,b)$内存在，节点$a\le x_0<x-1<...<x_n\le b$, $L_n(x)$是满足插值条件的插值多项式。则对任意的$x\in[a,b]$, 插值余项为*

$$R_{2n+1}(x)=f(x)-H_{2n+1}(x)=\frac{f^{(2n+1)}(\xi)}{(2n+1)!}\omega^2(x) \tag{6.31}$$

其中$\omega(x)=\prod_{j=0}^{n}(x-x_j)$.

Hermite插值问题的形式多样，一个具体问题解法往往不唯一。如果能充分利用问题的特点，则求解过程会得到简化。

例 6.3 给定数据表

x_i	1	2	3
$f(x_i)$	2	4	12
$f'(x_i)$		0.1562	0.0128

已知$f(x)$在区间$[1,3]$上具有连续的四阶导数，求满足条件

$$H(x_i)=f(x_i),(i=0,1,2),\quad H'(x_i)=f'(x_1)=3.$$

的插值多项式$H(x)$.

解：方法1： 由插值条件可确定一个次数不超过三次的插值多项式$H(x)$，设

$$H(x)=\sum_{i=0}^{2}h_i(x)f(x_i)+\sum_{i=0}^{2}\bar{h}_1(x)f'(x_1).$$

其中$h_i,\bar{h}_i$都是次数不超过三次的多项式，满足

$$h_i(x_j)=\delta_{ij},\qquad h_i'(x_1)=0\qquad \bar{h}_1(x_j)=0,\qquad \bar{h}_1'(x_1)=1,\qquad j=0,1,2.$$

由插值条件知，x_1,x_2都是$h(x)$的零点，且(x_1)为二重零点，故令

$$h_0(x)=a(x-x_1)^2(x-x_2)=a(x-2)^2(x-3).$$

其中a为待定常数，将$h_0(x_0)=1$代入上式，求得$a=\frac{-1}{2}$, 于是

$$h_0(x)=\frac{-1}{2}(x-2)^2(x-3).$$

对$h_1(x)$,知x_0,x_2是其零点，设

$$h_1(x)=(kx+b)(x-x_0)(x-x_2)=(kx+b)(x-1)^2(x-3).$$

由$h_1(x_1)=1$和 $h_1'(x_1)=0$，得到$k=0,b=-1$, 于是

$$h_1(x)=-(x-1)(x-3).$$

类似地，可求得

$$h_2(x)=\frac{1}{2}(x-x_1)(x-2)^2,\qquad \bar{h}_1(x)=-(x-1)(x-2)(x-3).$$

故$H(x)=2x^3-9x^2+15x-6.$

方法2 构造重节点插商表

x_i	$f(x_i)$	一阶差商	二阶差商	三阶差商
1	2	2	1	2
2	4	3	5	
2	4	8		
2	12			

由Newton插值公式得

$$H(x) = 2 + 2(x-1) + (x-1)(x-2) + 2(x-1)(x-2)(x-3) = 2x^3 - 9x^2 + 15x - 6.$$

6.2 最小二乘法

上节所讲的插值方法的目的是用简单多项式函数近似给定函数，要求插值函数与被插值函数在节点处有相同的函数值或导数值，如果要提高逼近程度，需要增加节点个数，而当数据量比较大时，大量的数据难以保证每个数据值都有好的精确性，当其中有些数据存在一定的误差时，由于插值条件的要求，其误差将完全被插值函数所继承。另外，有时不需要要求某点的误差为0，而是对整体的误差作限制。因此，需要讨论近似函数的另一种确定方法一逼近，也称为数据拟合。

对于观测数据$(x_i, y_i), i = 1, 2, ...N$,设

$$P(x) = \sum_{j=0}^{n} a_j \phi_j(x), \quad n \ll N \tag{6.32}$$

其中，$\phi_0(x), \phi_1(x), ..., \phi_n(x)$是基函数，根据实际观测的数据，可取为幂函数、三角函数、指数函数等，并称$P(x)$为拟合函数。要使$P(x)$很好的逼近实验数据，则应使得每一点x_i 处的偏差的平方和尽可能小,即有

$$\min_{a_0,a_1,...,a_n} Q(a_0, a_1, ..., a_n) = \min_{a_0,a_1,...,a_n} \sum_{i=0}^{N} \left(y_i - \sum_{j=0}^{n} a_j \phi_j(x_i) \right)^2. \tag{6.33}$$

称此多元极值问题(6.33)为最小二乘问题，也称最小二乘拟合，如果考虑到不同点的观测数据所具有的重要性不同，则得到加权最小二乘。

$$\min_{a_0,a_1,...,a_n} Q(a_0, a_1, ..., a_n) = \min_{a_0,a_1,...,a_n} \sum_{i=0}^{N} \omega_i \left(y_i - \sum_{j=0}^{n} a_j \phi_j(x_i) \right)^2. \tag{6.34}$$

特别地，当取$\phi_i(x) = x^i, i = 0, 1, 2, ..., n$ 时，称满足式(6.33)或(6.34)的解$P_n^*(x) = \sum_{i=0}^{n} a_i^* x_i^k$为最小二乘拟合多项式。

对(6.33)求最小值，令$\frac{\partial Q}{\partial a_k}=0, k=0,1,...,n$,得

$$\sum_{i=1}^{N}(y_i-\sum_{j=0}^{n}a_j\phi_j(x_i))\phi_k(x_i)=0, \qquad k=0,1,...,n \tag{6.35}$$

此方程称为法方程（正规方程）。若记

$$\Psi_j=(\phi_j(x_1),\phi_j(x_2),...,\phi_j(x_N))^T, \quad j=0,1,...,n$$
$$\mathbf{Y}=(y_1,y_2,...,y_N)^T$$

则法方程可写为

$$\sum_{j=0}^{n}a_i(\Psi_i,\Psi_k)=(\mathbf{Y},\Psi_k) \qquad k=0,1,...,n \tag{6.36}$$

其中(.,.)表示$\mathbb{R}^N$中的内积，此方程组的系数矩阵为

$$\begin{pmatrix} (\Psi_0,\Psi_0) & (\Psi_0,\Psi_1) & \cdots & (\Psi_0,\Psi_n) \\ (\Psi_1,\Psi_0) & (\Psi_1,\Psi_1) & \cdots & (\Psi_1,\Psi_n) \\ \vdots & \vdots & & \vdots \\ (\Psi_n,\Psi_0) & (\Psi_n,\Psi_1) & \cdots & (\Psi_n,\Psi_n) \end{pmatrix}$$

此矩阵为可逆的，从而说明正规方程组的解存在并唯一。

若记$A=(\Psi_0,\Psi_1,...\Psi_n)$, 则$A$为一个$N\times n$矩阵，以上最小二乘问题的法方程为

$$A^TAX=A^TY$$

其中$\mathbf{X}=(a_0,a_1,...,a_n)^T$。当$N$不是很大时，可以用平方根法进行求解，当$N$比较大，矩阵$A^TA$常常是病态的，此时可用正交分解的方法求解。

例 6.4 *对于数据表*

x_i	1	2	3	4	5
$f(x_i)$	4	4.5	6	8	8.5
ω_i	2	1	3	1	1

已知其经验公式为$y=a+bx$, 用最小二乘方法确定参数a,b.

解：根据最小二乘拟合思想，满足正规方程组

$$\begin{cases} a\sum\limits_{i=1}^{5}\omega_i+b\sum\limits_{i=1}^{5}\omega_ix_i=\sum\limits_{i=1}^{5}\omega_if(x_i), \\ a\sum\limits_{i=1}^{5}\omega_ix_i+b\sum\limits_{i=1}^{5}\omega_ix_i^2=\sum\limits_{i=1}^{5}\omega_ix_if(x_i). \end{cases}$$

代入数据表并计算，得到

$$\begin{cases} 8a + 22b = 47, \\ 22a + 74b = 145.5. \end{cases}$$

从而解得

$$a = 2.5648, b = 1.2037.$$

即满足观测数据表的线性拟合多项式为

$$y = 2.5648 + 1.2037x.$$

6.3 人工神经网络BP算法

前面所介绍的多项式插值是用已知的简单函数来逼近未知的、复杂的函数，人工神经网络(Artificial Neural Network, ANN)则提供了用多个简单线性函数来逼近某种函数或算法的思路，通过多个线性函数的叠加，达到了逼近非线性函数的目的。人工神经网络由大量的简单处理单元构成的广泛、并行、互联的网络，通过模拟生物神经系统，得到基于神经元(Neurons)的计算模型。神经元是人工神经网络的基本处理单元，是一个多输入单输出的非线性器件，其结构如图(6.1)所示。

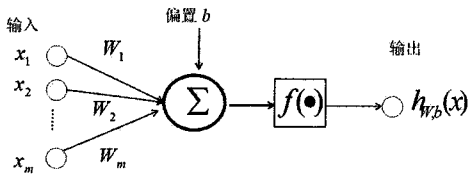


图 6.1: 神经元的数学模型

其数学模型具体表示为

$$h_{W,b} = f(W^T x) = f(\sum_{i=1}^{m} W_i x_i + b) \tag{6.37}$$

其中x_i是输入，W_i为权重，b是偏置量。f为激活函数(Active function)，常用的有Sigmoid函数、tanh函数、径向基函数、小波函数、修正线性单元(ReLU)函数等，对应的公式为

$$\begin{aligned} sigmoid(x) &= \frac{1}{1+e^{-x}} \\ \tanh(x) &= \frac{e^x - e^{-x}}{e^x + e^{-x}} \\ ReLU(x) &= \max(0, x) \end{aligned} \tag{6.38}$$

由许多神经元组成了神经网络结构，神经元也被称为节点或处理单元，每个节点均具有相同的结构。根据神经元的连接方式，以及网络连接的拓扑结构的不同，神经网络模型可分为前向网络（有向无环）和反馈网络（无向完备图，也称循环网络）。前向网络，是简单非线性函数的多次复合，网络结构简单，易于实现。分层前馈网络的神经元分层排列，总体上由三部分组成(图6.2)：输入层、隐藏层（为方便起见，图中只给出一层，实际中可以有多层）和输出层。输入层接收外界的输入信息，输入模式经过各层神经元的响应处理变为输出层的输入，输出层将信息在神经元中分析后传出，而隐层连接和输入和输出层，隐层可以有多层，每层的神经元只接受前一层神经元的输入，并输出给下一层，没有反馈。但在同一层内，各神经元之间无连接。节点分为两类，即输入单元和计算单元，每一计算单元可以任意多个输入，但只有一个输出（它可以耦合到任意多个其他节点作为其输入）。

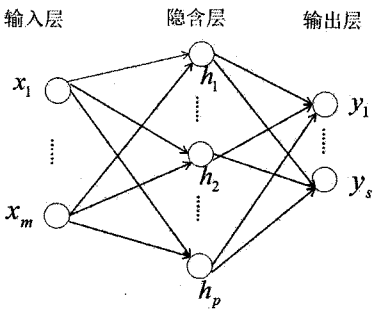


图 6.2: 单隐层前馈神经网络

下面首先以单隐层前向网络(图6.2)为例，介绍其具体算法。这是一个单隐层前馈神经网络，输入为$x_i, x_2, ..., x_m$，输出为$y_1, y_2, ..., y_s$，隐层结点用$h_1, h_1, ..., h_p$ 表示，共有p个。则隐层神经元的输出为

$$\begin{cases} \mathbf{h}^{(1)} = f^{(1)}(\sum_{i=1}^{s} \mathbf{x}_i W_i^{(1)} + \mathbf{b}^{(1)}) \\ \mathbf{y} = f^{(2)}(\sum_{j=1}^{p} \mathbf{h}_j^{(1)} W_i^{(2)} + \mathbf{b}^{(2)}) \end{cases} \tag{6.39}$$

其中$W_i^{(1)} \in \boldsymbol{R}^{m\times s}$，$\mathbf{b}^{(1)}$ 分别为输入到隐层的权值连接矩阵和偏置，$W_i^{(2)} \in \boldsymbol{R}^{p\times s}$，$\mathbf{b}^{(2)}$ 分别为隐层到输出层的权值连接矩阵和偏置，$f^{(1)}$和 $f^{(2)}$为相应的激活函数。

假设训练集个数为N,数据集为$\{\mathbf{x}^{(n)}, \mathbf{y}^{(n)}\}$，输入$\mathbf{x} \in \boldsymbol{R}^m$，输出$\mathbf{y} \in \boldsymbol{R}^s$。则输入

输出之间满足

$$\mathbf{y}=T(x,\theta)=f^{(2)}\left(\sum_{j=1}^{n}f^{(1)}(\sum_{i=1}^{m}\mathbf{x}_iW_i^{(1)}+\mathbf{b}^{(1)})W_j^{(2)}+b^{(2)}\right).$$

其中$\theta=(W^{(1)},\mathbf{b}^{(1)};W^{(2)},\mathbf{b}^{(2)})$。网络的训练过程是通过寻找$\theta$以更好地逼近某种函数，从而极小化误差，即优化如下的目标函数

$$min_\theta L(\theta)=\frac{1}{N}\sum_{n=1}^{N}\|\mathbf{y}^{(n)}-T(\mathbf{x}^{(n)};\theta)\|_F^2+\lambda\sum_{l=1}^{2}\|W^{(l)}\|_F^2.$$

利用梯度下降法求解目标函数，从而得到

$$\begin{cases}\theta^{(k)}=\theta^{k-1}-\alpha\nabla\theta|_{\theta=\theta^{k-1}}\\ \nabla\theta|_{\theta=\theta^{k-1}}=\dfrac{\partial L(\theta)}{\partial\theta}|_{\theta=\theta^{k-1}}\end{cases}$$

其中k为迭代次数。

当隐层个数超过2层时，称为多隐层前馈神经网络，或深度前馈神经网络。BP（Back Propagation）算法解决多隐层网络的目标函数优化问题，也称为误差反向传播算法。由此算法构成的网络，也称为BP 网络。BP网络是前向反馈网络的一种，也是当前应用最为广泛的一种网络。BP网络主要思想是利用输出后的误差来估计输出层的直接前导层的误差，再用这个误差估计更前一层的误差，如此一层一层的反传下去，就获得了所有其他各层的误差估计。学习过程通过神经网络在外界输入样本的刺激下不断改变网络的连接权值，以使网络的输出不断地接近期望的输出。具体而言，学习过程可分为信号的正向传播与误差的反向传播两个过程组成。

正向传播时，输入样本从输入层传入，经各隐层逐层处理后，传向输出层。若输出层的实际输出与期望的输出不符,则转入误差的反向传播阶段。 反向传播时，将输出以某种形式通过隐层向输入层逐层反传，并将误差分摊给各层的所有单元，从而获得各层单元的误差信号，此误差信号即作为修正各单元权值的依据。

以下以多隐层前馈神经网络为例，介绍BP算法。 设网络的训练集个数为N,数据集为$\{\mathbf{x}^{(n)},\mathbf{y}^{(n)}\}$，其中输入$\mathbf{x}\in\boldsymbol{R}^m$，输出$\mathbf{y}\in\boldsymbol{R}^s$， 隐层个数为$L-1$层，每层上的维数（神经元的个数）为$n_l$， 激活函数记为$f^{(l)}$, 则隐层的输出$\mathbf{h}^{(l)}$

$$\begin{cases}\mathbf{h}^{(l)}=f^{(l)}(\sum_{i=1}^{n_{l-1}}\mathbf{h}_i^{(l-1)}W_i^{(l)}+\mathbf{b}^{(l)}),l=1,2,...,L.\\ \mathbf{h}^{(0)}=\mathbf{x},\quad\mathbf{h}^{(L)}=\mathbf{y}.\end{cases}\tag{6.40}$$

网络的优化目标函数为

$$J(\mathbf{W},\mathbf{b})=\frac{1}{N}\sum_{n=1}^{N}\|\mathbf{y}^{(n)}-\mathbf{h}^{(L)}\|_F^2+\lambda\sum_{l=1}^{L}\|\mathbf{W}^{(l)}\|_F^2.\tag{6.41}$$

其中第一项为网络的误差项，第二项为正则项，防止网络过拟合。目标函数(6.41)式通常采用梯度下降的方法进行求解，通过如下的方法来更新参数。

$$\begin{cases} \mathbf{W}^{(l)} = \mathbf{W}^{(l)} - \alpha \dfrac{\partial J(\mathbf{W},\mathbf{b})}{\partial \mathbf{W}^{(l)}} \\ \mathbf{b}^{(l)} = \mathbf{b}^{(l)} - \alpha \dfrac{\partial J(\mathbf{W},\mathbf{b})}{\partial \mathbf{b}^{(l)}} \end{cases} \tag{6.42}$$

其中α为学习效率。根据梯度下降法的求解，以及求导的链式法则，得到损失项隐层输出关于对应参数的导数为

$$\begin{cases} \dfrac{\partial J(\mathbf{W},\mathbf{b})}{\partial \mathbf{W}^{(l)}} = \dfrac{\partial J}{\partial(\mathbf{h}^{(L)})} \dfrac{\partial \mathbf{h}^{(L)}}{\partial \mathbf{h}^{(L-1)}} \cdots \dfrac{\partial \mathbf{h}^{(l+1)}}{\partial \mathbf{h}^{(l)}} \dfrac{\partial \mathbf{h}^{(l)}}{\partial \mathbf{W}^{(l)}} \\ \dfrac{\partial J(\mathbf{W},\mathbf{b})}{\partial \mathbf{b}^{(l)}} = \dfrac{\partial J}{\partial(\mathbf{h}^{(L)})} \dfrac{\partial \mathbf{h}^{(L)}}{\partial \mathbf{h}^{(L-1)}} \cdots \dfrac{\partial \mathbf{h}^{(l+1)}}{\partial \mathbf{h}^{(l)}} \dfrac{\partial \mathbf{h}^{(l)}}{\partial \mathbf{b}^{(l)}} \end{cases} \tag{6.43}$$

其中$\delta^{(l)} = \frac{\partial J}{\partial(\mathbf{h}^{(l)})}$称为误差传播项，又

$$\begin{cases} \dfrac{\partial \mathbf{h}^{(l)}}{\partial \mathbf{W}^{(l)}} = \dfrac{\partial f^{(l)}(\mathbf{h}^{(l-1)})^T \mathbf{W}^{(l)} + \mathbf{b}^{(l)}}{\partial \mathbf{W}^{(l)}} = \mathbf{h}^{(l)} \circ (f^{(l)})' \\ \dfrac{\partial \mathbf{h}^{(l)}}{\partial \mathbf{b}^{(l)}} = \dfrac{\partial f^{(l)}(\mathbf{h}^{(l-1)})^T \mathbf{W}^{(l)} + \mathbf{b}^{(l)}}{\partial \mathbf{b}^{(l)}} = \mathbf{1} \circ (f^{(l)})' \end{cases} \tag{6.44}$$

其中$\circ$为Hadamard积。

正则项关于参数的导数为

$$\begin{cases} \dfrac{\partial \sum\limits_{l=1}^{L} \|W^{(l)}\|_F^2}{\mathbf{W}^{(l)}} = 2\mathbf{W}^{(l)} \\ \dfrac{\partial \sum\limits_{l=1}^{L} \|W^{(l)}\|_F^2}{\mathbf{b}^{(l)}} = 0 \end{cases} \tag{6.45}$$

从上述的算法中可看到，在计算优化目标函数$L(\mathbf{W},\mathbf{b})$中关于第l个隐层参数的梯度下降量时，通过第l个隐藏参数的梯度决定，并引入误差传播项来实现误差的反向传播。

BP算法的工作流程如下，对每个训练样本，BP算法执行以下操作：输入训练样本，然后逐层将进行传输，直到产生输出层的结果；然后计算输出层的误差，再将误差逆向传播至隐层神经元，最后根据隐层神经元的误差来对连接权和阈值进行调整，该迭代过程循环进行，直到达到某停止条件为止。

BP网络实现了一个从输入到输出的映射功能，而数学理论已证明它具有实现任何复杂非线性映射的功能。这使得它特别适合于求解内部机制复杂的问题。无需建立模型，或了解其内部过程，只需输入，获得输出。只要能提供足够多的样本模式对BP网络进行学习训练，它便能完成由n维输入空间到m维输出空间的非线性映

算法 6.1 (BP算法)

1: 输入训练集， 初始化网络参数及权重；
2: 根据当前参数，计算当前样本的输出；
3: 计算各层误差；
4: 调整权值，参数更新；
5: 检查网络总误差是否达到精度要求 满足，则训练结束；不满足，则返回步骤 2。

射。而且理论上，一个三层的神经网络，能够以任意精度逼近给定的函数，这是非常诱人的期望。另外，BP网络具有比较好的泛化能力和较强的容错能力。 但同时BP网络也存在问题，如 BP算法为一种局部搜索的优化方法，但它要解决的问题为求解复杂非线性函数的全局极值，因此，算法很有可能陷入局部极值，使训练失败。当训练次数增多时，学习效率可能下降，收敛速度慢（需做大量运算）。后续涌现出了不少改进的BP 算法，具体可参考相关文献。

6.4 小波变换简介

从上节可知，多项式插值因为函数的简单而得到广泛应用，当然，可以作为插值的函数不仅可以取多项式函数，有理函数，还可以取其他函数，如小波函数。小波变换也是一种函数逼近的方法，基本思想是用一簇多尺度、多分辨率的函数去表示或逼近一个信号或函数。这一簇函数是由母小波函数在不同尺度的平移和伸缩构成，平移参数和尺度参数的引入，同时提高了时间分辨率和频域分辨率，克服了傅里叶分析的时频局部性缺陷。小波变换先在比较大的分辨率下分解图像，随着分辨率的增强，尺度的减小，小波基函数的伸缩，图像的细节信息逐渐显现，这一多分辨率(Multi-Resolution， MR) 或多尺度(Multi-scale) 的特性符合人类视觉观察图像的特点，小波分析也被称为是“数学显微镜”。小波分析的多尺度分析理论更是广泛应用于信号分析、图像处理、理论物理、大型机械故障诊断等领域。

6.4.1 傅里叶变换与加窗傅里叶变换

傅里叶（Fourier）在求解热传导方程

$$\begin{cases} u_t(x,t) = u_{xx}(x,t) & t>0, 0 \le x \le \pi \\ u(x,0) = f(x) & 0 \le x \le \pi \\ u(0,t) = A \quad u(\pi,t) = B \end{cases}$$

时，得到此方程的解可表示为如下形式

$$f(x)=a_0+\sum_{k=1}^{\infty}a_k\cos kx+b_k\sin kx,\qquad x\in[-\pi,\pi] \tag{6.46}$$

称此级数为傅里叶级数。这意味着任何一个以2π 为周期的有限长信号,可以表示成三角函数的线性组合。注意到三角函数系$\{1,\cos x,\sin x,\cos 2x,\sin 2x,...\}$ 的正交性，组合系数可以通过两端做内积而得到。

$$\begin{aligned}a_1&=\langle f,1\rangle=\frac{1}{2\pi}\int_{-\pi}^{\pi}f(x)dx\\ a_k&=\langle f,\cos x\rangle=\frac{1}{\pi}\int_{-\pi}^{\pi}f(x)\cos kxdx\\ b_k&=\langle f,\sin x=\rangle\frac{1}{\pi}\int_{-\pi}^{\pi}f(x)\sin kxdx\end{aligned} \tag{6.47}$$

组合系数反应了信号$f(x)$中所含频率成分的多少。

设$f(t)$定义在整个实数轴，连续可微，且$\int_{-\infty}^{\infty}\|f(t)\|dt<\infty$, 取频率变量连续的从负无穷到正穷变化，结合欧拉公式和傅里叶级数，则得到傅里叶变换公式以及其逆变换

定义 6.3 *如果下面的积分变换存在，则称$\hat{f}(\omega)$ 为 $f(t)$的连续傅里叶变换，简称傅里叶变换*

$$\hat{f}(\omega)=\int_{-\infty}^{\infty}f(t)e^{-i\omega t}dt=\left\langle f,e^{i\omega t}\right\rangle. \tag{6.48}$$

而

$$f(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\hat{f}(\omega)e^{i\omega t}d\omega=\left\langle\hat{f},e^{-i\omega t}\right\rangle. \tag{6.49}$$

称为连续傅里叶逆变换，简称为傅里叶逆变换。

傅里叶变换系数$\hat{f}(\omega)$也称为频谱，表示了信号$f(t)$ 中频率为ω的分量。 傅里叶变换提供了从时间域到频率域的一种转换，将一个无限时宽的信号分解为频率为ω的一系列频率分量。这种表示信号的方法颠覆了人们之前从时间域观察信号的思维，提供了全新的频率域视角。

设f,g都是定义在实数轴上的可微函数，其傅里叶变换记为$\hat{f},\hat{g}$，则下列性质(表6.4)成立。 这些性质的证明都可以通过傅里叶变换的定义进行证明。其中卷积性质是信号处理中滤波器理论的基础，具有非常重要的作用。两个函数f,g的卷积定义如下

定义 6.4

$$(f*g)(t)=\int_{-\infty}^{\infty}f(t-x)g(x)dx"$$

表 6.4: 傅里叶变换性质

性质	时间域	频率域
线性性	$af+bg$	$a\hat{f}+b\hat{g}$
时域平移性	$f(t-t_0)$	$e^{-j\omega t_0}\hat{f}(\omega)$
频域平移性	$e^{j\omega t_0}f(t)$	$\hat{f}(\omega-\omega_0)$
尺度性	$f(at)$	$\frac{1}{a}\omega\hat{f}\left(\frac{\omega}{a}\right)$
时域微分	$f^{(n)}$	$(i\omega)^n\hat{f}$
频域微分	$(-it)^n f$	$\hat{f}^{(n)}$
卷积	$f*g$	$\hat{f}\hat{g}$
parsevela等式	$\int_{-\infty}^{\infty}\|f(t)\|^2dt=\frac{1}{2\pi}\int_{-\infty}^{\infty}\left\|\hat{f}(\omega)\right\|^2d\omega$	

虽然傅里叶变换可以有效的表达信号的频率特征，但同时也丢失掉了信号的时间信息，原因在于其基函数为三角函数，而三角函数是无始无终的周期性波，只具有频率信息，适合于分析平稳信号，而对于具有显著局部特征的非平稳信号，傅里叶变换变换就无能为力了。

为了克服傅里叶变换基函数没有时间局部性这一缺点，加窗傅里叶变换首先选取时间域上具有一定局部性的窗函数(图6.3)，再通过窗函数和复指数函数的乘积，构造了同时具有时间和频率局部信息的函数

$$g_{\omega,b}(t)=g(t-b)e^{i\omega t},$$

其中ω 和b分别表示频率参数和时间域上的平移参数，$g(t)$称为窗函数，满足 $tg(t)\in L^2(R)$, $\|g\|=1$。 如高斯函数，三角窗函数，汉明窗函数等都满可以作为窗函数。确定窗函数后，$g_{\omega,b}$的基本形状确定，通过频率参数ω 和平移参数b来调节具体形状。图(6.4)给出了$g_{\omega,b}$中，窗函数取高斯函数，b 固定，ω 变换对应的函数。需要注意的是，随着参数ω 的变化，$g_{\omega,b}$的频率发生改变，但其在时间轴上的窗口大小没有发生改变。

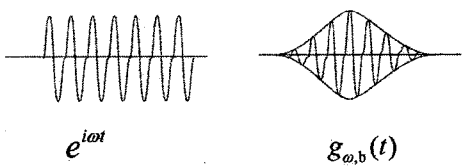
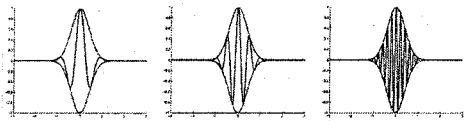
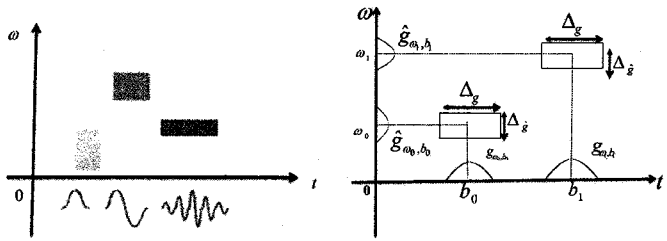


图 6.3: (a) 傅里叶变换基函数. (b) 加窗傅里叶变换函数.



$$g_{\omega,b}(t)=e^{-\pi(t-b)^2}+e^{2\pi i\omega t} \quad \text{with} \quad b=2,\omega=1,2,5$$

图 6.4: 不同频率的$g_{\omega,b}$.



(a) 时频窗 (b) 加窗傅里叶变换原子的时频窗

图 6.5: 不同的时频窗

由于$g_{\omega,b}$含有频率和时间两个参数，所以也称其为时频原子，特别地，当窗函数取高斯函数时，称为Gabor原子。对于一个时频原子，确定了其时间域的中心、方差（半径Δg）以及频率域的中心、方差$(\Delta\hat{g})$之后，可以得到时频平面的时频窗，也称为Heisenberg窗(6.5(a))。 时频窗的半径反应了这个原子在时间域和频率域的局部性，半径越小，对应的局部性越好，或者说具有比较好的时间或频率的分辨率。在分析信号时，人们希望能同时得到最好的时频局部性，但Heisenberg测不准原理说明，这是不可能实现的， 原子g 的时频窗的面积满足

$$\Delta g\Delta\hat{g}\geq\frac{1}{2}$$

其中$\hat{g}$是原子g的傅里叶变换。

图(6.5(b))给出了加窗傅里叶变换的时频原子g_{ω_0,b_0}和g_{ω_1,b_1} 的时频窗，可以看到，不管是对于低频还是高频信号，时频原子$g_{\omega,b}$ 的时频窗大小始终不变，只在时频窗内进行平移。也就是说，其在时间和频率上的分辨率是固定的，这点在分析非平稳信号，尤其是频率从低频到高频，或者从高频到低频快速变化的信号时，有一定的局限性。

给定了窗函数，则仿照傅里叶变换，加窗傅里叶变换以及逆变换定义为

$$\begin{aligned}(Gf)(\omega,b) &=< f, g_{\omega,b} >= \int_{-\infty}^{\infty} f(t)\overline{g_{\omega,b}}dt = \int_{-\infty}^{\infty} f(t)\overline{g(t-b)}e^{-i\omega t}dt \\ f(t) &= \frac{1}{2\pi}< Gf, \bar{g}_{\omega,b} >= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Gf(\omega,b)g_{\omega,b}d\omega db\end{aligned} \tag{6.50}$$

相比傅里叶变换，加窗傅里叶变换同时具有了时间和频率的局部性，变换后的系数反映了原信号在一定的时间范围内，局部的频率信息。但从图(6.5)的分析可知，由于其基函数在时频窗内具有固定的分辨率，因此不管是对高频信号，还是低频信号，都提取了同样大小的时间段和频率范围内的信息。

6.4.2 连续小波变换

对一个非平稳信号进行分析时，其短期内变化快的高频部分的时间局部性是非常重要的，反之，对于其变化较平稳的低频部分，此时更应该关注其频率信息。为了达到此目的，产生了一种具有时频局部性调节功能的函数，即小波函数。

定义 6.5 称满足下列允许性条件的$\psi(t)$为母小波(Mother Wavelet)

$$C_\psi = \int_{-\infty}^{+\infty} \frac{\left|\hat{\psi}(\omega)\right|^2}{\omega} d\omega < \infty. \tag{6.51}$$

从这个定义里，可以得到以下几点

1. $\lim\limits_{\omega\to 0}\hat{\psi}(\omega) = 0$，由此说明$\hat{\psi}(0) = 0$.
2. 由$\hat{\psi}(0) = \int_{-\infty}^{+\infty}\psi(t)e^{-i\cdot 0\cdot x}dt = \int_{-\infty}^{+\infty}\psi(t)dt = 0$

以上两点说明$\psi(t)$上下震荡，满足波的特性，并在时间域上快速衰减。而$\hat{\psi}(0) = 0$，这符合带通滤波器的性质。因此，任何均值为零且在频率增加时以足够快的速度衰减为零的带通滤波器，都可以作为一个母小波。

定义了母小波函数之后，对其进行尺度的伸缩和时间上平移，从而得到一族小波函数$\psi_{a,b}(t)$

$$\psi_{a,b}(t) = |a|^{-\frac{1}{2}}\psi\left(\frac{t-b}{a}\right) \tag{6.52}$$

其中$a > 0$为尺度参数，$a > 1$表示对信号进行压缩，$a < 1$则是对信号进行拉伸。b为时间上的平移参数。

同时频原子$g_{\omega,b}$一样，小波函数$\psi_{a,b}(t)$也是一种时频原子。图(6.6)显示了两个小波原子的时频窗，可以看到，随着尺度因子的变化，时频窗在时间和频率轴的半径发生了变化，当时间域的半径小时，频率域上的半径就大，反之，当时间域上的半

径大时，频率域上的半径则减小，即时间域和频率域的半径成反比。这一点可以由小波函数的傅里叶变换的性质得到。

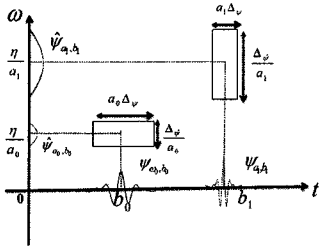


图 6.6: 小波原子的时频窗

小波的允许性条件给出了构造小波的必要条件，在这个条件上再加一些限制条件，则可构成各种小波函数。如Haar小波、Shannon小波、Coifmann小波、样条小波、Daubechies小波等。 图(6.7) 给出了几种小波函数的图形。

定义了小波函数之后，就可以给出连续小波变换的定义

定义 6.6 *设$f(t)$是平方可积的函数，$\psi(t)$是一个小波函数，则称下面的积分变换为连续小波变换*

$$Wf(a,b)=\int_{-\infty}^{\infty}f(t)\frac{1}{\sqrt{a}}\overline{\psi\left(\frac{t-b}{a}\right)}dt=\langle f(t),\psi_{a,b}(t)\rangle. \tag{6.53}$$

其中 $\tilde{\psi}$是ψ的共轭。$Wf(a,b)$称为小波变换的系数，衡量了信号f在时间t的局部领域b内，尺度因子为a的信息。小波变换是以小波函数的共轭为权函数，对原时间域信号在时间轴从负无穷到正无穷进行积分（求和）。而上式的后一个等号则说明，小波变换系数的计算是被变换函数和小波函数的内积，即计算的是两个函数的相似程度。

如果记 $\tilde{\psi}_a(.)=|a|^{\frac{-1}{2}}\bar{\psi}(\frac{-t}{a})$. 则小波变换可以写成卷积形式

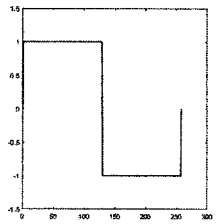
$$Wf(a,b)=(f*\tilde{\psi}_a)(b).$$

结合前面所提到的傅里叶变换的卷积性质，小波变换卷积形式的表示说明了小波变换也是一种滤波，并且可以通过卷积快速第实现，这点将会在快速小波变换的计算方法中被用到。

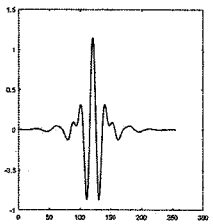
定义逆小波变换为

定义 6.7

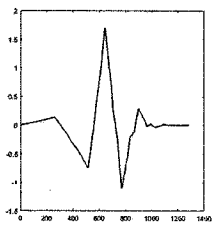
$$f(t)=\frac{1}{C_\psi}\int_{-\infty}^{\infty}\int_{0}^{\infty}Wf(a,b)\psi_{a,b}(t)\frac{dadb}{a^2}=\frac{1}{C_\psi}\langle Wf,\overline{\psi_{a,b}}\rangle_a. \tag{6.54}$$



(a) Haar小波函数



(b) Shannon小波函数



(c) Coifmann小波函数

图 6.7: 小波函数

其中

$$\langle f,g\rangle_a \overset{def}{=} \int_{-\infty}^{\infty}\int_0^{\infty} f(a,b)\overline{g(a,b)}\frac{dadb}{a^2}. \tag{6.55}$$

设f,g都是平方可积函数，其对应的小波变换记为Wf,Wg，连续小波变换具有下列性质

性质1 (线性性) $af+bg \longleftrightarrow aWf+bWg$;

性质2 (平移性) $f(t-t_0) \longleftrightarrow Wf(a,b-t_0)$;

性质3 (尺度性) $f(kt) \longleftrightarrow \frac{1}{\sqrt{(k)}}Wf(ak,kb)$;

性质4 (Parseval等式) $C_\psi\|f\|^2 = \|Wf\|^2$.

6.4.3 多尺度分析

定义 6.8 $L^2(\boldsymbol{R})$上的一系列闭子空间$\{V_j\}, j \in Z$如果满足以下的五个条件，则称为是一个多尺度分析(Multiscale Analysis) 或多分辨率分析(Multiresoulution Analysis)。

1. 单调性: $V_j \in V_{j-1}, \forall j \in Z$

2. 逼近性: $\lim\limits_{j\to+\infty} V_j = \bigcap\limits_{j\in z} V_j = \{0\}$, $\lim\limits_{j\to-\infty} V_j = \bigcup\limits_{j\in z} V_j = L^2(R)$

3. 尺度性: $u(t) \in V_j \Longleftrightarrow u(2t) \in V_{j-1}$

4. 平移不变性: $u(t) \in V_j \Longleftrightarrow u(t-k) \in V_j$

5. Riesz 基: $\exists\theta, \theta(t-k), k \in Z$构成$V_0$的一组Riese基。即设$\theta_k, k \in Z$是Hibert空间$H$的一组基，若存在两个常数$\lambda_{min}, \lambda_{max}$，满足$0 < \lambda_{min} \le \lambda_{max} < \infty$, 对$\forall x \in H, x = \sum_{j\in Z} a_k\theta_k$, 有

$$\lambda_{min}\|x\| \le \|a_k\| \le \lambda_{max}\|x\|$$

则称$\{\theta_k\}, k \in Z$构成了H空间的一组Riesz基。

多分辨率分析从分辨率（尺度）的角度对物理世界进行建模，V_j 空间代表了尺度为2^{-j},分辨率为2^j 的所有信息的集合，所以也称V_j为尺度空间,涵盖了从零子空间到整个$L^2(\boldsymbol{R})$ 空间。这说明，任意的一个函数属于$L^2(\boldsymbol{R})$，一定存在于某一个V_j 空间内。性质1说明，低分辨率的空间V_j 包含在高一级的分辨率空间V_{j-1}中。随着j的增大，分辨率逐步降低，反之，随着当j的减小，分辨率逐渐提高，最后到整个的$L^2(\boldsymbol{R})$空间($j \to -\infty$)。 由此得到了性质2。 而性质3说明，当对函数的细节放大2倍，则函数属于低一级的分辨率空间，反之亦然。从而得到

$$u(t) \in V_j \Longleftrightarrow u(2^j t) \in V_0$$

性质4说明，时间域上的平移不改变频率，即仍然在原来的空间里。综合性质3和性质4，可得到很重要的结论

$$u(t) \in V_0 \Longleftrightarrow u(2^{-j}(t-k)) \in V_j \tag{6.56}$$

式(6.56)表明，V_0 空间的元素和V_j空间的元素存在对应的关系，由此，研究一系列子空间V_j 的问题可转换为研究一个子空间V_0。 性质5保证了V_0 空间基的存在性，由线性代数的知识可知，通过标准正交化过程可将$\theta(t-k)$转化为V_0 空间的一组标准正交基$\phi(t)$。下面的定理给出了V_0空间的标准正交基$\phi(t)$的傅里叶变换构造。

定理 6.3 设$\{V_j\}, j \in Z$是一个多分辨率分析，V_0空间的基函数$\phi(t)$，也称为尺度函数，其傅里叶变换有下列的形式

$$\hat{\phi}(\omega) = \frac{\hat{\theta}(\omega)}{\left(\sum_{k=-\infty}^{+\infty} |\hat{\theta}(\omega + 2k\pi)|^2\right)^{1/2}}$$

例 6.5 Haar尺度函数

$$\phi(t) = \begin{cases} 1 & 0 \le t \le 1 \\ 0 & otherwise \end{cases}$$

例 6.6 几种多分辨率分析

1. 线性样条多分辨率分析：线性样条是连续的、逐段线性的函数，可以构造一个基于线性样条的多分辨率分析。其尺度函数$\phi(t)$ 满足

$$\phi(t) = \begin{cases} t+1 & -1 \le t0 \\ 1-t & 0 < t \le 1 \\ 0 & |t| > 1 \end{cases}$$

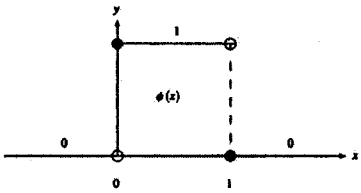


图 6.8: Haar尺度函数.

注意到，这组基函数不是正交的，但可以通过后面所讲的正交化方法构建一个新的尺度函数，以形成空间的标准正交基。

2. Shannon多分辨率：尺度函数空间由能量有限信号组成，在区间$[-2^{-j}, 2^j]$之外，函数的傅里叶变换为0.尺度函数为

$$\phi(t) = sinc(t) = \begin{cases} \dfrac{\sin(\pi t)}{\pi t}, & t \neq 0 \\ 1 \quad t = 0. \end{cases}$$

定理 6.4 设$\{V_j\}, j \in Z$是一个多分辨率分析，$\exists \phi(t) \in V_0, \ni \{\phi(t-k), \mid k \in Z\}$构成$V_0$的一组标准正交基，则 $\phi_{j,k} = 2^{-j/2}\phi(2^{-j}x - k)$ 构成V_j 的一组标准正交基。

定理6.4的证明可由性质3,4和公式(6.56)得到，这些性质和定理再次说明，V_j与V_0空间元素存在着对应关系，从而其基函数、标准正交基函数也存在同样的对应关系，因而，已知V_0 空间的一组标准正交基，则可以通过平移和伸缩变换，构造V_j 空间的一组标准正交基，从而一个多分辨率分析$\{V_j\}, j \in Z$完全由尺度函数$\phi_j(t)$所刻画。而尺度函数具有的性质将保证空间$\{V_j\}$满足多分辨率分析的所有条件。

$L^2(\boldsymbol{R})$空间的函数$f(t)$在空间V_{j-1} 的正交投影（逼近）可表示为

$$\mathrm{Proj}_{V_{j-1}}(f) = \sum_k \langle f, \phi_{j-1,k} \rangle \phi_{j-1,k} \tag{6.57}$$

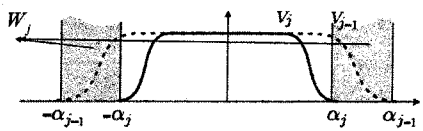


图 6.9: V_j 与V_{j-1}之间的频带.

但由多分辨率分析的定义知，高分辨率空间V_{j-1}包含了低分辨率空间V_j(图6.9)，也就是说，在V_j 之外，还有一部分信息包含在V_{j-1}中，记这部分空间为W_j，

则W_j是V_j在V_{j-1} 中的正交补空间，称为小波空间。即

$$V_{j-1} = V_j \oplus W_j$$

由此可看到小波空间W_j包含了不同分辨率尺度空间的差别，也就是细节信息。即高一级分辨率的空间等于低一级的分辨率空间加上细节，这些细节可通过一个带通滤波器得到，其频率在$[\alpha_j, \alpha_{j-1}]$ 之间。 依次类推，则得到以下定理。

定理 6.5 设$\{V_j\}, j \in Z$是一个依尺度函数$\phi_j(t)$的多分辨率分析，W_j是V_j在V_{j-1} 中的正交补空间，则有

$$\begin{aligned} V_{j-1} &= V_j \oplus W_j \\ &= V_{j+1} \oplus W_{j+1} \oplus W_j \\ &= ..... \\ &= ... \oplus W_2 \oplus W_1 \oplus W_0 \oplus W_{j+2} \oplus W_{j+1} \oplus W_j. \end{aligned}$$

和

$$L^2(\boldsymbol{R}) = \bigoplus_{j=-\infty}^{+\infty} W_j.$$

因而$f(t)$在空间V_{j-1} 的正交投影可进一步分解为其在V_j和W_j空间的正交投影之和

$$\begin{aligned} &\mathrm{Proj}_{V_{j-1}} = \mathrm{Proj}_{V_j} + \mathrm{Proj}_{W_j} \\ &= \mathrm{Proj}_{V_{j+1}} + \mathrm{Proj}_{W_{j+1}} + \mathrm{Proj}_{W_j} \\ &= ..... \\ &= ... + \mathrm{Proj}_{W_2} + \mathrm{Proj}_{W_1} + \mathrm{Proj}_{W_0} + ... + \mathrm{Proj}_{W_{j+2}} + \mathrm{Proj}_{W_{j+1}} + \mathrm{Proj}_{W_j}. \end{aligned}$$

为了表示$f(t)$在W_j空间的正交投影，需要找到小波空间W_j 的正交基，由$\{W_j\}, j \in Z$空间的定义，可知其具有和V_j 空间同样的尺度性以及平移不变性

$$u(t) \in W_0 \Longleftrightarrow u(2^{-j}(t-k)) \in W_j \tag{6.58}$$

因此，得到下面的定理

定理 6.6 设$\exists \psi(t) \in W_0, \ni \{\psi(t-k), \mid k \in Z\}$构成$W_0$的一组标准正交基，则 $\psi_{j,k} = 2^{-j/2}\psi(2^{-j}x - k)$构成$W_j$ 的一组标准正交基。

于是，$f(t)$在空间W_{j-1} 的正交投影（逼近）可表示为

$$\mathrm{Proj}_{W_{j-1}}(f) = \sum_k \langle f, \psi_{j-1,k} \rangle \psi_{j-1,k} \tag{6.59}$$

综上，由(6.57)和(6.59)式可得到：

$$\begin{aligned} f(t) &= \sum_k \langle f, \phi_{j,k}\rangle \phi_{j,k} + \sum_k \langle f, \psi_{j,k}\rangle \psi_{j,k} \\ &= \sum_k \langle f, \phi_{j+1,k}\rangle \phi_{j+1,k} + \sum_k \langle f, \psi_{j+1,k}\rangle \psi_{j+1,k} + \sum_k \langle f, \psi_{j,k}\rangle \psi_{j,k} \\ &= \cdots\cdots \end{aligned} \tag{6.60}$$

这表明多分辨率分析通过不同尺度空间和小波空间，得到了信号的逐次逼近的表示。而下一步的问题是，需要找到不同分辨率空间投影系数的关系，这需要引入下面的双尺度方程 。

设$\phi(t) \in V_0 \subset V_{-1}$, 则$\phi(t) = \sum_k h_k \phi_{-1,k}$, 又$\phi_{j,k} = 2^{-j/2}\phi(2^{-j}x - k)$, 因此得到

$$\phi(t) = \sqrt{2}\sum_k h_k \phi(2t - k) \tag{6.61}$$

相应地，对小波空间的基函数$\psi(t) \in W_0 \subset V_{-1}$, 则$\psi(t) = \sum_k h_k \phi_{-1,k}$, 又$\phi_{j,k} = 2^{-j/2}\phi(2^{-j}x - k)$, 因此得到

$$\psi(t) = \sqrt{2})\sum_k g_k \phi(2t - k) \tag{6.62}$$

(6.61)和(6.62) 式给出了相邻尺度的尺度空间，以及小波空间中基函数的关系，因此称为双尺度方程(Two-scale relations)。双尺度方程建立起了尺度函数$\phi(t)$，小波函数$\psi(t)$与离散滤波器组h_k ，g_k的对应关系，从而构造小波函数可归结为构造尺度函数，或者滤波器的问题。可以证明，下面条件给出了正交小波基存在的必要条件

$$\begin{cases} \hat{h}(0) = 1 \\ \left|\hat{h}(w)\right|^2 + \left|\hat{h}(w+\pi)\right|^2 = 1 \\ |\hat{g}(w)|^2 + |\hat{g}(w+\pi)|^2 = 1 \\ \hat{h}(w)\overline{\hat{g}(\omega)} + \hat{h}(w+\pi)\overline{\hat{g}(w+\pi)} = 0 \end{cases}$$

其中$\hat{h}, \hat{g}$ 是h_k, g_k的傅里叶变换，也称满足此条件对应的 $\phi(t), \psi(t)$为正交小波。特别地，当取 $\hat{g}(\omega) = \overline{\hat{h}(\omega+\pi)}$, 即 $g_k = (-1)^{k-1}\bar{h}_{1-k}$ 时，称滤波器$\hat{h}_k$和$\hat{g}_k$ 为共轭镜像滤波器。

6.4.4 小波分解与重构算法(Mallat 算法)

本节讨论信号的快速正交小波分解与重构，其算法也称为Mallat算法。快速小波变换不断地将每一逼近Proj_{V_j}表示成较粗糙的逼近 $\text{Proj}_{V_{j+1}}$ 与小波系数$\text{Proj}_{V_{j+1}}$的和。而另一方面，Proj_{V_j}可以由$\text{Proj}_{V_{j+1}}$ 和 $\text{Proj}_{V_{j+1}}$ 重构。

设$\phi_{j,k},\psi_{j,k}$ 分别是 $\phi_{j,k},\psi_{j,k}$ 的标准正交基，函数$f(t)$ 在 V_j,W_j 空间的正交投影系数可表示为记为$c_{j,k},d_{j,k}$, 由式(6.60)以及 $f(t)\in V_j=V_{j+1}\oplus W_{j+1}$,有

$$f(t)=\sum_k c_{j,k}\phi_{j,k}=\sum_k c_{j+1,k}\phi_{j+1,k}+\sum_k d_{j+1,k}\psi_{j+1,k} \tag{6.63}$$

若已知$c_{j,k}$, 求$c_{j+1,k},d_{j+1,k}$. 利用 $\phi_{j,k},\psi_{j,k}$ 的正交性，对式 (6.63) 两端分别用 $\phi_{j+1,k}$, $\psi_{j+1,k}$ 做内积，则得到

$$c_{j+1,k}=\langle f,\phi_{j+1,k}\rangle,\qquad d_{j+1,k}=\langle f,\psi_{j+1,k}\rangle \tag{6.64}$$

将(6.64)代入(6.63)的左端,再利用双尺度方程，得到信号的小波变换分解系数

$$\begin{aligned}c_{j+1,k}&=\left\langle\sum_{l\in Z}c_{j,l}\phi_{j,l}(x),\phi_{j+1,k}(x)\right\rangle\\&=\sum_{l\in Z}c_{j,l}<\phi_{j,l}(x),\phi_{j+1,k}(x)>=\sum_{l\in Z}\overline{h_{l-2k}}c_{j,l}\end{aligned}$$

以及

$$\begin{aligned}d_{j+1,k}&=\left\langle\sum_{l\in Z}c_{j,l}\phi_{j,l}(x),\psi_{j+1,k}(x)\right\rangle\\&=\sum_{l\in Z}c_{j,l}<\phi_{j,l}(x),\psi_{j+1,k}(x)>=\sum_{l\in Z}\overline{g_{l-2k}}c_{j,l}\end{aligned}$$

其中$\bar{h}_l=h_{-l},\bar{g}_l=g_{-l}$,而 h,g的定义见双尺度方程。

如果已知$c_{j+1,k},d_{j+1,k}$,求$c_{j,k}$。同样地，利用$\phi_{j,k},\psi_{j,k}$的正交性，对式(6.63)两端用$\phi_{j,k}$做内积，则得到

$$\begin{aligned}c_{j,k}&=\langle f,\phi_{j,k}\rangle=\left\langle\sum_k c_{j+1,k}\phi_{j+1,k}+\sum_k d_{j+1,k}\psi_{j+1,k},\phi_{j,k}(x)\right\rangle\\&=\sum_{l\in Z}c_{j+1,l}<\phi_{j+1,l}(x),\phi_{j,k}(x)>+\sum_{l\in Z}d_{j+1,l}<\psi_{j+1,l}(x),\phi_{j,k}(x)>\\&=\sum_{l\in Z}h_{k-2l}c_{j+1,l}+\sum_{l\in Z}g_{k-2l}d_{j+1,l}\end{aligned}$$

如果记

$$h'_k=\overline{h}_{-k},g'_k=\overline{f}_{-k}$$

则分解公式可写成

$$\begin{cases}c_{j+1,k}=\displaystyle\sum_{l\in Z}h'_{2k-l}c_{j,l}=(h'*c_j)\downarrow 2\\ d_{j+1,k}=\displaystyle\sum_{l\in Z}g'_{2k-l}c_{j,l}=(g'*c_j)\downarrow 2\end{cases} \tag{6.65}$$

其中$*$表示卷积运算。 若记

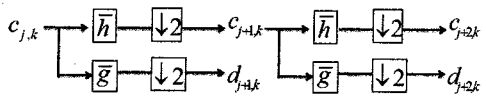
$$c'_{j+1,l}=\begin{cases}c_{j+1,l} & l=2p\\ 0 & l=2p+1,\end{cases}$$

$$d'_{j+1,l}=\begin{cases}d_{j+1,l} & l=2p\\ 0 & l=2p+1\end{cases}$$

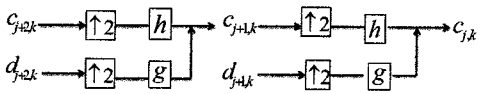
其重构公式可表示成

$$c_{j,k}=(h*c'_{j+1})\uparrow 2+(g*d'_{j+1})\uparrow 2 \tag{6.66}$$

公式(6.65)表明小波变换系数可通过迭代的离散卷积和降采样快速实现，即c_{j+1} 和 d_{j+1} 是c_j分别和滤波器$\bar{h}$ 、$\bar{g}$ 做卷积然后每隔一项做采样而得到的，如图(6.10(a))所示。滤波器$\bar{h}$ 去掉了序列c_j的高频，而$\bar{g}$是高通滤波器，提取了高频信息。在重构公式(6.66)中，首先对c_{j+1} 和 d_{j+1}进行补零插值扩充，然后将这些信号进行滤波，如图(6.10(b))所示。小波分解和重构的过程是迭代进行的，从尺度L出发，对$L\le j<J$迭代计算。



(a) 一维快速小波分解



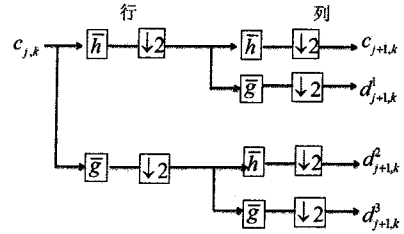
(b) 一维快速小波重构

图 6.10: 一维快速小波分解及重构

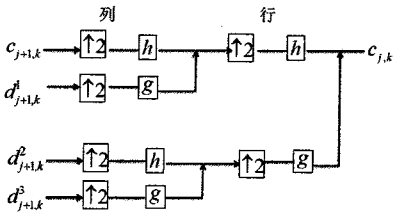
对于二维信号，其小波变换可通过行列分离的两个一维小波变换来实现，如图(6.11) 所示。经过一次小波分解后，二维信号（图像）分解为4个子图像，1个低频，3个高频。图(6.12) 给出了Lena图像的2次小波分解图。

6.4.5 小波变换的应用

小波变换这种多尺度、多分辨率思想，在分析非平稳信号时具有很大的优势。目前，小波变换在语音识别、图像处理等领域得到了广泛的应用，如一维信号的奇异点检测、故障检测、去噪等，二维信号的边缘检测、去噪、压缩、融合等。



(a) 二维快速小波分解



(b) 二维快速小波重构

图 6.11: 二维快速小波分解及重构

利用小波变换进行信号处理主要包括三个基本步骤. (1)对信号进行小波多尺度分解，得到每一尺度上的细节和近似系数；(2) 针对不同的应用目的，对小波系数（特别是细节系数）进行处理，提取信息或改变小波系数值；(3) 对处理后的小波系数进行多尺度小波逆变换，重构信号。对于信号的故障检测、特征提取等，只需进行到第(2)步即可，而如果要重构信号，则需要第(3)步逆变换。 本节简单地介绍小波变换在检测信号奇异点和信号去噪方面的应用。

信号经过小波变换分解后，其系数具有稀疏性，即大部分的系数为0，不为0的对应的信号中有变化的点，从而小波变换系数模极大对应了信号中的奇异点。如图(6.13)所示,信号$f(t)$中存在不同的奇异点，经过多尺度的小波分解后，得到的不同尺度上的小波系数均在奇异点处达到了极值。根据此原理，可用小波变换进行信号的奇异点检测，在二维信号中则对应图像的边缘检测等。另外，通过跟踪小波系数不同尺度上的模极大，可迭代的重构信号。

从频率上，信号可分为低频和高频两部分，其中噪声、二维图像的纹理、边缘等细节对应了高频部分，而信号的概貌、图像的平滑部分对应了低频部分。信号去噪是在去除噪声的同时尽量保留更多的原有信号，包括图像的细节，因此，在利用小波变换进行信号去噪时，通常保留其低频部分，只对高频系数进行改变，通过改变不同尺度下的小波高频系数来抑制高频，保留低频，从而达到去噪目的。 基于小

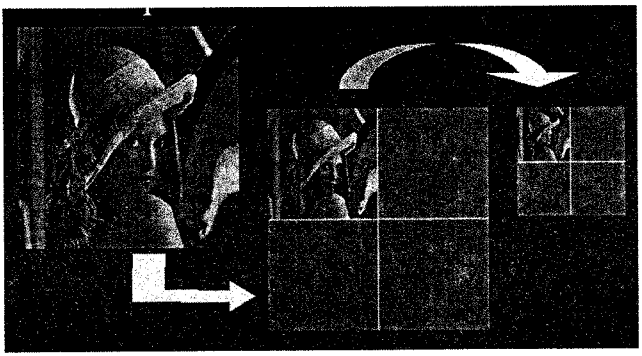


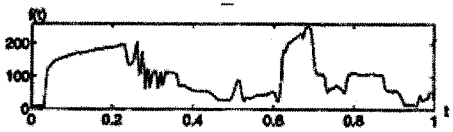
图 6.12: Lena小波分解示意图

波变换的信号去噪方法主要有三类：基于小波系数模极大方法、基于小波系数尺度相关的方法、基于小波阈值的方法。因为小波系数的模极大值对应信号的突变点(边缘)，信号的奇异性可以通过寻找小波模极大在细尺度下收敛的横坐标点来检测，这一理论奠定了小波系数模极大去噪和由边缘重构信号的理论基础。基于小波系数多尺度间相关性去噪算法依据信号和噪声在不同尺度上小波系数的不同特性：含噪信号在多尺度分解后，真实信号的各尺度系数间具有很强的相关性，尤其是在信号的边缘附近，其相关性更加明显；而噪声对应小波系数的相关性则很弱或者不相关。结合尺度间和尺度内相关性构造相关系数，相关系数大的对应信号，相关系数小的则为噪声。

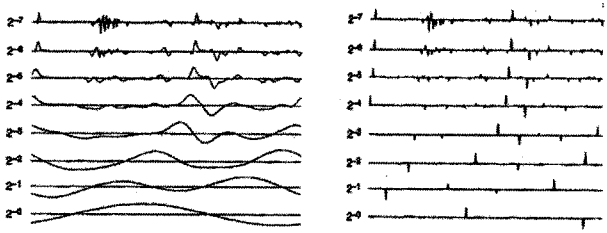
小波阈值去噪方法首先由Donoho提出并给出系统的理论推导，由于其方法简单，计算量小，去噪效果好，近来在图像去噪领域得到了广泛应用。对于受加性噪声污染的信号，经过小波变换后，包含重要信息的小波系数幅值较大，但数量较少，而噪声对应的小波系数幅值较小。通过选取适当的阈值，将小于阈值的小波系数置零，保留大于阈值的小波系数，再进行逆小波变换重构信号。阈值收缩方法的关键是阈值和阈值函数的确定。阈值太大，会丢失许多细节；阈值过小，图像中残留的噪声会更多。阈值有硬阈值、软阈值、半软阈值等，阈值函数的选取准则有基于Bayesian最大后验概率的、尺度自适应的、SURE(Stein Unbiased risk estimation)等。

硬阈值函数将系数幅值大于阈值的系数保留不变，不大于阈值的系数设为零。而软阈值函数具有了一定的连续性(图6.14)。

$$hardT = \begin{cases} c_{j,k} & |c_{j,k}| \geq T \\ 0 & |c_{j,k}| < T \end{cases}$$



(a) 原信号

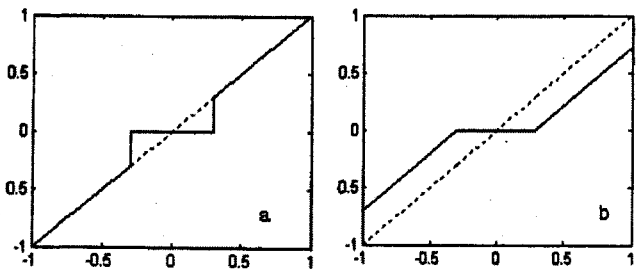


(b) 小波分解低频系数　　(c) 小波分解高频系数

图 6.13: 跟踪小波系数模极大进行边缘检测

$$softT = \begin{cases} c_{j,k} - sgn(c_{j,k}) & |c_{j,k}| \geq T \\ 0 & |c_{j,k}| < T \end{cases}$$

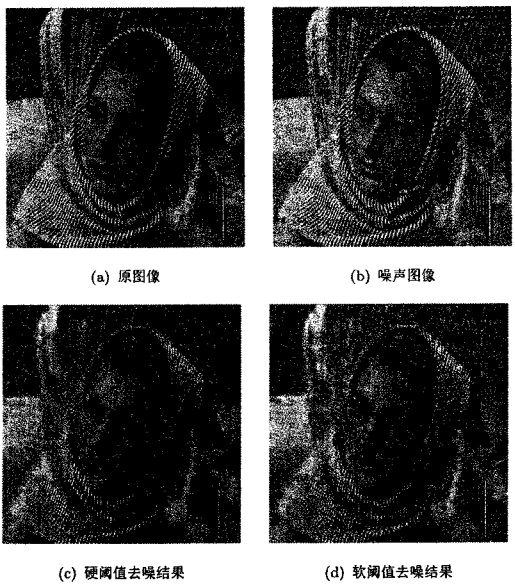
硬阈值方法去噪虽然简单易实现，但会在图像的边界处产生伪Gibbs现象。软阈值方法使得小波系数的结构性得以保持，有效减少了伪Gibbs现象。图(6.15)比较了Woman图像，加方差为0.1的高斯白噪声，经小波变换硬阈值和软阈值去噪的结果。



(a) 硬阈值　　(b) 软阈值

图 6.14: 阈值

当然，此处介绍的基于阈值去噪的小波方法只是最基本的小波分析在去噪方面的应用，近几年来，小波分析与其他方法相结合的信号处理方法受到了更多的专



(a) 原图像 (b) 噪声图像

(c) 硬阈值去噪结果 (d) 软阈值去噪结果

图 6.15: Woman图像

注，也取得了很好的效果，如小波变换与稀疏表示、与深度学习相结合的方法等。可以预见，小波分析将一直活跃在信号处理领域。

第六章习题

1. 分别用Lagrange插值公式表示经过点 $(4,10)$, $(5,5.25)$, $(6,1)$ 的二次代数多项式。

2. 已知 $f(x)=3^x$ 数据表

x_i	-1	0	1	2
$f(x_i)$	$\frac{1}{3}$	1	3	9

分别用Newton二次、三次插值多项式求 $f(\frac{1}{3})$ 的近似值。

3. 已知 $f(x)=\ln x$ 数据表

x_i	0.4	0.5	0.6	0.7	0.8
$\ln x$	-0.916291	-0.693147	-0.510826	-0.356675	-0.223144

作出差分表，用线性插值和二次插值，计算 $\ln 0.54$ 的近似值。

4. 求一个次数不高于四次的多项式 $P(x)$ 使其满足

$$P(0)=P'(0)=0, \qquad P(1)=P'(1)=1, \qquad P(2)=1.$$

并估计其误差。

5. 证明: (1) $\sum_{i=0}^{n-1}\Delta^2 y_i=\Delta y_n-\Delta y_0$. (2) $\Delta(f_i g_i)=f_i\Delta f_i+g_{i+1}\Delta f_i$.

6. 已知实验数据

x_i	0	1	2	3	4
$f(x_i)$	2	2.05	3	9.6	34

已知其经验公式为$y=a+bx^2$, 用最小二乘方法确定参数a, b.

7. 若记母函数$\psi(t)$的傅里叶变换为$\hat{\psi}(\omega)$，计算小波函数$\psi_{a,b}(t)$的傅里叶变换。

第七章　偏微分方程及其数值方法

科学和工程中的大多数实际问题都归结为偏微分方程的定解问题。本章首先介绍了偏微分方程定解问题的几种基本类型，然后介绍了求解简单定解问题的解析解的几种常用基本方法。然而，实际工程问题中绝大多数定解问题的解析解求解是非常困难的（有些问题在经典意义下甚至没有解），因此在本章中，我们重点介绍了求解几类经典的求解偏微分方程的有限差分方法，并对求解偏微分方程的有限元方法的基本原理和过程进行了介绍。

7.1　偏微分方程定解问题

微分方程本质上是函数的某种局部平衡关系，其中含有该函数导数。建立方程的过程主要有三步：先设所求解的量为未知数，然后找出所研究问题满足的等量关系式，最后利用一些基本的关系式将等量关系式两边用已知量和未知数来表示即可。在本节中，我们按照这个过程导出几个经典的实际问题所满足的微分方程和求的方程定解的条件，也使读者能够了解建立微分方程数学模型的基本方法。

7.1.1　波动方程的定解问题

问题的提出：一根长为 l 的柔软、均匀细弦，拉紧之后让它离开平衡位置，在垂直于弦线的外力作用下作微小横振动，求弦线上任一点在任一时刻的位移。

关于问题假设的说明：

(1)“充分柔软”：指当弦线发生变形时只抗伸长而不抗弯曲，即只考虑弦线上不同部分之间张力的相互作用，而对弦线反抗弯曲所产生的力矩忽略不计；

(2)“均匀”：弦线的线密度为常数；

(3)“横振动”：假设弦的运动发生在同一平面内，且弦线上各点位移与平衡位置垂直。

模型分析与建立：以弦线所处的平衡位置为 x 轴，垂直于弦线位置且通过弦线的一个端点的直线为 u 轴建立坐标系如图7.1所示，以 $u(x,t)$ 表示在 t 时刻弦线在横坐标为 x 的点离开平衡位置的位移。设 ρ 为弦的线密度（千克 / 米），f_0 为作用在弦线上且垂直于平衡位置的强迫力密度（牛顿 / 米）。任取一小段弦线（横坐标

从x到$x+\Delta x$的一段，不包括弦线的两个端点），$\vec{T}_1$ 和 $\vec{T}_2$ 分别是弦线在两个端点所受到的张力，即其余部分弦线对该小段弦线的作用力，α_1 为 $\vec{T}_1$ 与水平方向的夹角，α_2 为 $\vec{T}_2$ 与水平方向的夹角。

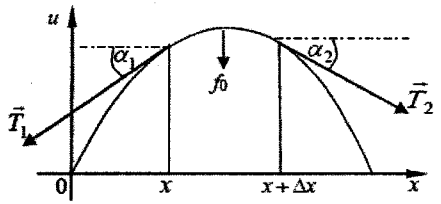


图 7.1: 弦震动示意图

将所取小段弦线近似视为质点,则其运动服从牛顿第二运动定律，即

$$F = ma \tag{7.1}$$

其中 a 为该段弦线运动加速度，m 为该弦段质量。记该段弦线所受垂直于平衡位置（即 u 轴方向）的合外力为 $F = F_1 + F_2 + F_3$，这里 F_1 为 $\vec{T}_1$ 在 u 轴方向的分量，F_2 为 $\vec{T}_2$ 在 u 轴方向的分量，F_3 为强迫外力。

设 $\vec{u}_0$ 为 u 轴正向的单位向量，$\sqrt{1+\left(\frac{\partial u}{\partial x}\right)^2}dx$ 为弦线的弧微分，则容易得到：

$$\begin{cases} F_1 = \vec{T}_1 \cdot \vec{u}_0 = |\vec{T}_1| cos(\vec{T}_1, \vec{u}_0) = |\vec{T}_1| sin\alpha_1 \\ F_2 = \vec{T}_2 \cdot \vec{u}_0 = |\vec{T}_2| cos(\vec{T}_2, \vec{u}_0) = |\vec{T}_2| sin\alpha_2 \\ F_3 = \int_x^{x+\Delta x} f_0(x,t)\sqrt{1+\left(\frac{\partial u}{\partial x}\right)^2}\,dx \end{cases} \tag{7.2}$$

当假设弦线作微小横振动时，α_1 和 α_2 都充分小，从而有

$$\begin{cases} \sin\alpha_1 \sim \tan\alpha_1 = \frac{\partial u(x,t)}{\partial x} \\ \sin\alpha_2 \sim \tan\alpha_2 = -\tan(\pi - \alpha_2) = \frac{\partial u}{\partial x}(x+\Delta x, t) \\ \sqrt{1+\left(\frac{\partial u}{\partial x}\right)^2} = 1 + \frac{1}{2}\left(\frac{\partial u}{\partial x}\right)^2 + o(\left|\frac{\partial u}{\partial x}\right|^2) \approx 1 \end{cases} \tag{7.3}$$

由于假设弦线是“均匀”、“充分柔软”的，可认为弦线每点处张力的方向为弦线的切线方向，弦线各点处张力的大小相等，故有$|\vec{T}_1| = |\vec{T}_2| = T_0$（常数）。

首先，利用(7.2)和(7.3)，可得

$$F = T_0\frac{\partial u}{\partial x}(x+\Delta x, t) - T_0\frac{\partial u}{\partial x}(x,t) + f_0(x,t)\Delta x \tag{7.4}$$

其次，将弦线近似为质点可得

$$a = \frac{\partial^2 u}{\partial t^2}(\bar{x}_1, t),\ \ m = \rho\int_x^{x+\Delta x}\sqrt{1+\left(\frac{\partial u}{\partial x}\right)^2}\,dx \approx \rho\Delta x, \tag{7.5}$$

其中$\bar{x}_1 \in [x, x+\Delta x]$。

然后，将 (7.4) 和 (7.5) 代入到 (7.1) 中，可得

$$\rho \Delta x \frac{\partial^2 u}{\partial t^2}(\bar{x}, t) = T_0 \frac{\partial u}{\partial x}(x+\Delta x, t) - T_0 \frac{\partial u}{\partial x}(x, t) + f_0(x, t)\Delta x \ . \tag{7.6}$$

假设 $u(x,t)$ 具有二阶连续偏导数，对 (7.6) 式右端前二项利用微分中值定理，可得

$$\rho \Delta x \frac{\partial^2 u}{\partial t^2}(\bar{x}_1, t) = T_0 \frac{\partial^2 u}{\partial x^2}(\bar{x}_2, t)\Delta x + f_0(x, t)\Delta x \ , \tag{7.7}$$

其中$\bar{x}_2 \in [x, x+\Delta x]$。

在 (7.7) 式两边同除 Δx，并令 $\Delta x \to 0$，可得到 $u(x,t)$ 所满足的偏微分方程

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} + f(x, t) \ , \tag{7.8}$$

其中$a^2 = \frac{T_0}{\rho}$，$f(x,t) = \frac{f_0(x,t)}{\rho}$。

方程 (7.8) 刻划了柔软均匀细弦在微小横振动时所服从的一般规律，人们称它为弦振动方程。弦线的特定振动情况除满足弦振动方程外，还依赖于初始时刻弦线的状态和在弦线两端所受到的外界约束。因此，为了确定一个具体的弦振动，除了列出它满足的方程 (7.8) 以外，还必须给出适合的初始条件和边界条。

初始条件： 给出了弦线在时刻 $t=0$ 时的关于初始位移 $u(x,0)$ 和初始速度 $u_t(x,t)$ 的约束：

$$u(x,0) = \varphi(x), u_t(x,0) = \psi(x), 0 \le x \le l \tag{7.9}$$

这里 $\varphi(x)$ 和 $\psi(x)$ 是已知函数。

边界条件： 给出了弦线两端 $x=0$ 和 $x=l$ 处的约束条件，最常用的边界约束条件一般有三种：

(1) 第一类边界条件：$u(0,t) = g_1(t)$，$u(l,t) = g_2(t)$，$t \ge 0$，该条件反映了端点的位移变化；

(2) 第二类边界条件：$-T_0 u_x(0,t) = g_1(t)$，$T_0 u_x(l,t) = g_2(t)$，$t \ge 0$，该条件反映了端点所受的垂直于弦线的外力；

(3) 第三类边界条件：$u_x(0,t) - \sigma_1 u(0,t) = g_1(t)$，$u_x(l,t) + \sigma_2 u(l,t) = g_1(t)$，$t \ge 0$，其中 $\sigma_1 = k_1/T_0$，$\sigma_2 = k_2/T_0$，这里 k_1, k_2 分别为两端连接的外界弹性物质弹性系数，该条件反映了端点与弹性物体连接时的弹性约束。

初始条件和边界条件通常称为定解条件，一个微分方程连同它相应的定解条件组成一个定解问题。当考虑的弦线比较长时，一般认为弦长是无穷大.这时定解条件中就没有边界条件而只有初始条件，这也是一个定解问题。

前面的例子是一维弦振动，如果考虑膜的振动或者是声波在空气中的传播，利用和弦振动方程类似的过程可以导出膜振动方程为

$$\frac{\partial^2 u}{\partial t^2} = a^2(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}) + f(x,y,t), \tag{7.10}$$

而声波在空气中传播所满足的方程为

$$\frac{\partial^2 u}{\partial t^2} = a^2(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}) + f(x,y,z,t). \tag{7.11}$$

方程(7.8)、(7.10)、(7.11)一般统称为波动方程（wave equation）。

7.1.2 热传导方程的定解问题

问题的提出： 在三维空间中考虑一个均匀、各向同性的导热体，假定它内部有热源，并且与周围介质有热交换，求物体内部温度的分布。

关于问题假设的说明：

(1) “均匀”：指导热体的密度为常数;

(2) “各向同性”：指导热体内任一点处在各个方向上的传热特性相同，例如当导热体是由同一种金属构成的，我们通常认为它是具有各向同性的。

模型分析与建立：设导热体在空间占据的区域为Ω，边界记为$\partial\Omega$ (如图 7.2 所示)。设导热体的密度为ρ（千克/米），比热为c（焦耳/度·千克），热源强度为$f_0(x,y,z,t)$（焦耳/千克· 秒）。 以$u(x,y,z,t)$ （度）表示导热体t时刻在(x,y,z)点处的温度。

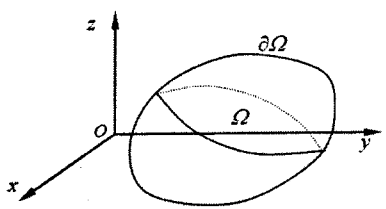


图 7.2:

任取一点$(x,y,z) \in \Omega$，并取该点的一个充分小邻域$G \in \Omega$，G的边界为∂G。在充分小的时段 $[t_1,t_2]$ 上，区域 G 的热量变化满足下面的热量守恒等量关系式:

$$Q_2 - Q_1 = W + \Phi \tag{7.12}$$

其中，Q_1 和 Q_2 分别为 t_1 和 t_2 时刻区域 G 中的热量，W 为在时间区间 $[t_1,t_2]$ 内由 G 内部的内热源生成的总热量，Φ 是在时间区间 $[t_1,t_2]$ 内从 G 的外部通过边界 ∂G 流入到 G内部的热量。

假设温度的变化是连续的，则当区域 G 充分小，时段 $[t_1,t_2]$ 也充分小时，可将 u 视为常数，根据物理学知识可得：

$$\begin{cases} Q_1 = \rho\Delta v \cdot c \cdot u(x_1,y_1,z_1,t_1) \\ Q_2 = \rho\Delta v \cdot c \cdot u(x_1,y_1,z_1,t_2) \\ W = f_0(x_1,y_1,z_1,\bar{t}_1)\cdot\rho\Delta v\cdot\Delta t \end{cases} \tag{7.13}$$

其中$(x_1,y_1,z_1)\in G$，$\bar{t}_1\in[t_1,t_2]$，Δv为区域G的体积，$\Delta t=t_2-t_1$为时间步长。

根据根据Fourier热定律和通量计算公式，还可得

$$\Phi = \iint\limits_{\partial G} \vec{q}\cdot(-\vec{n})ds\Delta t \tag{7.14}$$

其中$\vec{q}=-k(x,y,z)\cdot\nabla u$ 为导热体内热流量，这里$k(x,y,z)$为导热体的导热系数，负号表示热量从温度高处向温度低处流动。$\vec{n}$为 ∂G 的单位外法向量，负号表示计算通过边界 ∂G 的流入热量。

由于我们考虑的是均匀、各向同性的导热体，因此导热体密度和导热系数都是常数，即 $\rho(x,y,z)=\rho$，$k(x,y,z)=k$。另外，假设 u 对空间变量具有二阶连续偏导数，对时间变量具有一阶连续偏导数，则利用高斯公式可得

$$\begin{aligned} \Phi &= \iint\limits_{\partial G}\vec{q}\cdot(-\vec{n})ds\Delta t = \iint\limits_{\partial G} k\nabla u\cdot\vec{n}ds\Delta t \\ &= \iint\limits_{\partial G} k\frac{\partial u}{\partial n}ds\Delta t = k\iiint\limits_{G}\Delta u dv\Delta t \\ &= k\Delta u(x_2,y_2,z_2,\bar{t}_1)\cdot\Delta v\cdot\Delta t \end{aligned} \tag{7.15}$$

其中 $\Delta u = u_{xx}+u_{yy}+u_{zz}$, $(x_2,y_2,z_2)\in\bar{G}$。

将 (7.13) 和 (7.15) 带入到 (7.12)，并在两边同除以 $\Delta v\Delta t$，并令 $\Delta v\to 0,\Delta t\to 0$，可得

$$\frac{\partial u}{\partial t} = a^2\Delta u + f\ , \tag{7.16}$$

其中$a^2=\frac{k}{\rho c}$，$f(x,y,z,t)=\frac{1}{c}f_0(x,y,z,t)$。

为了具体确定某特定物体内部的温度分布，还需要知道初始条件和边界条件:

初始条件： 导热体内在初始时刻 $t=0$ 时的温度分布，即

$$u(x,y,z,0)=\varphi(x,y,z),\quad (x,y,z)\in\bar{\Omega}$$

边界条件： 记 $\sum = \partial\Omega \times [0, +\infty)$，则常用的边界条件有：

(1) 第一类边界条件：$u|_{\Sigma} = g(x, y, z, t)$，该条件反映了边界$\partial\Omega$上的温度分布;

(2) 第二类边界条件：$k\frac{\partial u}{\partial n}\Big|_{\Sigma} = g(x, y, z, t)$，该条件反映了通过边界$\partial\Omega$的热流量，其中$g \geq 0$ 表示流入，$g \leq 0$ 表示流出，$g = 0$ 表示在边界绝热;

(3) 第三类边界条件：

$$\begin{cases} k\dfrac{\partial u}{\partial n} - k_1(u_1 - u) = 0\ , \\ \dfrac{\partial u}{\partial n} + \sigma u = g, (x, y, z, t) \in \Sigma. \end{cases}$$

该条件刻画了导热体通过边界与周围介质的热交换。

方程 (7.16) 刻画了导热体内温度分布和变化所服从的一般规律，人们称其为热传导方程。需要指出的是，热传导方程绝不只是仅用来表示热传导现象。事实上，自然界中还有很多现象都可用方程形如（7.16）的方程来刻画，如分子在诸如空气和水中的扩散等，因此（7.16）也通常被称为扩散方程。

7.1.3 Poisson方程的定解问题

在前面所研究的导热体温度分布问题 (7.8) 中，如果 f 和边界条件中约束条件都与 t 无关，则经过相当长时间后，区域 G 内各点温度趋于定值（不随时间而变化），因此 $u_t = 0$，从而得到

$$-\Delta u = \frac{1}{a^2} f \tag{7.17}$$

该方程称为 Poisson 方程。特别地，当$f = 0$时，称为 Laplace 方程。

由于 u 和 f 与 t 无关，所以 Poisson 方程的定解条件只有边界条件而无初始条件。根据边界条件的不同，可得到如下三种特殊的 Poisson 方程的三个定解问题：

(1)Poisson 方程的第一边值问题（Dirichlet 问题）：

$$\begin{cases} -\Delta u = f,\ (x, y, z) \in \Omega \\ u = \varphi,\ (x, y, z) \in \partial\Omega. \end{cases}$$

(2)Poisson 方程的第二边值问题（Neumann 问题）：

$$\begin{cases} -\Delta u = f,\ (x, y, z) \in \Omega \\ \dfrac{\partial u}{\partial n} = \varphi,\ (x, y, z) \in \partial\Omega\ . \end{cases}$$

(3)Poisson方程的第三边值问题：

$$\begin{cases} -\Delta u = f,\ (x, y, z) \in \Omega \\ (\dfrac{\partial u}{\partial n} + \sigma\ u) = \varphi,\ (x, y, z) \in \partial\Omega\ . \end{cases}$$

在上面的三个定解问题中，$f(x,y,z),\varphi(x,y,z)$都是已知函数。特别地，如果$f(x,y,z)=0$，则称 u 为区域 Ω 内的调和函数，这类函数在偏微分方程理论和应用研究中有着非常重要的作用。在实际工程中，例如当我们考虑带有稳定电流的导体，如果内部无电流源，可以证明导体内的电位势满足Laplace 方程。类似地，对于带有稳定电荷的介质，稳定电荷产生的静电势也满足Poisson方程。因而Laplace方程和Poisson方程有时也称为位势方程。

7.1.4 二阶偏微分方程的分类

为简单起见，先考虑常系数的二阶线性微分方程，其中两个自变量的二阶线性方程的一般形式为

$$a_{11}\frac{\partial^2 u}{\partial {x_1}^2}+2a_{12}\frac{\partial^2 u}{\partial x_1\partial x_2}+a_{22}\frac{\partial^2 u}{\partial x_2^2}+b_1\frac{\partial u}{\partial x_1}+b_2\frac{\partial u}{\partial x_2}+cu=f, \tag{7.18}$$

这里 f 是自变量 x_1,x_2 的函数，$a_{ij}(1\le i,j\le 2),b_i(1\le i\le 2)$ 和 c 均为常数，且有$a_{12}=a_{21}$。

引入下面二阶常系数线性偏微分算子

$$L\cdot=a_{11}\frac{\partial^2 \cdot}{\partial {x_1}^2}+2a_{12}\frac{\partial^2 \cdot}{\partial x_1\partial x_2}+a_{22}\frac{\partial^2 \cdot}{\partial x_2^2}+b_1\frac{\partial \cdot}{\partial x_1}+b_2\frac{\partial \cdot}{\partial x_2}+c\cdot \tag{7.19}$$

则 (7.18) 可简单地表示为 $Lu=f$.

方程的化简主要是简化二阶导数项的形式。设 A 为 (7.19) 中二阶导数项系数所构成的二阶常数矩阵，即

$$A=\begin{pmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{pmatrix}.$$

其中 $a_{12}=a_{21}$。若记梯度算子为 $\nabla_x=(\frac{\partial}{\partial x_1},\frac{\partial}{\partial x_2})^T$，则可得到 (7.19) 的简洁形式为

$$\nabla_x^T A\nabla_x u+(b_1\ b_2)\cdot\nabla_x u+cu=f. \tag{7.20}$$

对于任意二阶矩阵 $B=(b_{11},b_{12};b_{21},b_{22})$，作自变量线性变换 $(y_1,y_2)^T=B(x_1,x_2)^T$，并利用多元复合函数的链导法可得

$$\frac{\partial u}{\partial x_1}=b_{11}\frac{\partial u}{\partial y_1}+b_{21}\frac{\partial u}{\partial y_2},\ \frac{\partial u}{\partial x_2}=b_{12}\frac{\partial u}{\partial y_1}+b_{22}\frac{\partial u}{\partial y_2}$$

从而得到 ∇_x 和 ∇_y 在线性变换下一阶导数的关系为

$$\nabla_x=B^T\nabla_y. \tag{7.21}$$

将 (7.21) 代入到 (7.20) 中，可得

$$\nabla_y^T BAB^T\nabla_y u+(b_1\ b_2)B^T\nabla_y u+cu=f. \tag{7.22}$$

由于 A 为实对称矩阵，故可以通过正交变换将其对角化，即存在正交矩阵 B 使得 BAB^T 为对角阵，从而使 (7.22) 中第一项表示的二阶导数项成为如下标准形式

$$\nabla_y^T BAB^T \nabla_y u = \lambda_1 u_{y_1 y_1} + \lambda_2 u_{y_2 y_2}, \tag{7.23}$$

其中 λ_1, λ_2 为矩阵 A 的特征值。根据标准形式特征值，可将将方程 (7.19) 划归为不同类型：

(1) 当 A 正定或负定，即 λ_1, λ_2 同号，称 (7.19) 为 椭圆型方程；

(2) 当 A 非奇异且 λ_1, λ_2 异号，称 (7.19) 为 双曲型方程;

(3) 当 A 奇异且 λ_1, λ_2 其中之一为零时，称 (7.19) 为 抛物型方程。

对于任意一个带有 n 个自变量常系数的二阶线性微分方程，都可用线性代数中二次型化简方法将方程简化为标准型，其具体过程是：

(1) 首先，求方程二阶导数项系数矩阵 A 的特征向量；

(2) 其次，利用 Schmidt 正交化方法将所得到的 n 个特征向量单位正交化，并以这 n 个单位正交向量作为 n 个列向量生成正交矩阵 B；

(3) 最后，通过一个正交变换 $y = Bx$，在整个 R^n 上就将方程化为标准型了。

对于变系数情形，即下面方程中系数 $a_{ij}(1 \le i, j \le 2), b_i(1 \le i \le 2), c$ 均为自变量 $x = (x_1\ x_2)^T$ 的函数，x 属于平面上的某个区域Ω，此时对应的两个自变量的二阶线性方程的一般形式仍可表示为

$$a_{11}\frac{\partial^2 u}{\partial {x_1}^2} + 2a_{12}\frac{\partial^2 u}{\partial x_1 \partial x_2} + a_{22}\frac{\partial^2 u}{\partial x_2^2} + b_1\frac{\partial u}{\partial x_1} + b_2\frac{\partial u}{\partial x_2} + cu = f \tag{7.24}$$

这里，与 (7.18) 的区别在于这里的系数与自变量有关。

此时，变系数方程 (7.24) 中二阶导数项系数所构成的二阶矩阵为

$$A(x) = \begin{pmatrix} a_{11}(x_1, x_2) & a_{12}(x_1, x_2) \\ a_{21}(x_1, x_2) & a_{22}(x_1, x_2) \end{pmatrix}.$$

记判别式

$$\Delta(x) = -|A(x)| = a_{12}^2(x) - a_{11}(x)a_{22}(x)$$

其中 $|A(x)|$ 为矩阵 $A(x)$ 的行列式。根据该判别式，可定义变系数方程的分类如下：

定义 7.1 设Ω为平面(x_1, x_2)上的某个区域，$x_0 = (x_1^0\ x_2^0)^T \in \Omega$，则有

(1) 若$\Delta(x_0) < 0$，称 (7.24) 在点x_0 为椭圆型方程；

(2) 若$\Delta(x_0) = 0$，称 (7.24) 在点x_0 为抛物型方程；

(3) 若$\Delta(x_0)>0$，称 (7.24) 在点x_0为双曲型方程；

如果 (7.24) 在区域 Ω 的某一子集 E 中每一点都是椭圆型的，就称 (7.24) 在E是椭圆型的。类似地还可定义 E 上的抛物型或双曲型方程。

对于 n 个自变量的二阶线性微分方程，其一般形式为

$$\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j}+\sum_{j=1}^{n}b_j(x)\frac{\partial u}{\partial x_j}+c(x)=f(x)\ . \tag{7.25}$$

其中自变量 $x=(x_1\ x_2\ \cdots\ x_n)^T\in\Omega$，$\Omega$ 为 R^n 的某个区域。

若记 (7.25) 中二阶导数项系数所构成的 n 阶矩阵为 $A(x)=(\ a_{ij}(x)\ )$，则类似于定理7.1，可定义下面的微分方程的分类：

定义 7.2 对于给定的点 $x_0=(x_1^0\ x_2^0\cdots x_n^0)^T\in\Omega$，有

(1) 若 $A(x_0)$ 为正定或负定的，称 (7.25) 在点 x_0 是椭圆型的；

(2) 若 $A(x_0)$ 为非奇异矩阵，且 n 个特征值中 $(n-1)$ 个同号，剩下那一个异号，称 (7.25) 在点x_0是双曲型的；否则，就称(7.25)在点x_0是狭义双曲型的；

(3) 若 $A(x_0)$ 为奇异矩阵，且 n 个特征值中只有一个为零，其余的 $(n-1)$ 个同号，称 (7.25) 在点x_0是抛物型的；否则，就称(7.25)在点x_0是狭义抛物型的；

如果 (7.25) 在区域 Ω 的某一子集 E 中每一点都是椭圆型的，就称(7.25) 在E是椭圆型的。类似可定义 E 上的抛物型或双曲型方程。

根据上述定义，可以很容易验证前面介绍的波动方程通常为双曲型方程，热传导方程通常为抛物型型方程，而 Poisson 方程和 Laplace 方程都是椭圆型方程。

7.2 偏微分方程的解析解

偏微分方程的求解是偏微分方程研究中的核心内容之一，其中偏微分方程的解析解对于偏微分方程的严格理论分析具有重要意义。事实上，偏微分方程的解析解的求解一直是偏微分方程研究中的难点，也没有一套统一的方法。在本节中，我们介绍几种求解偏微分方程的解析解的常用方法，主要包括分离变量法、积分变换法以及格林函数法。

7.2.1 分离变量法

分离变量法的基本思想就是通过变量的分离，将偏微分方程的求解问题转化为

多个常微分方程的求解，从而达到简化计算的目的，这也是偏微分方程解析解求解的最常用的方法之一。

7.2.1.1 齐次方程齐次边界条件定解问题的分离变量法

以下面的一维弦的自由横振动第一类齐次边界条件定解问题为例：

$$\begin{cases} u_{tt}(x,t)-a^2u_{xx}(x,t)=0, & (0\le x\le l, t>0) \\ u(0,t)=0, \quad u(l,t)=0, & (t\ge 0) \\ u(x,0)=\phi(x),\ u_t(x,0)=\psi(x), & (0\le x\le l) \end{cases} \tag{7.26}$$

该问题的方程是齐次的(即方程(7.8)中的$f=0$)，边界条件也是齐次的（即边界初始位移均为零），这也是直接使用分离变量法所需要的两个必要条件。

设问题有变量分离形式的解：

$$u(x,t)=X(x)T(t), \tag{7.27}$$

这里$X(x)$与变量t无关，$T(t)$与变量x无关，将它代入方程（7.26）得到

$$\frac{T''(t)}{a^2T(t)}=\frac{X''(x)}{X(x)} \tag{7.28}$$

这是一个变量分离的恒等式，等式左边仅仅是 t 的函数，等式右边则仅仅是 x 的函数，而 x,t 是两个无关的独立变量，所以这个等式只能是常数（记为 $-\lambda$），于是有

$$\frac{T''(t)}{a^2T(t)}=\frac{X''(x)}{X(x)}=-\lambda, \tag{7.29}$$

从而得到两个常微分方程：

$$\begin{cases} T''(t)+\lambda a^2T(t)=0, \\ X''(x)+\lambda X(x)=0. \end{cases} \tag{7.30}$$

根据 (7.26) 中的齐次边界条件，也可得到

$$X(0)T(t)=0, \qquad X(l)T(t)=0, \tag{7.31}$$

注意到我们需要求得的是非零解，则 $T(t)\neq 0$，从而只有 $X(0)=0, X(l)=0$，由此就得到关于 $X(x)$ 的施斗姆-刘维尔（Sturm-Liouville）本征值问题：

$$\begin{cases} X''(x)+\lambda X(x)=0 \\ X(0)=0, \quad X(l)=0 \end{cases} \tag{7.32}$$

对于该本征值问题，我们有下面结论：

(1) 当 $\lambda<0$ 时，λ 不是问题（7.32）的本征值;

(2) 当 $\lambda=0$ 时，可得 $X(x)=Ax+B$，其中 A,B 为待定常数，由 $X(0)=0$ 得 B=0，另外由 $X(l)=0$ 且 $l\neq 0$，可得 $A=0$，从而有 $X\equiv 0$，这表明 $\lambda=0$ 也不是问题 (7.32) 的本征值;

(3) 当 $\lambda>0$ 时，方程的通解为

$$X(x)=C\cos\sqrt{\lambda}x+D\sin\sqrt{\lambda}x$$

由 $X(0)=0$ 得 $C=0$；由 $X(l)=0$ 得关于 λ 的方程

$$D\sin\sqrt{\lambda}l=0 \tag{7.33}$$

由于求问题的非零解，所以 $D\neq 0$, 因此只有$\sin\sqrt{\lambda}l=0$，从而得到问题（7.32）的可列个本征值：

$$\lambda_n=(\frac{n\pi}{l})^2,(n=1,2,3,...) \tag{7.34}$$

以及对应的本征函数（把非零常数D省去）

$$X_n(x)=\sin\frac{n\pi x}{l},\quad (n=1,2,3,...) \tag{7.35}$$

现将本征值 $\lambda_n=(\frac{n\pi}{l})^2$ 代入关于 $T(t)$ 的方程得到

$$T''_n(t)+(\frac{n\pi a}{l})^2T_n(t)=0,$$

这是一个二阶的常系数线性齐次方程，它的通解为

$$T_n(t)=C_n\cos\frac{n\pi at}{l}+D_n\sin\frac{n\pi at}{l},$$

从而得到变量分离状态的解，称之为驻波：

$$u_n(x,t)=X_n(x)T_n(t)=(C_n\cos\frac{n\pi at}{l}+D_n\sin\frac{n\pi at}{l})\sin\frac{n\pi x}{l}.$$

现在要求满足初始条件的解，一般而言，这可列个驻波解并不满足初始条件，为使得到满足初始条件的混合问题的解，按照叠加原理，将可列个驻波解叠加得到

$$u(x,t)=\sum_{n=1}^{+\infty}(C_n\cos\frac{n\pi at}{l}+D_n\sin\frac{n\pi at}{l})\sin\frac{n\pi x}{l}.$$

于是，由条件 $u(x,0)=\phi(x)$ 可得

$$\phi(x)=\sum_{n=1}^{+\infty}C_n\sin\frac{n\pi x}{l}.$$

注意到本征函数系 $\{\sin\frac{n\pi x}{l}, n=1,2,3...\}$ 在 $[0,l]$ 上是正交的完全的函数系，且$\|\sin\frac{n\pi x}{l}\| = \sqrt{1/2},\ (n=1,2,3...)$，故而有

$$C_n = \frac{2}{l}\int_0^l \phi(\xi)\sin\frac{n\pi\xi}{l}d\xi, \qquad n=1,2,3...$$

同理，由 $u_t(x,0)=\psi(x)$ 可得

$$\psi(x) = \sum_{n=1}^{+\infty}\frac{n\pi a}{l}D_n\sin\frac{n\pi x}{l}$$

从而得到

$$D_n = \frac{2}{n\pi a}\int_0^l \psi(\xi)\sin\frac{n\pi\xi}{l}d\xi, \qquad n=1,2,3...$$

因此分离变量法又叫傅立叶解法.

分离变量法对偏微分方程与边界条件都要进行变量分离，所以方程与边界条件都应是齐次的。在求解过程中会得到施斗姆-刘维尔本征值问题，由此确定可列个本征值与相应的本征函数系，这也是分离变量法的核心问题。

例 7.1 *考虑下面的二维问题：假设边长为 a,b 的矩形薄板，两板面不透热，它的一边 $y=b$ 为绝热，其余三边保持零度温度。设板的初始温度分布是 $\phi(x,y)$，试求板内的温度变化。*

解：以 $u(x,y,t)$ 为此矩形板内点 (x,y) 处时刻 t 的温度，这时此混合问题为：

$$\begin{cases} u_t(x,y,t) = a^2[u_{xx}(x,y,t)+u_{yy}(x,y,t)], (0\le x\le a, 0\le y\le b) \\ u|_{x=0}=0, \quad u|_{x=a}=0, \quad u|_{y=0}=0, \quad u_y|_{y=b}=0, \\ u|_{t=0}=\phi(x,y). \end{cases}$$

这是一个齐次方程、齐次边界条件的问题，可设有变量分离形式的解。

（1）设解为

$$u(x,y,t) = X(x)Y(y)T(t),$$

代入齐次方程中，分离变量后有

$$\frac{T'(t)}{a^2T(t)} = \frac{X''(x)}{X(x)} + \frac{Y''(y)}{Y(y)}.$$

由于变量 x,y,t 都是独立变量，所以令

$$\frac{X''(x)}{X(x)} = -\lambda, \quad \frac{Y''(y)}{Y(y)} = -\mu,$$

这里λ, μ都是待定的常数，并且

$$T'(t)+a^2(\lambda+\mu)T(t)=0. \tag{7.36}$$

齐次边界条件分离变量后得

$$X(0)=X(a)=0, \quad Y(0)=0, \quad Y'(0)=0,$$

于是得到两个施斗姆-刘维尔本征值问题：

$$\begin{cases} X''(x)+\lambda X(x)=0, \\ X(0)=0,\ X(a)=0. \end{cases}$$

与

$$\begin{cases} Y''(y)+\mu Y(y)=0, \\ Y(0)=0,\ Y'(b)=0. \end{cases}$$

（2）解上述两个本征值问题，容易得到它的本征值和相应的本征函数系.

$$\lambda_n=(\frac{n\pi}{a})^2, X_n(x)=\sin\frac{n\pi x}{a}, \quad (n=1,2,3,...)$$

$$\mu_m=(m+\frac{1}{2})^2\frac{\pi^2}{b^2}, \ Y_m(y)=\sin(m+\frac{1}{2})\frac{\pi y}{b}, \quad (m=0,1,2,...)$$

（3）将本征值代入关于$T(t)$的一阶方程 (7.36) 中，得到

$$T_{nm}(t)=C_{nm}e^{-\left[\frac{n^2}{a^2}+\frac{1}{b^2}(m+\frac{1}{2})^2\right]a^2\pi^2t}$$

（4）为得到满足初始条件的混合问题的解，将

$$u_{nm}(x,y,t)=X_n(x)Y_m(y)T_{nm}(t)$$

对nm叠加，有

$$u(x,y,t)=\sum_{n=1}^{+\infty}\sum_{m=0}^{+\infty}C_{nm}e^{-\left[\frac{n^2}{a^2}+\frac{1}{b^2}(m+\frac{1}{2})^2\right]a^2\pi^2t}\sin\frac{n\pi x}{a}\sin(m+\frac{1}{2})\frac{\pi y}{b}$$

由初始条件得

$$\phi(x,y)=\sum_{n=1}^{+\infty}\sum_{m=0}^{+\infty}C_{nm}\sin\frac{n\pi x}{a}\sin(m+\frac{1}{2})\frac{\pi y}{b}$$

同样，函数系 $\{\sin\frac{n\pi x}{l}, n=1,2,3...\}$ 在 $[0,l]$ 上是正交的完全的函数系，且$\|\sin\frac{n\pi x}{l}\|=\sqrt{1/2}$, $(n=1,2,3...)$，从而得

$$C_{nm}=\frac{4}{ab}\int_0^a\int_0^b\phi(\xi,\eta)\sin\frac{n\pi\xi}{a}\sin(m+\frac{1}{2})\frac{\pi\eta}{b}d\xi d\eta.$$

7.2.1.2 非齐次方程齐次边界条件的解法

这里把分离变量法推广用于求解非齐次方程齐次边界条件的定解问题，介绍的方法叫本征函数法，其基本思想就是将函数在齐次方程齐次边界问题的本征函数系下进行表示，然后利用本征函数系的正交性来求出表示系数，从而得到方程的解。

例 7.2 *考虑一维波动方程的强迫振动问题，具有第一类齐次边界条件的定解问题*

$$\begin{cases} u_{tt}(x,t)-a^2u_{xx}(x,t)=f(x,t), \quad (0\le x<l, t>0) \\ u(0,t)=0, \quad u(l,t)=0\ , \ (t>0) \\ u(x,0)=\phi(x), \quad u_t(x,0)=\psi(x), \ (0\le x<l) \end{cases}$$

解：（1）首先，得到对应齐次方程齐次边界条件的问题为：

$$\begin{cases} u_{tt}(x,t)-a^2u_{xx}(x,t)=0, \quad (0\le x<l, t>0) \\ u(0,t)=0, \quad u(l,t)=0\ , \ (t>0) \\ u(x,0)=\phi(x), \quad u_t(x,0)=\psi(x), \ (0\le x<l) \end{cases}$$

并求出本征值与本征函数系为

$$\lambda_n=\left(\frac{n\pi}{l}\right)^2, \ X_n(x)=\sin\frac{n\pi x}{l}, \quad (n=1,2,3,...)$$

（2）设问题的解 $u(x,t)$ 在本征函数系 $\left\{X_n(x)=\sin\frac{n\pi x}{l}\right\}$, $(n=1,2,3,...)$ 的傅立叶级数表示为

$$u(x,t)=\sum_{n=1}^{+\infty}T_n(t)\sin\frac{n\pi x}{l}$$

其中 $T_n(t)$ $(n=1,2,3,...)$ 是待定的函数。

为了求出函数 $T_n(t)$，将已知函数 $f(x,t)$, $\phi(x)$, $\psi(x)$ 分别在本征函数系 $\left\{\sin\frac{n\pi x}{l}\right\}$, $(n=1,2,3,...)$ 下展开为傅立叶级数：

$$f(x,t)=\sum_{n=1}^{+\infty}f_n(t)\sin\frac{n\pi x}{l}, \ \phi(x)=\sum_{n=1}^{+\infty}\phi_n\sin\frac{n\pi x}{l}, \ \psi(x)=\sum_{n=1}^{+\infty}\psi_n\sin\frac{n\pi x}{l}$$

其中

$$f_n(t)=\frac{2}{l}\int_0^l f(\xi,t)\sin\frac{n\pi\xi}{l}d\xi, \ \phi_n=\frac{2}{l}\int_0^l\phi(\xi)\sin\frac{n\pi\xi}{l}d\xi, \ \psi_n=\frac{2}{l}\int_0^l\psi(\xi)\sin\frac{n\pi\xi}{l}d\xi$$

都是已知的。

在上述傅立叶级数展开基础上，将 $u(x,t)=\sum_{n=1}^{+\infty}T_n(t)\sin\frac{n\pi x}{l}$ 也代入方程与初始条件中，可得关于$T_n(t)$的常微分方程初值问题：

$$\begin{cases} T''_n(t)+\left(\frac{n\pi a}{l}\right)^2T_n(t)=f_n(t) \\ T_n(0)=\phi_n, \quad T'_n(0)=\psi_n \end{cases}, (n=1,2,3...)$$

用拉普拉斯积分变换解得

$$T_n(t)=\phi_n(t)\cos\frac{n\pi at}{l}+\frac{l}{n\pi a}\psi_n\sin\frac{n\pi at}{l}+\frac{l}{n\pi a}\int_0^t f_n(\tau)\sin\frac{n\pi a(t-\tau)}{l}d\tau.$$

这样就得到一维波动方程强迫振动齐次边界条件的解

$$\begin{aligned}u(x,t)&=\sum_{n=1}^{+\infty}T_n(t)\sin\tfrac{n\pi x}{l}\\&=\sum_{n=1}^{+\infty}(\phi_n\cos\tfrac{n\pi at}{l}+\psi_n\tfrac{l}{n\pi a}\sin\tfrac{n\pi at}{l})\sin\tfrac{n\pi x}{l}\\&+\sum_{n=1}^{+\infty}\left[\tfrac{l}{n\pi a}\int_0^t f_n(\tau)\sin\tfrac{n\pi a(t-\tau)}{l}d\tau\right]\sin\tfrac{n\pi x}{l}.\end{aligned}$$

除了上述利用傅里叶级数展开的本征值方法外，对于非齐次方程的求解，还有一种方法称为齐次化方法。该方法的理论基础被称为齐次化原理，其基本思想是通过变量替换将非齐次问题转化为齐次方程的求解。对于该方法本章不作详细介绍，感兴趣的读者可自行学习。

7.2.1.3 非齐次边界条件的定解问题的解法

对于非齐次边界条件的定解问题的解法，应先将边界条件齐次化，把定解问题转化为齐次边界条件，用前面叙述的方法求解。

例 7.3 设长为 l，端点按某种规律依时间t变化的弦强迫振动，弦振动位移 $u(x,t)$ 满足定解问题：

$$\begin{cases}u_{tt}-a^2u_{xx}=f(x,t), & (0<x<l,t>0)\\ u(0,t)=\mu(t), \quad u(l,t)=\nu(t), & (t>0)\\ u(x,0)=\phi(x), \quad u_t(x,0)=\psi(x), & (0<x<l)\end{cases}\tag{7.37}$$

分析：为了能用本征函数法求解，先将边界条件齐次化，为此设

$$u(x,t)=v(x,t)+w(x,t),$$

这里 $v(x,t)$ 是引进的新的函数，使 $v(x,t)=u(x,t)-w(x,t)$ 满足齐次边界条件 $v(0,t)=0, \quad v(l,t)=0$，就必须让函数 $w(x,t)$ 满足

$$w(0,t)=\mu(t), \quad w(l,t)=\nu(t).$$

对此最简单的取法是令 $w(x,t)=A(t)x+B(t)$，其中 $A(t)=\frac{1}{l}[\nu(t)-\mu(t)]$，$B(t)=\mu(t)$，从而得到

$$w(x,t)=\frac{x}{l}[\nu(t)-\mu(t)]+\mu(t).$$

这样，若 $u(x,t)$ 是定解问题 (7.37) 的解，那么 $v(x,t)$ 满足定解问题

$$\begin{cases} v_{tt}-a^2v_{xx}=f(x,t)-(w_{tt}-a^2w_{xx}), \quad (0<x<l,t>0) \\ v(0,t)=0, \quad v(l,t)=0, \quad (t>0) \\ v(x,0)=\phi(x)-w(x,0), \quad v_t(x,0)=\psi(x)-w_t(x,0), \ (0<x<l) \end{cases}$$

这个问题就可以用本征函数法或齐次化原理求解。

应当指出，偏微分方程理论上已经证明了$u(x,t)$ 的解与 $w(x,t)$ 的选取法无关。这里仅给出了对边界条件是第一类边界条件的 $w(x,t)$ 的选取形式，对其他类型的边界条件完全可以用类似的方法找出相应的 $w(x,t)$，使边界条件齐次化。常用的几种边界条件下对应的函数 $w(x,t)$ 的表达式为：

（1）若 $u(0,t)=\mu(t), \quad u_x(l,t)=\nu(t)$，则$w(x,t)$ 可取为

$$w(x,t)=x\nu(t)+\mu(t).$$

（2）若 $u_x(0,t)=\mu(t), \quad u(l,t)=\nu(t)$，则$w(x,t)$ 可取为

$$w(x,t)=(x-l)\nu(t)+\mu(t).$$

（3）若 $u_x(0,t)=\mu(t), \quad u_x(l,t)=\nu(t)$，则 $w(x,t)$ 可取为

$$w(x,t)=x\mu(t)+\frac{x^2}{2l}[\nu(t)-\mu(t)].$$

（4）若 $u_x(0,t)-\sigma u(0,t)=\mu(t), \quad u_x(l,t)+\sigma u(l,t)=\nu(t)$，这里 $\sigma>0$，则 $w(x,t)$ 可取为

$$w(x,t)=x\mu(t)+x^2\frac{\nu(t)-(1+\sigma l)\mu(t)}{2l+\sigma l^2}.$$

例 7.4 求解混合问题

$$\begin{cases} u_t-a^2u_{xx}=e^{-\frac{a^2}{4}t}\sin\frac{x}{2}, \quad (0<x<\pi,t>0) \\ u(0,t)=0, \quad u_x(\pi,t)=b, \\ u(x,0)=\phi(x) \end{cases}$$

解:（1）先把边界条件齐次化

令 $u(x,t)=v(x,t)+bx$，则该问题转化为关于 $v(x,t)$ 的定解问题

$$\begin{cases} v_t-a^2v_{xx}=e^{-\frac{a^2}{4}t}\sin\frac{x}{2}, \quad (0<x<\pi,t>0) \\ v(0,t)=0, \quad v_x(\pi,t)=0 \\ v(x,0)=\phi(x)-bx \end{cases}$$

（2）用本征函数法或齐次化原理解此问题得

$$v(x,t)=te^{-\frac{a^2}{4}t}\sin\frac{x}{2}+\sum_{n=0}^{+\infty}a_ne^{-\frac{(2n+1)^2a^2}{4}t}\sin\frac{(2n+1)x}{2}$$

其中 $a_n=\frac{2}{\pi}\int\limits_0^{\pi}[\phi(\xi)-\beta\xi]\sin\frac{(2n+1)\xi}{2}d\xi,\quad(n=0,1,2,...)$

最后得原问题的解为

$$u(x,t)=bx+te^{-\frac{a^2}{4}t}\sin\frac{x}{2}+\sum_{n=0}^{+\infty}a_ne^{-\frac{(2n+1)^2a^2}{4}t}\sin\frac{(2n+1)x}{2}.$$

7.2.2 积分变换法

积分变换是在偏微分方程求解中的常用方法之一，通过积分变换可以把偏微分方程的定解问题化为常微分方程的定解问题，从而使问题得以解决。目前，最常用的积分变换主要有傅里叶变换和 Laplace 变换。

7.2.2.1 傅立叶积分变换在偏微分方程定解问题中的应用

傅立叶积分变换是一种非常重要的积分变换，在这里我们一维波动方程的定解问题的求解为例，来说明傅立叶积分变换在偏微分方程定解问题中的应用。

例 7.5 *求解无界弦振动方程的初值问题*

$$\begin{cases}u_{tt}-a^2u_{xx}=0,\ (-\infty<x<+\infty,t>0)\\ u(x,0)=\phi(x),\ (-\infty<x<+\infty)\\ u_t(x,0)=\psi(x),\ (-\infty<x<+\infty)\end{cases}$$

解：设 $u(x,t)$ 关于变量 x 的傅立叶变换为 $\mathfrak{F}[u(x,t)]=U(\omega,t)$，则定解问题经傅立叶变换后为

$$\begin{cases}\frac{d^2U(\omega,t)}{dt^2}+a^2\omega^2U(\omega,t)=0\\ U(\omega,0)=\Phi(\omega),\quad U_t(\omega,0)=\psi(\omega)\end{cases}$$

这里 $\Phi(\omega)=\mathfrak{F}[\phi(x)]$，$\Psi(\omega)=\mathfrak{F}[\psi(x)]$。

这是一个含参数的二阶常微分方程的初值问题，容易求得该方程的通解为

$$U(\omega,t)=C_1e^{ia\omega t}+C_2e^{-ia\omega t}.$$

由初始条件，易得 $C_1+C_2=\Phi(\omega),\quad i\omega a(C_1-C_2)=\Psi(\omega)$，从而联立求解可得到

$$C_1=\frac{1}{2}\left[\Phi(\omega)+\frac{1}{i\omega a}\Psi(\omega)\right],\quad C_2=\frac{1}{2}\left[\Phi(\omega)-\frac{1}{i\omega a}\Psi(\omega)\right],$$

由此得到

$$U(\omega,t)=\frac{1}{2}\left[\Phi(\omega)+\frac{1}{i\omega a}\Psi(\omega)\right]e^{i a\omega t}+\frac{1}{2}\left[\Phi(\omega)-\frac{1}{i\omega a}\Psi(\omega)\right]e^{-i a\omega t}.$$

对上式等号两边同时进行傅立叶逆变换，可得

$$u(x,t)=\frac{1}{2}\mathfrak{F}^{-1}\left[\Phi(\omega)e^{i a\omega t}\right]+\frac{1}{2a}\mathfrak{F}^{-1}\left[\frac{1}{i\omega}\Psi(\omega)e^{i a\omega t}\right]$$

$$+\frac{1}{2}\mathfrak{F}^{-1}\left[\Phi(\omega)e^{-i a\omega t}\right]-\frac{1}{2a}\mathfrak{F}^{-1}\left[\frac{1}{i\omega}\Psi(\omega)e^{-i a\omega t}\right]$$

根据延迟定理和积分定理，容易得到

$$\mathfrak{F}^{-1}[e^{\pm i\omega a t}\Phi(\omega)]=\phi(x\pm at),\ \mathfrak{F}^{-1}\left[e^{\pm i\omega a t}\frac{1}{i\omega}\Psi(\omega)\right]=\int_{\infty}^{\pm}\psi(\xi)\,d\xi,$$

从而得到定解问题的解为

$$\begin{aligned}u(x,t)&=\tfrac{1}{2}[\phi(x-at)+\phi(x+at)]+\tfrac{1}{2a}\left[\int_{-\infty}^{x+at}\psi(\xi)d\xi-\int_{-\infty}^{x-at}\psi(\xi)d\xi\right]\\&=\tfrac{1}{2}[\phi(x-at)+\phi(x+at)]+\tfrac{1}{2a}\int_{x-at}^{x+at}\psi(\xi)d\xi.\end{aligned}$$

7.2.2.2 拉普拉斯变换在数学物理定解问题中的应用

Laplace变换是一种重要的积分变换形式，在这里我们以一维热传导问题来说明拉普拉斯变换在数学物理定解问题中的应用。

例 7.6 *求解一维无界空间的有源的热传导问题，且初始温度已知，即解定解问题：*

$$\begin{cases}u_t-a^2u_{xx}=f(x,t), & (-\infty<x<+\infty,t>0)\\ u(x,0)=\phi(x), & (-\infty<x<+\infty)\end{cases}$$

解：基于Laplace 变换，记

$$\begin{cases}\mathfrak{L}[u(x,t)]=U(x,p)=\int_0^{+\infty}u(x,t)e^{-pt}dt\\ \mathfrak{L}[f(x,t)]=F(x,p)=\int_0^{+\infty}f(x,t)e^{-pt}dt\end{cases}$$

方程两边作拉普拉斯变换，有

$$pU(x,p)-\phi(x)-a^2\frac{d^2U(x,p)}{dx^2}=F(x,p),$$

化简为

$$\frac{d^2U(x,p)}{dx^2}-\frac{p}{a^2}U(x,p)=-\frac{1}{a^2}\left[\phi(x)+F(x,p)\right],$$

由常数变易法得解

$$U(x,p)=c_1e^{\frac{\sqrt{p}}{a}x}+c_2e^{-\frac{\sqrt{p}}{a}x}-\frac{1}{2a\sqrt{p}}e^{\frac{\sqrt{p}}{a}x}\int[\phi(\xi)+F(\xi,p)]\,e^{-\frac{\sqrt{p}}{a}\xi}d\xi$$
$$+\frac{1}{2a\sqrt{p}}e^{-\frac{\sqrt{p}}{a}x}\int[\phi(\xi)+F(\xi,p)]\,e^{\frac{\sqrt{p}}{a}\xi}d\xi\,.$$

由自然边界条件 $\lim\limits_{|x|\to+\infty}u(x,p)$有界，故有 $\lim\limits_{|x|\to+\infty}U(x,p)$有界，从而$c_1=c_2=0$，因此得解

$$\begin{aligned}U(x,p)&=-\frac{1}{2a\sqrt{p}}\int_{+\infty}^{x}e^{-\frac{\sqrt{p}}{a}(\xi-x)}\left[\phi(\xi)+F(\xi,p)\right]d\xi\\&\quad+\frac{1}{2a\sqrt{p}}\int_{-\infty}^{x}e^{-\frac{\sqrt{p}}{a}(x-\xi)}\left[\phi(\xi)+F(\xi,p)\right]d\xi\\&=\frac{1}{2a\sqrt{p}}\left[\int_{-\infty}^{x}e^{-\frac{\sqrt{p}}{a}(x-\xi)}\phi(\xi)d\xi+\int_{x}^{+\infty}e^{\frac{\sqrt{p}}{a}(x-\xi)}\phi(\xi)d\xi\right]\\&\quad+\frac{1}{2a\sqrt{p}}\left[\int_{-\infty}^{x}F(\xi,p)e^{-\frac{\sqrt{p}}{a}(x-\xi)}d\xi+\int_{x}^{+\infty}F(\xi,p)e^{\frac{\sqrt{p}}{a}(x-\xi)}d\xi\right]\end{aligned}$$

注意到 $\mathcal{L}^{-1}\left[\frac{1}{\sqrt{p}}e^{-a\sqrt{p}}\right]=\frac{1}{\sqrt{\pi t}}e^{-\frac{a^2}{4t}}$，利用卷积性质，可得原定解问题的解为：

$$\begin{aligned}u(x,t)&=\frac{1}{2a\sqrt{\pi t}}\left[\int_{-\infty}^{x}\phi(\xi)e^{-\frac{(x-\xi)^2}{4a^2t}}d\xi+\int_{x}^{+\infty}\phi(\xi)e^{-\frac{(\xi-x)^2}{4a^2t}}d\xi\right]\\&\quad+\frac{1}{2a\sqrt{\pi}}\left[\int_{-\infty}^{x}\int_{0}^{t}\frac{f(\xi,\tau)e^{-\frac{(x-\xi)^2}{4a^2(t-\tau)}}}{\sqrt{t-\tau}}d\tau d\xi+\int_{x}^{+\infty}\int_{0}^{t}\frac{f(\xi,\tau)e^{-\frac{(\xi-x)^2}{4a^2(t-\tau)}}}{\sqrt{t-\tau}}d\tau d\xi\right]\\&=\frac{1}{2a\sqrt{\pi t}}\int_{-\infty}^{+\infty}\phi(\xi)e^{-\frac{(x-\xi)^2}{4a^2t}}d\xi+\frac{1}{2a\sqrt{\pi}}\int_{0}^{t}d\tau\int_{-\infty}^{+\infty}\frac{f(\xi,\tau)e^{-\frac{(x-\xi)^2}{4a^2(t-\tau)}}}{\sqrt{t-\tau}}d\xi\,.\end{aligned}$$

7.2.3 格林函数法

本节利用高等数学中的格林(Green)公式导出调和函数的积分表达式，引进格林函数(又叫点源函数，是一种广义函数)，并利用格林函数来求解稳态边值问题，这种方法叫格林函数法，它是解数学物理问题时常用的方法之一.

(1) 格林公式

设 D 是以分片光滑的曲面 S 为其边界的有界区域，函数 $P(x,y,z)$，$Q(x,y,z)$，$R(x,y,z)$ 是在 $\overline{D}$ 上连续，在区域 D 内有连续偏导数的任意函数，则成立格林公式

$$\iiint_D\left(\frac{\partial P}{\partial x}+\frac{\partial Q}{\partial y}+\frac{\partial R}{\partial z}\right)dV=\oiint_S[P\cos(n,x)+Q\cos(n,y)+R\cos(n,z)]dS \quad (7.38)$$

这里 dV 是体积元，n 是曲面 S 的外法线方向，dS 为 S 上的面积元.

设函数 $u(x,y,z),v(x,y,z)$ 以及它们的所有的一阶偏导数在闭区域 $\bar{D}=D\bigcup S$ 上是连续的，u 和 v 在 D 内具有连续的二阶偏导数。令

$$P=u\frac{\partial v}{\partial x},\ Q=u\frac{\partial v}{\partial y},\ R=u\frac{\partial v}{\partial z}$$

带入到格林公式中，可得

$$\iiint_D (u\Delta v - v\Delta u)dV = \oiint_S (u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n})dS. \tag{7.39}$$

其中 Δ 是拉普拉斯(Laplace)算子，$\frac{\partial}{\partial n}$ 表示曲面S的外法线方向导数。

公式 (7.38) 通常也被称为第一格林公式，公式 (7.39) 则被称为第二格林公式。

（2）拉普拉斯方程的基本解

在三维空间内，若记 $r = \sqrt{(x-\xi)^2 + (y-\eta)^2 + (z-\varsigma)^2} = r(M,N)$ 表示点 $M(x,y,z)$ 和 $N(\xi,\eta,\varsigma)$ 之间的距离，利用复合函数求导的链式法则，对空间中任意固定的一点 N，函数 $\frac{1}{r}$ 除点 N 外关于变量x,y,z 处处满足如下的拉普拉斯方程：

$$\Delta(\frac{1}{r}) = 0, \quad M \neq N. \tag{7.40}$$

函数 $\frac{1}{r}$ 在求解拉普拉斯方程和泊松(Poisson)方程时有极重要的作用，通常把函数 $\frac{1}{r}$ 称为三维拉普拉斯方程或者泊松方程的基本解。

对于二维空间，函数

$$\ln\frac{1}{r} = \ln\frac{1}{\sqrt{(x-\xi)^2 + (y-\eta)^2}} = \ln\frac{1}{r(M,N)}$$

叫做二维拉普拉斯方程或泊松方程的基本解.

(3) 调和函数的积分表达式

还是以三维问题为例，利用格林公式不难得到三维空间调和函数的积分表达式。

定理 7.1 （调和函数的积分表达式）设函数$u(x,y,z)$在闭区域$\bar{D}$上有连续的一阶偏导数，且$u(x,y,z)$ 在区域D内调和（即$\Delta u = 0$ 在D内成立），那么对于D内任意固定的一点$M_0(x_0,y_0,z_0)$ 就有

$$u(M_0) = \frac{1}{4\pi}\oiint_S [\frac{1}{r}\frac{\partial u}{\partial n} - u\frac{\partial(\frac{1}{r})}{\partial n}]dS, \ M_0 \in D, \tag{7.41}$$

这里 $r = r(M,M_0)$ 为点 $M(x,y,z)$ 到 $M_0(x,y,z)$ 的距离。

证明： 事实上，设 $M_0(x_0,y_0,z_0)$ 为区域 D 内任意固定的一点，$M(x,y,z)$ 为 $\bar{D}$ 上的一个动点，动点 M 到定点M_0 的距离

$$r = r(M,M_0) = \sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2}$$

注意到函数 $\frac{1}{r}$ 的调和区域并不包含 M_0，因此为了要应用格林公式，必须将点 M_0 挖去。为此，以 M_0 点为球心，充分小的正数 $\rho > 0$) 为半径作球体 $B_\rho^{M_0}$，用 $S_\rho^{M_0}$ 表示这个小球 $B_\rho^{M_0}$ 的球面。记区域 $D_1 = D \setminus B_\rho^{M_0}$ (通常称区域 D 内挖去点 M_0)，这时区域 D_1 的表面为 $S \bigcup S_\rho^{M_0}$。

于是，函数 $u, v = \frac{1}{r}$ 在闭区域 $\bar{D}_1 = D_1 \bigcup S \bigcup S_\rho^{M_0}$ 上可用格林公式，从而可得到

$$\iiint_{D_1} [u\Delta(\frac{1}{r}) - \frac{1}{r}\Delta u]dV = \oiint_S (u\frac{\partial(\frac{1}{r})}{\partial n} - \frac{1}{r}\frac{\partial u}{\partial n})dS + \oiint_{S_\rho^{M_0}} (u\frac{\partial(\frac{1}{r})}{\partial n} - \frac{1}{r}\frac{\partial u}{\partial n})dS$$

因为在区域 D_1 内$\Delta u = 0, \Delta(\frac{1}{r}) = 0$，因此上式左边等于零，从而

$$\oiint_S (u\frac{\partial(\frac{1}{r})}{\partial n} - \frac{1}{r}\frac{\partial u}{\partial n})dS + \oiint_{S_\rho^{M_0}} u\frac{\partial(\frac{1}{r})}{\partial n}dS - \oiint_{S_\rho^{M_0}} \frac{1}{r}\frac{\partial u}{\partial n}dS = 0$$

我们注意到对区域 D_1 而言，小球面 $S_\rho^{M_0}$ 的外法线方向应指向球心 M_0，与半径 r 的方向刚好相反，因此在球面 $S_\rho^{M_0}$ 上有

$$\frac{\partial(\frac{1}{r})}{\partial n} = -\frac{\partial(\frac{1}{r})}{\partial r} = \frac{1}{r^2} = \frac{1}{\rho^2}$$

这样上式第二项积分有

$$\oiint_{S_\rho^{M_0}} u\frac{\partial(\frac{1}{r})}{\partial n}dS = \oiint_{s_\rho^{M_0}} \frac{u}{\rho^2}dS = \frac{1}{\rho^2}u(M_1)4\pi\rho^2 = 4\pi\, u(M_1)$$

这里用到积分中值定理，M_1 为球面 $S_\rho^{M_0}$ 上的某一点。

对于上式第三项积分，用积分中值定理有

$$\oiint_{S_\rho^{M_0}} \frac{1}{r}\frac{\partial u}{\partial n}dS = \frac{1}{\rho}\cdot 4\pi\rho^2 \cdot \frac{\partial u}{\partial n}|_{M_2} = 4\pi\rho \cdot \frac{\partial u}{\partial n}|_{M_2}$$

这里M2为$S_\rho^{M_0}$上的某一点。

又因为 $\frac{\partial u}{\partial n}$ 在 M_0 点的邻域内是有界的，让 $\rho \to 0$，则 M_1, M_2 趋于球心 M_0，所以第三项积分也趋于零，由此得到

$$\oiint_S (u\frac{\partial(\frac{1}{r})}{\partial n} - \frac{1}{r}\frac{\partial u}{\partial n})dS + 4\pi\, u(M_0) = 0.$$

从而得到有界区域 D 内调和函数 u 的积分表达式：

$$u(M_0) = \frac{1}{4\pi}\oiint_S (\frac{1}{r}\frac{\partial u}{\partial n} - u\frac{\partial(\frac{1}{r})}{\partial n})dS, M_0 \in D.$$

公式 (7.41)说明，调和函数 u 在区域 D 内任意一点 M_0 处的值可以由它的边界 S 上的值和它在边界 S 上的法向导数 $\frac{\partial u}{\partial n}$ 的值来确定，这对解边值问题提供了方便。

推论 7.1 *若 u 在有界区域D内是二阶连续的可微函数，则有积分表达式*

$$u(M_0)=\frac{1}{4\pi}\iint_S(\frac{1}{r}\frac{\partial u}{\partial v}-u\frac{\partial(\frac{1}{r})}{\partial n})dS-\frac{1}{4\pi}\iiint_D\frac{\Delta u}{r}dV,\quad M_0\in D. \tag{7.42}$$

这是因为在闭区域$\bar{D}_1$上用格林公式，有

$$-\iiint_{D_1}\frac{1}{r}\Delta u dV=\iint_S(u\frac{\partial(\frac{1}{r})}{\partial n}-\frac{1}{r}\frac{\partial u}{\partial n})dS+\iint_{S_\rho^{M_0}}(u\frac{\partial(\frac{1}{r})}{\partial n}-\frac{1}{r}\frac{\partial u}{\partial n})dS$$

类似上述的讨论，上式右端当 $\rho\to 0$ 时，区域 $D_1\to D$，从而可以得到推论的结论。

对于二维情形，由于基本解为 $\ln\frac{1}{r}$，可得到在二维有界区域 D 内调和的函数 u 的积分表达式：

$$u(M_0)=\frac{1}{2\pi}\oint_C[\ln(\frac{1}{r})\frac{\partial u}{\partial n}-u\frac{\partial(\ln\frac{1}{r})}{\partial n}]dS, M_0\in D, \tag{7.43}$$

这里 C 为区域 D 的边界.

对一般的在区域 D 内有二阶连续可微函数 u，则积分表达式为

$$u(M_0)=\frac{1}{2\pi}\oint_C[\ln(\frac{1}{r})\frac{\partial u}{\partial n}-u\frac{\partial(\ln\frac{1}{r})}{\partial n}]dl-\frac{1}{2\pi}\iint_D(\ln\frac{1}{r})\Delta u dS, M_0\in D. \tag{7.44}$$

本章对于对于公式 (7.43) 和 (7.44) 的证明不做详细赘述，作为习题留给读者自己去证明。

(4) 狄里克雷问题的格林函数法

我们下面的狄里克雷问题来讨论格林函数法的应用：

$$\begin{cases}\Delta u=0, & \text{M在区域D内}\\ u|_S=f_1(M) & \text{M在边界S上}\end{cases}$$

从调和函数 u 的积分表达式出发，在区域 D 内的调和函数 u 的积分表达式为：

$$u(M_0)=\frac{1}{4\pi}\iint_S(\frac{1}{r}\frac{\partial u}{\partial n}-u\frac{\partial(1/r)}{n})dS,\quad M_0\in D.$$

对于狄里克雷问题，积分表达式中的第二项 u 在边界面 S 上的值已知，用 $f(M)$ 代替，从而有

$$u(M_0)=\frac{1}{4\pi}\oint\!\!\!\!\!\!\iint_S(\frac{1}{r}\frac{\partial u}{\partial n}-f(M)\frac{\partial(1/r)}{n})dS, M_0\in D,$$

这样求解的关键是如何从上式中消去带 $\frac{\partial u}{\partial n}$（未知的）这一项。

由格林公式出发，要在区域 D 内求一个函数 g，它在区域 D 内调和（即 $\Delta g=0$），则格林公式为：

$$0=\oint\!\!\!\!\!\!\iint_S(u\frac{\partial g}{\partial n}-g\frac{\partial u}{\partial n})dS$$

用 $\frac{1}{4\pi}$ 乘以上式，再和积分表达式相加，就有

$$u(M_0)=\frac{1}{4\pi}\oint\!\!\!\!\!\!\iint_S[(\frac{1}{r}-g)\frac{\partial u}{\partial n}-f(M)\frac{\partial(1/r-g)}{n}]dSM_0\in D$$

如果上式中在边界面 S 上有 $\frac{1}{r}-g=0$，即 $g|_S=\frac{1}{r}$，那末狄里克雷问题的解就是：

$$u(M_0)=-\frac{1}{4\pi}\oint\!\!\!\!\!\!\iint_S[f(M)\frac{\partial(1/r-g)}{n}]dSM_0\in D.$$

7.3 偏微分方程求解的有限差分法

前面介绍了偏微分方程的基本理论以及求解偏微分方程解析解的几种常用方法。但在实际工程中，求解偏微分方程的解析解往往是非常困难的，这时候人们更希望能够得到能够实际计算的数值解，在本节中我们主要介绍偏微分方程数值求解方法中的一种最常用，也是最重要的一种方法：有限差分法。其基本思想都是将连续问题离散化，然后利用计算机进行数值求解。有限差分法首先将求解区域做网格剖分，用有限网格节点代替连续区域，然后从定解问题的微分或积分形式出发，用数值微分或数值积分公式导出相应的线性代数方程组，然后对有限形式的线性代数方程组进行求解，从而得到所求函数在网格节点上的值。

在本节中我们将针对三类标准的偏微分方程，分别讨论如下基本问题：(1) 如何对求解区域做网格剖分; (2) 怎样构造逼近微分方程定解问题的差分格式; (3) 差分格式解的收敛性和稳定性分析。

7.3.1 椭圆型方程的有限差分法

椭圆型方程中最简单的典型方程是在7.1.3节中描述的Poisson方程及Laplace方程的定解问题。本小节主要讨论下面的 Poisson 问题的有限差分法

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -f(x,y) \tag{7.45}$$

7.3.1.1 矩形网格的差分格式

（一）五点差分格式

设 Ω 为 xOy 面上一有界区域, $\partial\Omega$ 为其边界, 取沿 x 轴和 y 轴方向的步长分别为 h_1 和h_2, 分别做与坐标轴平行的直线族 $x_i = ih_1, y_i = jh_2$, $(i = 0, \pm 1, \cdots, j = 0, \pm 1, \cdots)$ 对平面区域 Ω 进行剖分，两族直线的交点 (ih_1, jh_2) 称为网格点或节点, 记为 (x_i, y_i)。 如图 7.3所示，其中图(a) 表示了区域分割示意图，图(b)给出剖分节点示意图。

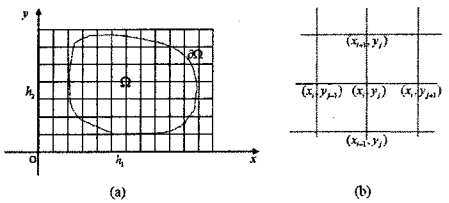


图 7.3: 区域离散剖分示意图

仅考虑于 $\Omega \cup \partial\Omega$ 的网格点, 如果两网格点(x_i, y_i) 和(x_i', y_i')满足不等式

$$\left|\frac{x_i - x_i'}{h_1}\right| + \left|\frac{y_j - y_j'}{h_2}\right| \leq 1$$

称这两个网格点是相邻的. 如果一个节点的四个相邻的点都属于 $\Omega \cup \partial\Omega$, 则称此节点为内邻节点, 或简称内点, 全部内点集合记作 Ω_h, 如果一个节点的四个相邻节点至少有一个不属于 $\Omega \cup \partial\Omega$ 则称其为边界节点, 或简称界点, 全体界点的集合记为 $\partial\Omega_h$.

设 (x_i, y_i) 为任意内点, 利用泰勒(Taylor)展开, 有

$$\frac{u(x_{i+1}, y_i) - 2u(x_i, y_i) + u(x_{i-1}, y_i)}{h_1^2} = \frac{\partial^2 u(x_i, y_i)}{\partial x^2} + O(h_1^2) \tag{7.46}$$

$$\frac{u(x_i, y_{i+1}) - 2u(x_i, y_i) + u(x_i, y_{i-1})}{h_2^2} = \frac{\partial^2 u(x_i, y_i)}{\partial y^2} + O(h_2^2) \tag{7.47}$$

舍掉无穷小, 并取 u_{ij} 为 $u(x_i,y_i)$ 的近似值, $f_{ij}=f(x_i,y_i)$, 则得

$$\frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{h_1^2}+\frac{u_{i,j+1}-2u_{ij}+u_{i,j-1}}{h_2^2}=-f_{ij} \tag{7.48}$$

由于差分方程 (7.48) 中只出现 u 在 (x_i,y_i) 及其四个邻点上的值（如图7.3(a) 所示）, 故称之为五点差分格式, 显见其截断误差为$O(h_1^2+h_2^2)$.

特别取正方形网格$h_1=h_2=h$, 则泊松方程差分格式成为

$$u_{ij}-\frac{1}{4}(u_{i-1,j}+u_{i,j-1}+u_{i+1,j}+u_{i,j+1})=\frac{h^2}{4}f_{ij} \tag{7.49}$$

拉普拉斯方程差分格式成为

$$u_{ij}=\frac{1}{4}(u_{i-1,j}+u_{i,j-1}+u_{i+1,j}+u_{i,j+1}) \tag{7.50}$$

除了上面的依据泰勒公式推导五点差分格式外，还可以利用积分插值法推导五点差分格式。

首先积分插值法需做对偶剖分，为此记 $x_{i-\frac{1}{2}}=(i-\frac{1}{2})h_1$, $y_{i-\frac{1}{2}}=(j-\frac{1}{2})h_2$, 在前面剖分上做两族平行于坐标轴的直线: $x=x_{i-\frac{1}{2}}$ 和$y=y_{j-\frac{1}{2}}, i,j,=0,\pm1,\cdots$。 考虑任一内点 (x_i,y_i), 则对偶剖分形成以 (x_i,y_i) 为中以的矩形区域, 如图 7.4 的阴影部分, 即以 A, B, C, D 表示矩形顶点, 坐标分别是 $A(x_{i-\frac{1}{2}},y_{j-\frac{1}{2}})$, $B(x_{i+\frac{1}{2}},y_{j-\frac{1}{2}})$, $C(x_{i+\frac{1}{2}},y_{j+\frac{1}{2}})$, $D(x_{i-\frac{1}{2}},y_{j+\frac{1}{2}})$, 以 D_{ij} 表示内部区域, 并在 D_{ij} 上对泊松方程 (7.45) 两边积分。

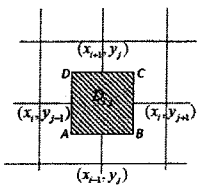


图 7.4: 积分区域示意图

利用格林(Green)公式, 可得

$$\int_{\widehat{ABCDA}}\frac{\partial u}{\partial n}\mathrm{d}s=-\iint\limits_{D_{ij}}f\mathrm{d}x\mathrm{d}y \tag{7.51}$$

式中 $\dfrac{\partial u}{\partial n}$ 表示 u 沿矩形 $\widehat{ABCDA}$ 的外法向导数, 用中矩形公式代替沿矩形四边形的线积分, 并用中心差商代替外法向导数, 则

$$\int_{\widehat{ABCDA}}\frac{\partial u}{\partial n}\mathrm{d}s\approx\frac{u_{i,j-1}-u_{ij}}{h_2}h_1+\frac{u_{i+1,j}-u_{ij}}{h_1}h_2+\frac{u_{i,j+1}-u_{ij}}{h_2}h_1+\frac{u_{i-1,j}-u_{ij}}{h_1}h_2$$

代入(7.51)式, 两边除以h_1h_2, 就得到形如 (7.49) 的五点差分格式

$$\frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{h_1^2}+\frac{u_{i,j+1}-2u_{ij}+u_{i,j-1}}{h_2^2}=-f_{ij}$$

这里

$$f_{ij}\approx\frac{1}{h_1h_2}\iint\limits_{D_{ij}}f\mathrm{d}x\mathrm{d}y$$

在椭圆形方程的有限差分法中，五点差分格式是最常用的。当然我们这里只给出了五个点的一种取法，也还有其他的五个点的取法，就可以得到不同的五点差分格式，有兴趣的读者自行查阅相关文献。

（二）边界条件的处理

五点差分格式需要用到中心点周围的四个点，这对于内点而言没有任何问题，但对于边界附近的节点，就会出现部分节点不属于原始区域的问题。因此，在对内点列出差分方程后, 为了方程的求解, 还必须对边界条件进行处理, 以给出边界上的关系式。

（1）对第一类边界条件

$$u\Big|_{\partial\Omega}=\alpha(x,y),\qquad (x,y)\in\partial\Omega$$

由于在构造网格时, $\partial\Omega_h$ 通常都不与 $\partial\Omega$ 相重合, 可能有少数边界节点落在 $\partial\Omega$ 上, 而大部分边界节点不落在 $\partial\Omega$ 上（如图 7.5所示，黑点为边界节点，曲线为 $\partial\Omega$ 的一段），由此而产生边界条件的转换:

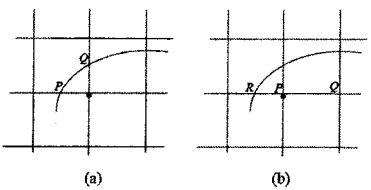


图 7.5: 第一类边界条件处理边界点示意图

(1) 直接转移法

若边界节点 (x_i,y_i) 正好落在 $\partial\Omega$上, 则$u_{ij}=\alpha(x_i,y_i)$; 如果 (x_i,y_i) 不在 $\partial\Omega$上, 此时 $\partial\Omega$ 与网格线交于 P 和 Q 两点（如图 7.5(a) 所示），则取与点 (x_i,y_i) 距离最近的点$\alpha(x,y)$ 的值为 u_{ij}, 采用这种替代方法所产生的误差阶为 $O(h)$。

(2) 线性插值

如图 7.5(b)所示, 对界点 P 可用 $\partial\Omega$ 上的点 R 与内点 Q 沿 x 方向做线性插值

$$u(P)=\frac{h}{h+\delta}u(R)+\frac{\delta}{h+\delta}u(Q)$$

其中 $\delta=RP<h$. 可以验证采用这种方法产生的误差阶为 $O(h^2)$。

（2）第二类、第三类边界条件

第二类和第三类边界条件都涉及到导数信息：

$$\left(\frac{\partial u}{\partial n}+ku\right)\Big|_{\partial\Omega}=\gamma(x,y)$$

的处理，显然 $k=0$, 即为第二类边界条件。

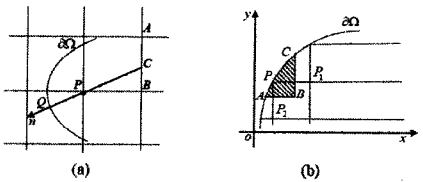


图 7.6: 第二、三类边界条件处理边界点示意图

设 P 为界点, 过 P 做 $\partial\Omega$ 的垂线交 $\partial\Omega$ 于 Q, 交内部网格线 AB 于 C(图7.6(a)所示), 令

$$AC=t_1h,\qquad CB=t_2h,\qquad PC=th$$

其中 h 为正方形网格的边长, $t_1+t_2=1$, $1\le t\le\sqrt{2}$, 则在点 P 处外法向导数 $\dfrac{\partial u}{\partial n}$ 近似为

$$\frac{u(P)-u(C)}{th}=\left(\frac{\partial u}{\partial n}\right)p+O(h)$$

$u(C)$可用节点A、B上的 u 值的线性插值表示

$$u(C)=t_1u(B)+t_2u(A)+O(h^2)$$

略去无穷小量, 则可得到点 P 处第三类边界条件的差分方程为

$$\frac{1}{th}\left[u(P)-t_1u(B)-t_2u(A)\right]+ku(P)=\gamma(Q)$$

若界点 P 在 $\partial\Omega$ 上恰为两族网格交点, 如图 7.6(b)所示, $P(x_i,y_i)$ 是界点, $P_1(x_i,y_{j-1})$ 和 $P_2(x_i,y_{j-1})$ 是相邻内点, 则用积分插值法处理第三类边界条件较为方便。

过点 $(x_{i+\frac{1}{2}}, y_j)$, $(x_i, y_{j-\frac{1}{2}})$ 分别作为 y 轴和 x 轴的平行线, 与 $\partial\Omega$ 截出的一曲边三角形 ABC. 于 ABC 积分 (7.45) 式两端, 并利用格林公式, 得

$$\int\limits_{ABCA} \frac{\partial u}{\partial n} \mathrm{d}s = -\iint\limits_{\triangle ABC} f \mathrm{d}x\mathrm{d}y \tag{7.52}$$

由于

$$\int_{\overline{AB}} \frac{\partial u}{\partial n} \mathrm{d}s = \frac{u(P_2) - u(P)}{h_2} \cdot \overline{AB}$$

$$\int_{\overline{BC}} \frac{\partial u}{\partial n} \mathrm{d}s = \frac{u(P_1) - u(P)}{h_1} \cdot \overline{BC}$$

$$\int_{\overline{CA}} \frac{\partial u}{\partial n} \mathrm{d}s = \int_{\overline{CA}} (\gamma - ku) \mathrm{d}s \approx [\gamma(P) - ku(P)]\overline{CA}$$

这里 h_1, h_2 分别是沿 x 轴, y 轴方向的步长。将上面三式代入 (7.52) 式, 即得第三类边界条件的差分方程

$$\frac{u(P_2) - u(P)}{h_2}\overline{AB} + \frac{u(P_1) - u(P)}{h_1}\overline{BC} + [\gamma(P) - ku(P)]\overline{CA} = -\iint\limits_{\triangle ABC} f \mathrm{d}x\mathrm{d}y$$

7.3.1.2 差分方程解的收敛性及差分方程组的解法

仅以泊松方程的第一类边界条件的五点差分格式进行讨论, 设

$$\begin{cases} -\Delta_h u_{ij} = f_{ij}, \ (x_i, y_i) \in \Omega_h \\ u_{ij} = \alpha_{ij}, \ (x_i, y_i) \in \partial\Omega_h \end{cases} \tag{7.53}$$

其中Δ_h称为差分算子

$$\Delta_h u_{ij} \triangleq \frac{u_{i+1,j} - 2u_{ij} + u_{i,j+1}}{h_1^2} + \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{h_2^2} \tag{7.54}$$

差分格式解的收敛性, 是指当步长 $h_1, h_2 \to 0$ 时, 差分方程的解 u_{ij} 是否逼近微分方程的解 $u(x_i, y_i)$。证明解的收敛性要借助于差分算子Δ_h的极值原理。

定理 7.2 (极值原理) 设 u_{ij} 是定义在 $\Omega_h \cup \partial\Omega_h$ 上的函数, 则有

(1) 如果 $\Delta_h u_{ij} \geq 0, (x_i, y_i) \in \Omega_h$, 则

$$\max_{\Omega_h} u_{ij} \leq \max_{\partial\Omega_h} u_{ij}$$

即 Ω_h 内的所取 u 值不会大于 u 在边界 $\partial\Omega_h$ 上的最大值, 即 u 在 $\Omega_h \cup \partial\Omega_h$ 上的最大值必定在边界 $\partial\Omega_h$ 上取得。

(2) 如果 $\Delta_h u_{ij} \leq 0, (x_i, y_i) \in \Omega_h$, 则

$$\min_{\Omega_h} u_{ij} \geq \min_{\partial\Omega_h} u_{ij}$$

即 Ω_h 内的所取 u 值不会大于 u 在边界 $\partial\Omega_h$ 上的最小值，即 u 在 $\Omega_h \cup \partial\Omega_h$ 上的最小值必定在边界 $\partial\Omega_h$ 上取得。

证明：(反证法)假设(1)的结论不成立，即在 Ω_h 内一点 (x_{i_0}, y_{j_0}) 能够取得 $\Omega_h \cup \partial\Omega_h$ 上的最大值 M，即 $u(x_{i_0}, y_{j_0}) = M$，其中

$$M \geq u_{ij}, (x_i, y_i) \in \Omega_h, \text{ 且 } M > u_{ij}, (x_i, y_i) \in \partial\Omega_h$$

根据差分算子定义，在 (x_{i_0}, y_{j_0}) 处，有

$$\Delta_h u_{i_0 j_0} = \frac{u_{i_0+1,j_0} - 2u_{i_0,j_0} + u_{i_0-1,j_0}}{h_1^2} + \frac{u_{i_0,j_0+1} - 2u_{i_0,j_0} + u_{i_0,j_0-1}}{h_2^2} \geq 0$$

从而得到

$$M = u_{i_0 j_0} \leq \frac{1}{\frac{1}{h_1^2} + \frac{1}{h_2^2}} \left[\frac{u_{i_0+1,j_0} + u_{i_0-1,j_0}}{2h_1^2} + \frac{u_{i_0,j_0+1} + u_{i_0,j_0-1}}{2h_2^2} \right]$$

注意到$M \geq u_{ij}, (x_i, y_i) \in \Omega_h$，因此得到

$$u_{i_0+1,j_0} = u_{i_0-1,j_0} = u_{i_0,j_0+1} = u_{i_0,j_0-1} = M$$

把已得四点重复上述证明可得同样的结论(即每个点周围四个点的值都等于 M)，从而得到

$$u_{ij} = M, \qquad (x_i, y_i) \in \Omega_h \cup \partial\Omega_h$$

但这与 $M > u_{ij}, (x_i, y_i) \in \partial\Omega_h$矛盾。

对于(2)的证明，注意到有 $\max(-u_{ij}) = -\min u_{ij}$ 及 $\Delta_h(-u_{ij}) = -\Delta_h u_{ij}$，因此只需要令 $u_{i,j} = -v_{i,j}$，即可按照上述步骤证明（2）成立。

利用极值原理，我们就可以考察差分方程解的收敛性。事实上，由极值原理，对定义在 $\Omega_h \cup \partial\Omega_h$上的函数$u_{ij}$，有

$$\max_{\Omega_h} |u_{ij}| \leq \max_{\partial\Omega_h} |u_{ij}| + \frac{h^2}{2} \max_{\Omega_h} |\Delta_h u_{ij}|$$

其中 h 为矩形区域 Ω 的 x 方向的步长，从而得到估计式

$$\max_{\Omega_h} |u_{ij} - u(x_i, y_i)| \leq O(h_1^2 + h_2^2)$$

因此，当 $h_1, h_2 \to 0$，差分方程的解收敛于微分方程的解.

由前面的讨论可知，椭圆型方程边值问题经离散化导出的差分方程是一个线性代数方程组，这个方程组有两个明显特点:

①相应的系数矩阵有较高的阶数;

②系数矩阵是带状矩阵, 即在整个矩阵中, 零元素占绝大多数, 非零元素很少, 且非零元素位于主对角线两侧, 呈带状分布.

对这类方程组, 通常采用迭代法求解, 常用的迭代法有 Jacobi 方法, Gauss-Seidel方法, 逐次超松驰方法等等。

例 7.7 *求解方程*

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2}+\dfrac{\partial^2 u}{\partial y^2}=0 \\ u(x,0)=u(0,y)=0 \\ u(x,0.5)=200x \\ u(0.5,y)=200y \end{cases}$$

解： 采用五点差分格式

$$4u_{ij}-u_{i+1,j}-u_{i-1,j}-u_{i,j+1}-u_{i,j-1}=0, \qquad i,j=1,2,3$$

取正方形网格部分区域$\Omega, h=0.125$, 如图7.7所示 用网格点表示上述量, 则方程可写成

$$\begin{cases} 4u_1-u_2-u_4=u_{0,3}+u_{1,4} \\ 4u_2-u_3-u_1-u_5=u_{2,4} \\ 4u_3-u_2-u_6=u_{4,3}+u_{3,4} \\ 4u_4-u_5-u_1-u_7=u_{0,2} \\ 4u_5-u_6-u_4-u_2-u_8=0 \\ 4u_6-u_5-u_3-u_9=u_{4,2} \\ 4u_7-u_8-u_4=u_{0,1}+u_{1,0} \\ 4u_8-u_9-u_7-u_5=u_{2,0} \\ 4u_9-u_8-u_6=u_{3,0}+u_{4,1} \end{cases}$$

其中右端项可以按照边界条件的直接转移法得到

$$\begin{cases} u_{1,0}=u_{2,0}=u_{3,0}=u_{0,1}=u_{0,2}=u_{0,3}=0 \\ u_{1,4}=u_{4,1}=25, \\ u_{2,4}=u_{4,2}=50, \\ u_{3,4}=u_{4,3}=75. \end{cases}$$

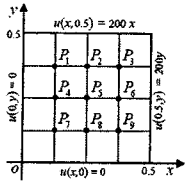


图 7.7: 正方形网格区域分割

从而得线性方程组

$$\begin{pmatrix} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \end{pmatrix} = \begin{pmatrix} 25 \\ 50 \\ 150 \\ 0 \\ 0 \\ 50 \\ 0 \\ 0 \\ 25 \end{pmatrix}$$

用高斯-赛德尔方法求解上述线性方程组, 其结果如下:

i	1	2	3	4	5	6	7	8	9
u_i	18.75	37.50	56.25	12.50	25.00	37.50	6.25	12.50	18.75

事实上，本例的解析为 $u(x,y) = 400xy$, 可以验证上表计算得到的值和在节点处的实际值是比较吻合的，这也从数值实验上验证了该差分格式是收敛的。

7.3.2 抛物型方程的有限差分法

抛物型方程中最简单的典型方程就是扩散方程，本节以扩散方程的定解问题为例，来介绍抛物型方程的有限差分法。

在平面区域 $\Omega = \{(x,t)|0 \le x \le l, 0 \le t \le T\}$ 上考虑下面的扩散方程的定解问题

$$\begin{cases} Lu = \dfrac{\partial u}{\partial t} - a\dfrac{\partial^2 u}{\partial x^2} = 0, \ (x,t) \in \Omega \\ u(x,0) = \varphi(x), \qquad 0 \le x \le l \\ u(0,t) = u(l,t) = 0, \quad 0 \end{cases} \tag{7.55}$$

用平行直线族 $x_i = ih, t_k = k\tau$, $(i = 0, 1, \cdots, N,\ k = 0, 1, \cdots, M)$ 对区域 Ω 进行矩形剖分, h, t 分别称空间步长和时间步长, 对于固定的 $t = t_k$ 上的全体节点 (x_i, t_k) 称为网格的第 k 层, 并记全体内部节点为 Ω_h。

7.3.2.1 扩散方程的差分格式

(一)古典显格式与古典隐格式

首先，考虑函数 $u(x,t)$ 在节点 $(x_i, t_k) \in \Omega_h$ 处分别在时间和空间维度上的 Taylor 展开

$$u(x, t_k) = u(x_i, t_k) + \left[\frac{\partial u}{\partial x}\right]_i^k (x - x_i) + \left[\frac{\partial^2 u}{\partial x^2}\right]_i^k (x - x_i)^2 + O((x - x_i)^3), \tag{7.56}$$

$$u(x_i, t) = u(x_i, t_k) + \left[\frac{\partial u}{\partial t}\right]_i^k (t - t_k) + O((t - t_k)^2) \tag{7.57}$$

式中$[\ ]_i^k$表示函数在节点(x_i, t_k)处的值。

由此，可得到函数 $u(x,t)$ 在 $(x_{i+1}, t_k), (x_{i-1}, t_k), (x_i, t_{k+1})$ 三个点上的值分别为:

$$\begin{cases} u(x_{i-1}, t_k) = u(x_i, t_k) + \left[\frac{\partial u}{\partial x}\right]_i^k (-h) + \left[\frac{\partial^2 u}{\partial x^2}\right]_i^k h^2 + O(h^3) \\ u(x_{i+1}, t_k) = u(x_i, t_k) + \left[\frac{\partial u}{\partial x}\right]_i^k h + \left[\frac{\partial^2 u}{\partial x^2}\right]_i^k h^2 + O(h^3) \\ u(x_i, t_{k+1}) = u(x_i, t_k) + \left[\frac{\partial u}{\partial t}\right]_i^k \tau + O(\tau^2) \end{cases} \tag{7.58}$$

从而可得

$$\begin{cases} \left[\frac{\partial^2 u}{\partial x^2}\right]_i^k = \frac{1}{h^2}\left[u(x_{i-1}, t_k) - 2u(x_i, t_k) + u(x_{i+1}, t_k)\right] + O(h^2) \\ \left[\frac{\partial u}{\partial t}\right]_i^k = \frac{1}{\tau}\left[u(x_i, t_{k+1}) - u(x_i, t_k)\right] + O(\tau) \end{cases} \tag{7.59}$$

将 (7.59) 带入到 (7.55) 中的扩散方程中，可得

$$\begin{aligned} 0 = L[u]_i^k &= \left[\tfrac{\partial u}{\partial t}\right]_i^k - a\left[\tfrac{\partial^2 u}{\partial x^2}\right]_i^k \\ &= \frac{1}{\tau}\left[u(x_i, t_{k+1}) - u(x_i, t_k)\right] - \frac{a}{h^2}\left[u(x_{i-1}, t_k) - 2u(x_i, t_k) + u(x_{i+1}, t_k)\right] \\ &\quad + O(\tau + h^2) \end{aligned}$$

舍去上式中的局部截断误差 $O(\tau + h^2)$, 并以 u_i^k 表示 $u(x_i, t_k)$ 就可得到差分方程

$$L_h[u]_i^k = \frac{1}{\tau}(u_i^{k+1} - u_i^k) - \frac{a}{h^2}(u_{i+1}^k - 2u_i^k + u_{i-1}^k) = 0 \tag{7.60}$$

这里 $L_h[u]_i^k$ 是 Lu 在点 (x_i, t_k) 的处的有限差分。

引进记号 $r=\frac{\tau}{h^2}$, 整理 (7.60) 可得

$$u_i^{k+1}=u_i^k+ra(u_{i+1}^k-2u_i^k+u_{i-1}^k) \tag{7.61}$$

这里 r 称网格比, 差分方程(7.61)的局部截断误差为$O(\tau+h^2)$。

由差分方程 (7.61)可以看出, 计算第 $k+1$ 层任一节点处的值 u_i^{k+1}, 可由第 k 层的三个相邻节点处的值$u_{i+1}^k,u_i^k,u_{i-1}^k$ 得到（如图7.8 所示），其中第 0 层和边界节点上的值可由 (7.55) 式的初始条件和边界条件得到

$$\begin{cases} u_i^0=\varphi(x_i), & i=1,2,\cdots,N-1, \\ u_i^k=u_N^k=0, & k=1,2,\cdots,M-1. \end{cases} \tag{7.62}$$

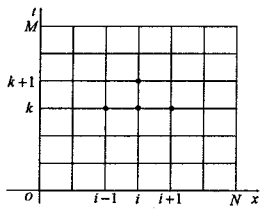


图 7.8: 古典显格式节点分布图

从初边值条件 (7.62) 出发，就可按格式 (7.61) 沿着 t 的方向逐点逐层计算得到所有的 u_i^k。由于每次计算可以通过第 k 层的值直接计算得到第 $k+1$ 的值,因此差分方程 (7.61) 通常被称为古曲显格式。

上述古典差分格式还可以写成如下的矩阵形式

$$\boldsymbol{u}^{k+1}=A\boldsymbol{u}^k, \qquad k=1,2,\cdots,M \tag{7.63}$$

其中

$$A=\begin{pmatrix} 1-2ra & ra & & & \\ ra & 1-2ra & ra & & \\ & \ddots & \ddots & \ddots & \\ & & ra & 1-2ra & ra \\ & & & ra & 1-2ra \end{pmatrix}, \boldsymbol{u}^k=\begin{pmatrix} u_1^k \\ u_2^k \\ \vdots \\ u_{N-1}^k \end{pmatrix}$$

从公式 (7.63) 可以看到，古典显格式的计算只涉及到矩阵向量的乘法，计算复杂度很低。

除了古典显格式外，古典隐格式也是常用的差分格式。古典隐格式的节点分布如图 7.9 所示

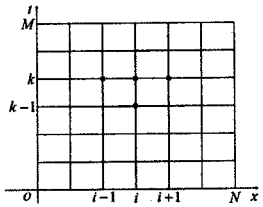


图 7.9: 古典隐格式节点分布图

和古典显格式的推导过程类似，利用Taylor公式 (7.56) 和 (7.57) 可得 函数 $u(x,t)$ 在图 7.9所示的 $(x_{i+1},t_k),(x_{i-1},t_k),(x_i,t_{k-1})$ 三个点上的值分别为：

$$\begin{cases} u(x_{i-1},t_k)=u(x_i,t_k)+\left[\frac{\partial u}{\partial x}\right]_i^k(-h)+\left[\frac{\partial^2 u}{\partial x^2}\right]_i^k h^2+O(h^3) \\ u(x_{i+1},t_k)=u(x_i,t_k)+\left[\frac{\partial u}{\partial x}\right]_i^k h+\left[\frac{\partial^2 u}{\partial x^2}\right]_i^k h^2+O(h^3) \\ u(x_i,t_{k-1})=u(x_i,t_k)+\left[\frac{\partial u}{\partial t}\right]_i^k(-\tau)+O(\tau^2) \end{cases} \tag{7.64}$$

从而得到

$$\begin{aligned} 0=L[u]_i^k&=\left[\frac{\partial u}{\partial t}\right]_i^k-a\left[\frac{\partial^2 u}{\partial x^2}\right]_i^k \\ &=\frac{1}{\tau}[u(x_i,t_k)-u(x_i,t_{k-1})]-\frac{a}{h^2}[u(x_{i-1},t_k)-2u(x_i,t_k)+u(x_{i+1},t_k)] \\ &\quad +O(\tau+h^2) \end{aligned}$$

同样，舍去上式中的局部截断误差 $O(\tau+h^2)$, 并以 u_i^k 表示 $u(x_i,t_k)$ 就可得到差分方程

$$u_i^k-r(u_{i+1}^k-2u_i^k+2u_{i-1}^k)=u_i^{k-1} \tag{7.65}$$

其中 r 为网格比，该格式的局部截断误差为 $O(\tau+h^2)$。 考虑到初始条件与边界条件, 就有

$$\begin{cases} -rau_{i-1}^k+(1+2ra)u_i^k-rau_{i+1}^k=u_i^{k-1}, & i=1,2,\cdots,N-1 \\ u_i^0=\varphi(x_i), & i=1,2,\cdots,N-1 \\ u_i^k=u_N^k=0, & j=1,2,\cdots,M \end{cases} \tag{7.66}$$

写成矩阵形式为

$$A\boldsymbol{u}^k=\boldsymbol{u}^{k-1} \tag{7.67}$$

其中

$$A=\begin{pmatrix} 1+2ra & -ra & & & \\ -ra & 1+2ra & -ra & & \\ & \ddots & \ddots & \ddots & \\ & & -ra & 1+2ra & -ra \\ & & & -ra & 1+2ra \end{pmatrix}$$

从 (7.66)式可以看到，已知第 $k-1$ 层的值之后，必须要通过求解线性方程组才能得到第 k 层的值，不能像显格式那样直接解出，这种格式被称为古典隐格式。因矩阵 A 是对角占优矩阵, 一般用追赶法求解，虽然计算也比较方便，但显然计算复杂度要高于古典显格式。

（二）理查森(Richardson)格式

无论是古典显格式还是古典隐格式，局部截断误差均为 $O(\tau+h^2)$。局部阶段误差反映了差分格式与原连续问题之间的误差，因此人们往往希望能够为了提高截断误差的阶, 得到高精度的差分格式。理查森(Richardson)格式就是基于这种想法提出来的，该格式采用的节点分布如图 7.10 所示。

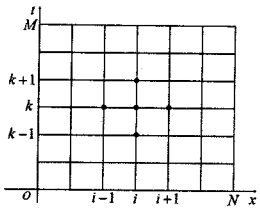


图 7.10: 理查森格式节点分布图

理查森格式在空间维度上的处理与古典隐格式相同，但在时间方向上，理查森格式不仅用到第 $k-1$ 层的值，而且还用到了第 $k+1$ 层的值。按照Taylor 公式 (7.57) 可得

$$\begin{cases} u(x_i,t_{k-1}) = u(x_i,t_k) + \left[\frac{\partial u}{\partial t}\right]_i^k(-\tau) + O(\tau^2) \\ u(x_i,t_{k+1}) = u(x_i,t_k) + \left[\frac{\partial u}{\partial t}\right]_i^k\tau + O(\tau^2) \end{cases}$$

将上面的两式相减可以得到

$$\left[\frac{\partial u}{\partial t}\right]_i^k = \frac{1}{2\tau}\left[u(x_i,t_{k+1}) - u(x_i,t_{k-1})\right] + O(\tau^2)$$

从而得到

$$\begin{aligned} 0 = L[u]_i^k &= \left[\tfrac{\partial u}{\partial t}\right]_i^k - a\left[\tfrac{\partial^2 u}{\partial x^2}\right]_i^k \\ &= \frac{1}{2\tau}\left[u(x_i,t_{k+1}) - u(x_i,t_{k-1})\right] - \frac{a}{h^2}\left[u(x_{i-1},t_k) - 2u(x_i,t_k) + u(x_{i+1},t_k)\right] \\ &\quad + O(\tau^2+h^2) \end{aligned}$$

同样，舍去上式中的局部截断误差 $O(\tau^2+h^2)$, u_i^k 表示 $u(x_i,t_k)$ 就可得到理查森差分方程

$$u_i^{k+1} = u_i^{k-1} - 2ra(u_{i+1}^k - 2u_i^k + u_{i-1}^k), \qquad i = 1,2,\cdots,N-1 \tag{7.68}$$

其中 r为网格比，该格式的局部截断误差为$O(\tau^2+h^2)$。从这里可以看到，理查森差分格式的局部截断误差阶高于古典显格式和古典隐格式。

理查森差分方程的矩阵形式为

$$\boldsymbol{u}^{k+1}=\boldsymbol{u}^{k-1}+B\boldsymbol{u}^{k} \tag{7.69}$$

其中

$$B=\begin{pmatrix} -4ra & 2ra & & & \\ 2ra & 1-4ra & 2ra & & \\ & \ddots & \ddots & \ddots & \\ & & 2ra & 1-4ra & 2ra \\ & & & 2ra & -4ra \end{pmatrix}$$

从 (7.68) 或 (7.69) 可以看到，理查森格式也是一种显格式，在计算第 $k+1$ 层的节点值时，需要用到第 k 层和第 $k-1$ 层上的节点值。实际计算时，除初始条件(第 0 层)及边界条件外，要事先用其他格式求得第一层上的值$\boldsymbol{u}^1$，方能按(7.68) 或 (7.69) 式计算出所有层的节点值。

（三）加权六点格式

为了得到稳定性好，且有较高精度的差分格式，我们还可以把古典显格式和古典隐格式进行线性组合，得如下差分格式:

$$\frac{u_i^k-u_i^{k-1}}{\tau}-\frac{a}{h^2}[\theta(u_i^{k+1}-2u_i^k=u_{i-1}^k)+(1-\theta)(u_{i+1}^{k-1}-2u_i^{k-1}+u_{i-1}^{k-1})]=0 \tag{7.70}$$

其中 $0\ le\theta\le 1$，这一差分方程用到相邻两层六个节点的函数值，如图 7.11 所示，因此该格式也被称为加权六点格式。特别地，当 $\theta=0$ 时，该格式退化为古典显格式；当 $\theta=0$ 时，该格式退化为古典隐格式。

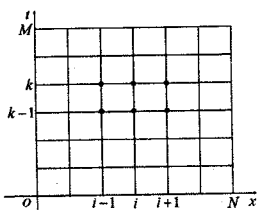


图 7.11: 加权六点格式节点分布图

同样利用Taylor展开，当方程的解 $u(x,t)$ 充分光滑时，可以得到该格式在 (x_i,t_k) 处的局部阶段误差为

$$R=a\tau\left(\frac{1}{2}-\theta\right)\left[\frac{\partial^3 u}{\partial x^2\partial t}\right]_i^k+O(\tau^2+h^2)$$

可以看出, $\theta \neq \frac{1}{2}$时, 截断误差为$O(\tau + h^2)$; $\theta = \frac{1}{2}$时, 截断误差是$O(\tau^2 + h^2)$, 此差分格式常被使用, 被称为克兰克-尼科尔森(Crank-Nicolson)格式, 其具体形式为

$$u_i^k - u_i^{k-1} - \frac{1}{2}ra\left[(u_{i+1}^k - 2u_i^k + u_{i-1}^k) + (u_{i+1}^{k-1} - 2u_i^{k-1} + u_{i-1}^{k-1})\right] = 0 \tag{7.71}$$

其中 r 为网格比。

克兰克-尼科尔森可用矩阵表示为

$$A_1 \boldsymbol{u}^k = A_2 \boldsymbol{u}^{k-1} \tag{7.72}$$

其中

$$A_1 = \begin{pmatrix} 1+ra & -\frac{ra}{2} & & & \\ -\frac{ra}{2} & 1+r & -\frac{ra}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{ra}{2} & 1+r & -\frac{ra}{2} \\ & & & -\frac{ra}{2} & 1+r \end{pmatrix}$$

$$A_2 = \begin{pmatrix} 1-r & \frac{ra}{2} & & & \\ \frac{ra}{2} & 1-r & \frac{ra}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{ra}{2} & 1-r & \frac{ra}{2} \\ & & & \frac{ra}{2} & 1-ra \end{pmatrix}$$

可以看到，克兰克-尼科尔森是一个隐格式, 但 $\boldsymbol{u}^k$ 的系数矩阵 A_1 是对角占优的三对角阵, 结合定解条件, 可以用追赶法来快速求解方程组。

例 7.8 考虑定解问题

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2}, & 0 < x < 1, \quad t > 0 \\ u(x,0) = \sin \pi x, & 0 \le x \le 1 \\ u(0,t) = u(1,t) = 0, & t > 0 \end{cases}$$

取$h = 0.1$, $\tau = 0.01$, 网格比$r = \dfrac{\tau}{h^2} = 1$

① 用古典显格式;

② 古典隐格式;

③ 克兰克-尼科尔森格式求解.

并比较其结果(其精确解为$u(x,t) = e^{-\pi^2 t}\sin \pi x$)。

解：① 古典显格式

$$\begin{cases} \dfrac{u_i^{k+1}-u_i^k}{\tau}=\dfrac{1}{h^2}\left(u_{i+1}^k-2u_i^k+u_{i-1}^k\right), & i=1,2,\cdots,9 \\ u_i^0=\sin\pi x_i, & ,i=0,1,\cdots,10 \\ u_0^k=u_{10}^k=0 \end{cases}$$

由于$r=\dfrac{\tau}{h^2}=1$, 所以差分格式为

$$u_i^{k+1}=u_{i-1}^k-u_i^k+u_{i+1}^k, \qquad i=1,2,\cdots,9$$

矩阵形式

$$A\boldsymbol{u}^k=\boldsymbol{u}^{k+1}$$

其中

$$A=\begin{pmatrix} -1 & 1 & & & \\ 1 & -1 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -1 & 1 \\ & & & 1 & -1 \end{pmatrix}, \qquad \boldsymbol{u}^k=\begin{pmatrix} u_1^k \\ u_2^k \\ \ddots \\ u_8^k \\ u_9^k \end{pmatrix}$$

由初始条件开始, 直接计算可得到$t=0.5$的一组解, 见表7.1.

② 古典隐格式

$$\frac{u_i^{k+1}-u_i^k}{\tau}=\frac{1}{h^2}\left(u_{i+1}^{k+1}-2u_i^{k+1}+u_{i-1}^{k+1}\right), \qquad i=1,2,\cdots,9$$

因为$r=\dfrac{\tau}{h^2}=1$, 所以上述方程简为

$$-u_{i+1}^{k+1}+3u_i^{k+1}-u_{i-1}^{k+1}=u_i^k, \qquad i=1,2,\cdots,9$$

矩阵形式的系数矩阵A为对角占优的三对角阵

$$A=\begin{pmatrix} 3 & -1 & & & \\ -1 & 3 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 3 & 1 \\ & & & -1 & 3 \end{pmatrix}$$

用追赶法求解, 重复计算, 可得$t=0.5$的一组解, 见表7.1.

③ 克兰克-尼科尔森格式

$$\frac{u_i^{k+1}-u_i^k}{\tau}=\frac{1}{2h^2}[(u_{i+1}^k-2u_i^k-u_{i-1}^k)+(u_{i+1}^{k+1}-2u_i^{k+1}+u_{i-1}^{k+1})], \qquad i=1,2,\cdots,9$$

由于$r=\frac{\tau}{h^2}=1$, 所以上述差分格式写成

$$-u_{i+1}^{k+1}+4u_i^{k+1}-u_{i-1}^{k+1}=u_{i+1}^k+u_{i-1}^k, i=1,2,\cdots,9$$

此线性代数方程组的系数矩阵仍是对角占优的三角矩阵, 用追赶法重复计算, 就可得到$t=0.5$的一组解, 见表7.1.

表 7.1: $t=0.5$的一组解u_i^{50}

x_i	精确解$u(x_i,0.5)$	显格式	隐格式	C-N格式
0	0	0	0	0
0.1	0.00222241	8.19876×10^{-7}	0.00289802	0.00230512
0.2	0.00422728	-1.55719×10^{-8}	0.00551236	0.00438461
0.3	0.00581836	2.13833×10^{-8}	0.00758711	0.00603489
0.4	0.00683989	-2.50642×10^{-8}	0.00891918	0.00709440
0.5	0.00719188	2.62685×10^{-8}	0.00937818	0.00745954
0.6	0.0683989	-2.49015×10^{-8}	0.00891918	0.00709440
0.7	0.00581836	2.11200×10^{-8}	0.00758711	0.00603489
0.8	0.00422728	-1.53086×10^{-8}	0.00551236	0.00438461
0.9	0.00222241	8.03604×10^{-7}	0.00289802	0.00230512
1	0	0	0	0

从上面的计算结果可以看出, 在网格比相同情况下, 用隐格式计算结果比显格式精确。事实上，在这个例子中，显格式得到的解显然和真实解有很大的差距，如果将网格比缩小为 $r=0.05$, 则显格式也可得到与克兰克-尼科尔森格式相近的结果。为什么会产生这样的情况呢？这就是下面要研究的差分格式的稳定性和收敛性问题。

7.3.2.2 差分格式的稳定性和收敛性

（一）差分格式的稳定性概念

对一个有限差分格式, 当初始数据有微小误差时, 按时间上逐层计算, 是否保留差分方程的解也是微小变化, 即计算过程能否控制误差的增长或传播, 这就是差分格式的稳定性问题.

例 7.9 *考察古典显格式对于舍入误差的传播情况。*

解: 这里采用的方法被称 ε-图法, 就是在某层任意一个节点处给出一个误差 ε, 然后通过图表观察这一误差的发展情形。

设 u_i^{k+1} 是古典显格式

$$u_i^{k+1} = u_i^k + ra(u_{i+1}^k - 2u_i^k + u_{i-1}^k) \tag{7.73}$$

的精确解, $\bar{u}_i^{k+1}$ 是 (7.73) 式近似解, 用 e_i^{k+1} 表示计算 u_i^{k+1} 时产生的误差, 即$e_i^{k+1} = \bar{u}_i^{k+1} - u_i^{k+1}$。假定在第 k 层的点 (i_0, k) 处产生误差 $e_{i_0}^k = \varepsilon$, 而这一层其他各点的无误差, 且以后计算中也不引入新的误差, 显然由于 u_i^k 的 $\bar{u}_i^k$ 都满足差分格式 (7.73), 则 e_i^{k+1} 也满足下面的误差传播方程

$$e_i^{k+1} = e_i^k + ra(e_{i+1}^k - 2e_i^k + e_{i-1}^k) \tag{7.74}$$

现分析$ra = \frac{1}{2}$ 和 $ra = 1$时, 误差 ε 随 k 增加而变化的情形。

当 $ra = \frac{1}{2}$ 时, e_i^{k+1} 满足的误差传播方程为

$$e_i^{k+1} = \frac{1}{2}(e_{i+1}^k + e_{i-1}^k) \tag{7.75}$$

利用此公式, 易计算误差传播情况, 如表7.2 所列。由表 7.2 可知, 初始数据的误差, 在以后层次逐渐减小, 说明$ra = \frac{1}{2}$ 时, 差分格式是稳定的。

表 7.2: $ra = \frac{1}{2}$时误差的传播情况

k \ i	i_0-4	i_0-3	i_0-2	i_0-1	i_0	i_0+1	i_0+2	i_0+3	i_0+4
k					ε				
$k+1$				$\frac{\varepsilon}{2}$	0	$\frac{\varepsilon}{2}$			
$k+2$			$\frac{\varepsilon}{4}$	0	$\frac{\varepsilon}{2}$	0	$\frac{\varepsilon}{4}$		
$k+3$		$\frac{\varepsilon}{8}$	0	$\frac{3\varepsilon}{8}$	0	$\frac{3\varepsilon}{8}$	0	$\frac{\varepsilon}{8}$	
$k+4$	$\frac{\varepsilon}{16}$	0	$\frac{4\varepsilon}{16}$	0	$\frac{6\varepsilon}{16}$	0	$\frac{4\varepsilon}{16}$	0	$\frac{\varepsilon}{16}$

当 $ra = 1$ 时, e_i^{k+1} 满足的误差传播方程为

$$e_i^{k+1} = e_{i+1}^k - e_i^k + e_{i-1}^k \tag{7.76}$$

计算结果如表7.3所示.

此时初始数据的误差, 随 k 的增大而迅速增大, 因此 $ra = 1$ 时古典显格式是不稳定的。

由这个例子可以看出, 一个差分格式是否稳定与网格比的大小有。差分格式的稳定性在差分方法的研究中, 具有特别重要的意义, 需进一步讨论, 并找出判定方法。为此, 这里首先给出稳定的严格定义:

表 7.3: $ra=1$时误差的传播情况

k \ i	i_0-4	i_0-3	i_0-2	i_0-1	i_0	i_0+1	i_0+2	i_0+3	i_0+4
k					ε				
$k+1$				ε	$-\varepsilon$	ε			
$k+2$			ε	-2ε	3ε	-2ε	ε		
$k+3$		ε	-3ε	6ε	-7ε	6ε	-3ε	ε	
$k+4$	ε	-4ε	10ε	-16ε	19ε	-16ε	10ε	-4ε	ε

定义 7.3 对于任意给定的 $\varepsilon>0$, 如果总存在与h,τ 无关且只赖于 ε 的正数 δ, 使当 $||\bar{\boldsymbol{u}}^0-\boldsymbol{u}^0||<\delta$ 时, 不等式 $||\bar{\boldsymbol{u}}^n-\boldsymbol{u}^n||<\varepsilon$ 对任何对 $n(0\le n\tau\le T)$ 都成立, 则称差分格式是稳定的, 其中范数 $||\boldsymbol{u}^n||=\{\max_i(u_i^n)^2h\}^{\frac{1}{2}}$。

简单地说，所谓差分格式稳定，就是指误差的传播是衰减的或至少是“可控”的。

对逼近初边值问题 (7.55) 的两层差分格式, 可统一写成矩阵形式

$$\boldsymbol{u}^{k+1}=H\boldsymbol{u}^k \tag{7.77}$$

则误差向量 $E^k=\bar{\boldsymbol{u}}^k-\boldsymbol{u}^k$ 所满足的误差传播方程为

$$E^{k+1}=HE^k$$

通过逐次迭代, 可推出

$$E^{k+1}=H^{k+1}E^0$$

在上式两边同时取范数，可得

$$||E^{k+1}||=||H^{k+1}E^0||\le||H^{k+1}||\cdot||E^0||\le||H||^{k+1}\cdot||E^0||$$

由此可见, 差分格式稳定的充要条件是, $||H||^n\le K$ 对任何 n 成立。于是得到差分格式稳定性的等价定义:

定义 7.4 差分格式(7.76)称对初值是稳定的, 如果存在常数K及t_0, 使得$)<\tau<\tau_0$时, 一致地有

$$||\boldsymbol{u}^n||\le K||\boldsymbol{u}^0|| \tag{7.78}$$

通过对 H 的直接估计就可以探求差分格式的稳定性条件, 常用的方法有矩阵方法,即只要矩阵 H 的谱半径小于1即可，但在实际问题中矩阵 H 的特征值很难求，其谱半径也很难估计。因此，在本节中我们并不介绍判别差分格式稳定性的矩阵方法，而主要介绍比较简便实用的傅里叶(Fourier) 方法。

（二）判别稳定性的傅里叶方法

傅里叶方法的适用范围是线性常系数的定解问题，下面从具体例子入手，介绍该方法的基本思想。

为便于讨论，不失一般性，可假设函数 $u(x,t)$ 空间变化范围为 $[0,1]$ 区间，如果不是 $[0,1]$ 区间，也可以通过线性变化标准化为 $[0,1]$ 区间。

对于古典显格式

$$u_i^{k+1} = u_i^k + ra(u_{i+1}^k - 2u_i^k + u_{i-1}^k) \tag{7.79}$$

注意到，无论对于齐次方程还是非齐次方程，我们在稳定性分析时只需要讨论误差传播方程，而古典显格式的误差传播方程就是其次差分方程 (7.79)，因此我们在稳定性讨论的时候，只需要讨论该齐次差分方程即可。

首先，利用第 k 层的节点值 $\{u_i^k, i = 0,1,2,\cdots,N\}$ 构造 $[0,1]$ 区间上的函数 $u^k(x)$，使得 $u^k(x) = u_i^k, x \in [x_i - \frac{h}{2}, x_i + \frac{h}{2})$。在稳定性研究中，可以假设边界条件是齐次的，故 $u^k(0) = u^k(1) = 0$，所以可以将函数 $u^k(x)$ 进一步以 1 为周期进行周期延拓至$(-\infty, +\infty)$，将延拓后的周期函数仍然记为 $u^k(x)$。显然，按照这种方式延拓得到的周期函数 $u^k(x)$ 在一个周期内仅有有限和跳跃间断点，可展开为Fourier级数：

$$u^k(x) = \sum_{n=-\infty}^{+\infty} v^k(n) e^{j2n\pi x}, \tag{7.80}$$

其中 $j = \sqrt{-1}$。

对原节点系 $\{u_i^k, i = 0,1,2,\cdots,N\}$ 进行延拓之后，离散节点上的差分方程 (7.79) 也可写为连续形式

$$u^{k+1}(x) = u^k(x) + ra(u^k(x+h) - 2u^k(x) + u^k(x-h)) \tag{7.81}$$

将 $u^k(x)$ 的Fourier级数带入到公式 (7.81) 中，可得：

$$\sum_{n=-\infty}^{+\infty} v^{k+1}(n) e^{j2n\pi x} = \sum_{n=-\infty}^{+\infty} v^k(n) \left[(1-2ra) + ra(e^{j2n\pi h} + e^{-j2n\pi h})\right] e^{j2n\pi x}$$

注意到函数系 $\{e^{j2n\pi x}\}$ 是一个正交函数系，这就意味着上式等式两边的求和项中的对应求和分量都是相等的。若记 $\sigma = 2n\pi$，则等式两边求和分量都满足

$$v^{k+1}(n)e^{j\sigma x} = v^k(n)e^{j\sigma x} + ra[v^k(n)e^{j\sigma(x+h)} - 2v^k(n)e^{j\sigma x} + v^k(n)e^{j\sigma(x-h)}] \tag{7.82}$$

公式 (7.82) 可以简单看作是将 $u^k(x) = v^k(n)e^{j\sigma x}$ 带入到连续型公式 (7.81) 而得到的，事实上我们在实际中就是通过这种简单的带入来简化上面复杂的分析过程得

到公式 (7.82) 的。根据公式 (7.82)，可以得到

$$v^{k+1}(n) = G(\sigma,\tau)v^k(n) \tag{7.83}$$

其中 $G(\sigma,\tau)$ 称为差分格式的 误差传播因子，这里对于古典显格式有 $G(\sigma,\tau) = 1 - 4ra\sin^2\dfrac{\sigma h}{2}$。

公式 (7.83) 定义的误差传播因子 $G(\sigma,\tau)$ 与差分格式的稳定性有着密切的关系，且有如下重要结论：

定理 7.3 *如果存在常数 $c>0$, 使得对一切 σ, 成立*

$$|G(\sigma,\tau)| \le 1 + c\tau \tag{7.84}$$

则差分格式是稳定的。

事实上，根据 (7.83)，经过逐层递推可得

$$v^k(n) = G^k(\sigma,\tau)v^0(n)$$

当 $G(\sigma,\tau)$ 满足定理条件时，有

$$|G^k(\sigma,\tau)| = |G(\sigma,\tau)|^k \le (1+c\tau)^k \le \mathrm{e}^{cT}, \qquad k\tau \le T$$

即传播因子的任意 k 次幂都是有界的。

条件 (7.84) 通常被称为冯 • 诺伊曼(von Neumann)条件, 简称为 V-N 条件。应用该条件判别差分格式稳定性的方法, 称为傅里叶方法。该条件意味着误差的传播是可以被“控制”的。特别地，如果 $|G(\sigma,\tau)| < 1$，则误差随着层数的递增是趋向于0的。对于稳定性而言，我们有充分条件就足够了，因此在实际计算中，我们通常将 V-N 条件进一步简化为

$$|G^k(\sigma,\tau)| \le 1 \tag{7.85}$$

现在我们用傅里叶方法给出古典显格式的稳定性条件, 前面已求得传播因子为

$$G(\sigma,\tau) = 1 - 4ra\sin^2\frac{\sigma h}{2}$$

要使差分格式稳定, 只需

$$|G(\sigma,\tau)| = \left|1 - 4r\sin^2\frac{\sigma h}{2}\right| \le 1$$

从而导出古典显格式的稳定性条件为 $ra \le \frac{1}{2}$。

例 7.10 *考察古典隐格式*

$$u_i^k - r(u_{i+1}^k - 2u_i^k + u_{i-1}^k) = u_i^{k-1}$$

的稳定性。

解： 应用傅里叶方法, 令$u(x_i)=v^k(\sigma)\mathrm{e}^{j\sigma ih}$ 代入

$$v^k(\sigma)[\mathrm{e}^{j\sigma ih}-r(\mathrm{e}^{j\sigma(i+1)h}-2\mathrm{e}^{j\sigma ih}+\mathrm{e}^{j\sigma(i-1)h})]=v^{k-1}(\sigma)\mathrm{e}^{j\sigma ih}$$

约去公因子, 便得传播因子

$$G(\sigma,\tau)=\frac{1}{1-r(\mathrm{e}^{j\sigma h}-2+\mathrm{e}^{-j\sigma h})}=\frac{1}{1+4r\sin^2\dfrac{\sigma h}{2}}$$

显然, 对任何网格比r, 都有

$$|G(\sigma,\tau)|\le 1$$

古典隐格式对于任意网格比都是稳定的。

需要指出的是 V-N 条件只适合于判断两层分格式是否稳定。如果是三层差分格式, 不能直接使用上述方法, 要首先把它化成两层差分格式, 但这时得到的传播因子是一个矩阵。对传播矩阵, 有下面一些结论:

定理 7.4 *差分格式稳定的充分必要条件是当$k\tau\le T$ 时, 传播矩阵的模$G(\sigma,\tau)$一致有界, 即*

$$||G^k(\sigma,\tau)||\le M \tag{7.86}$$

其中M是与τ无关的常数。

设$\rho(G)$为矩阵$G(\sigma,\tau)$的谱半径, 由于$\rho(G)\le||G||$, 得到

定理 7.5 *差分格式稳定的必要条件是*

$$\rho(G)\le 1+c\tau \tag{7.87}$$

其中c常数。

注意: 对一个两层差分格式, $G(\sigma,\tau)$ 为一个数, 所以有 $|\lambda|=|G(\sigma,\tau)|$, 结合定理1, 知 V-N 条件是稳定性的充分且必要的条件.

当 G 为正规矩阵时, 即 $G^*G=GG^*$, G^*是 G 的伴随矩阵, 可以证明$\rho(G)=||G||$, 得出

定理 7.6 *如果传播矩阵 $G(\sigma,\tau)$ 为正规矩阵, 则*V-N*条件是差分格式稳定性的充分必要条件。*

例 7.11 *利用傅里叶方法讨论理查森格式的稳定性。*

解： 理查森格式

$$u_i^{k+1}=u_i^{k-1}+2ra(u_{i+1}^k-2u_i^k+u_{i-1}^k)$$

是一个三层格式, 用傅里叶方法讨论其稳定性, 首先把它化成等价的两层差分方程组

$$\begin{cases} u_i^{k+1} = v_i^k + 2ra(u_{i+1}^k - 2u_i^k + u_{i-1}^k) \\ v_i^{k+1} = u_i^k \end{cases}$$

令$\boldsymbol{u} = [u, v]^T$, 可把上式写成

$$\boldsymbol{u}_i^{k+1} = \begin{pmatrix} 2ra & 0 \\ 0 & 0 \end{pmatrix} \boldsymbol{u}_{i+1}^k + \begin{pmatrix} -4ra & 1 \\ 1 & 0 \end{pmatrix} \boldsymbol{u}_i^k + \begin{pmatrix} 2ra & 0 \\ 0 & 0 \end{pmatrix} \boldsymbol{u}_{i-1}^k$$

将 $\boldsymbol{u}^k(x) = v^k(n)\mathrm{e}^{j\sigma x}$ 代入上式消去因子 $\mathrm{e}^{j\sigma x}$, 整理后得到

$$v^{k+1}(n) = \begin{pmatrix} -8ra\sin^2\dfrac{\sigma h}{2} & 1 \\ 1 & 0 \end{pmatrix} v^k(n)$$

传播矩阵为

$$G(\sigma, \tau) = \begin{pmatrix} -8ra\sin^2\dfrac{\sigma h}{2} & 1 \\ 1 & 0 \end{pmatrix}$$

其特征值为

$$\lambda_{1,2} = -4ra\sin^2\frac{\sigma h}{2} \pm \left(1 + 16r^2a^2\sin^4\frac{\sigma h}{2}\right)^{\frac{1}{2}}$$

显然有

$$|\lambda_1| = \left|-4ra\sin^2\frac{\sigma h}{2} - \left(1 + 16r^2a^2\sin^4\frac{\sigma h}{2}\right)^{\frac{1}{2}}\right| > 1 + 4ra\sin^2\frac{\sigma h}{2}$$

不满足V-N条件, 所以理查森格式是不稳定的。

对方程是高维情形, 也可用傅里叶方法判别其差分格式的稳定性.

例 7.12 推导二维热传导方程定解问题的古典显格式, 并判别稳定性。

$$\begin{cases} \frac{\partial u}{\partial t} = a\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right), \quad 0 \le x, y \le 1 \\ u(x, y, 0) = \varphi(x, y), \quad 0 < x, y < 1 \\ u(0, y, t) = u(1, y, t) = u(x, 0, t) = u(x, 1, t) = 0 \end{cases} \tag{7.88}$$

解: 取 t 的步长为 τ, x, y 的步长均为 h, 则 (7.88) 式的差分显格式为

$$u_{ij}^{k+1} = (1 - 4ra)u_{ij}^k + ra(u_{i+1,j}^k + u_{i-1,j}^k) + ra(u_{i,j+1}^k + u_{i,j-1}^k) \tag{7.89}$$

其对应的延拓连续方程为

$$\begin{aligned} u^{k+1}(x, y) = (1 - 4ra)u^k(x, y) + ra(u^k(x + h, y) \\ + u^k(x - h, y)) + ra(u^k(x, y + h) + u^k(x, y - h)) \end{aligned} \tag{7.90}$$

令 $u^k(x,y)=v^k(m,n)\mathrm{e}^{\mathrm{i}(\sigma_1 x+\sigma_2 y)}$ 代入上式，并消去公因子，得到

$$G(\boldsymbol{\sigma},\tau)=1-4ra\left(\sin^2\frac{\sigma_1 h}{2}+\sin^2\frac{\sigma_1 h}{2}\right)$$

其中 $\boldsymbol{\sigma}=[\sigma_1,\sigma_2]^T$，利用V-N条件，要求

$$\left|1-4ra\left(\sin^2\frac{\sigma_1 h}{2}+\sin^2\frac{\sigma_1 h}{2}\right)\right|\leq 1$$

得到差分格式稳定的充分条件为 $ra\leq\frac{1}{4}$。

上面用傅里叶方法考察了一些差分格式的稳定性，为了区别稳定性的不同情况，我们称差分格式在网格比取 一定条件为稳定为条件稳定；差分格式对于任何网格比都稳定为无条件稳定格式；对任何网络比都不稳定为无条件不稳定。一般来说，显示差格式往往是条件稳定的，而隐式差分格式无条件稳定的要多，但因显格式计算简单，因此计算具体问题时，常常把显、隐格式交替使用。另外，通过对一些条件稳定甚至是无条件不稳定的格式，通过一些改造之后，也可能稳定。如果对理查森格式做如下修改同以 $u_i^{k+1}+u_i^{k-1}$代替 (7.68) 式中的 $2u_i^k$，则得杜福特-弗兰克尔(DuFort-Frankel)格式

$$\frac{u_i^{k+1}-u_i^{k-1}}{2\tau}-\frac{a}{h^2}[u_{i+1}^k-(u_i^{k+1}+u_i^{k-1})+u_{i-1}^k]=0 \tag{7.91}$$

其截断误差为$O(\tau^2+h^2)$，并且当$\frac{\tau}{h}\to 0$时，它是无条件稳定的差分格式，该稳定性留给读者自行证明。

（三）收敛性定理

差分格式的稳定性不仅对实际计算十分重要，对差分格式的收敛性研究也有着重要作用。

在此先给出差分格式的相容性概念.

定义 7.5 如果当 $\tau,h\to 0$ 时，差分格式的截断误差也趋于零，称差分格式与微分方程是相容的。

显然，相容性条件是用差分方程求解微分方程的必备条件，可以看出，前面介绍的古典显格式、古典隐格式、理查森格式都与扩散方程 (7.55) 是相容的。下面给出由稳定性推出收敛性的重要定理。

定理 7.7 设

$$Lu=\frac{\partial u}{\partial t}=-a\frac{\partial^2 u}{\partial x^2}=0 \tag{7.92}$$

其差分格式为

$$L_k[u]_i^k=0 \tag{7.93}$$

若 Lu与$L_k[u]_i^k$ 相容, 且差分格式对初值稳定, 则 (7.93) 式的解 u_i^k 收敛于 (7.92) 式的解 $u(x_i,t_k)$, 即

$$||\boldsymbol{u}^k-[\boldsymbol{u}]^k||\to 0,\qquad 当\tau,h\to 0时$$

式中

$$\boldsymbol{u}=[u(x_1,t_k),u(x_2,t_k),\cdots,u(x_{N-1},t_k)]^T$$

$$[\boldsymbol{u}]^k=[u_1^k,u_2^k,\cdots,u_{N-1}^k.]$$

若(7.93)式的截断误差阶为$O(\tau^\alpha+h^\beta)$, 则收敛速度有估计式

$$\boldsymbol{u}^k-[\boldsymbol{u}]^k=O(\tau^\alpha+h^\beta).$$

该定理也称拉克斯(Lax)等价定理, 我们不给予证明, 只给出一般叙述: 给定一个适定的线性初值问题, 如果逼近它的差分格式和它相容, 则差分格式的收敛性等价于差分格式的稳定性。根据这个定理, 我们可以着重讨论差分格式的稳定性, 一般不再讨论收敛性问题, 只要差分格式是稳定的, 就可以用差分格式计算微分方程近似解。

7.3.3 双曲型方程的有限差分解法

本节就最简单的双曲型方程——对流方程和波动方程, 建立差分格式, 并对其稳定性进行讨论.

7.3.3.1 双曲型方程的特征

对流方程是一阶双曲型方程, 其初值问题是

$$\begin{cases}\dfrac{\partial u}{\partial t}+a\dfrac{\partial u}{\partial x}=0,\ -\infty<x<+\infty,t\geq 0\\ u(x,0)=\varphi(x),\ -\infty<x<+\infty\end{cases}\tag{7.94}$$

其中a为常数, 令

$$x-at=\xi$$

则(7.94)式的解$u(x,t)$满足

$$\frac{\mathrm{d}u}{\mathrm{d}t}=\frac{\mathrm{d}u(at+\xi,t)}{\mathrm{d}t}=a\frac{\partial u}{\partial x}+\frac{\partial u}{\partial t}=0$$

显见, 方程的解为仅依赖于ξ的函数, 由此可知, 初值问题(7.94)的解为

$$u(x,t)=\varphi(x-at)\tag{7.95}$$

直线$x-at=c$称方程的特征线, (7.95)式指出, 初值问题(7.94)的解沿特征线是常数, 其依赖域为特征线与x轴交点的坐标, 是一点ξ, 如图7.12所示.

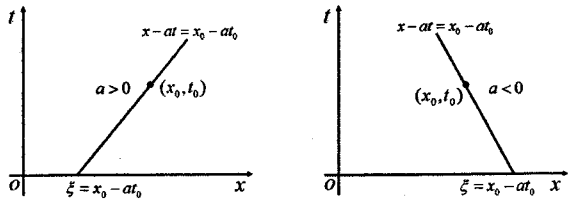


图 7.12: 双流方程特征线

波动方程为二阶双曲型方程, 初值问题是

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = a^2 \dfrac{\partial^2 u}{\partial x^2} = 0, -\infty < x < +\infty, t \geq 0 \\ u(x,0) = \varphi(x), -\infty < x < +\infty \\ \dfrac{\partial u(x,0)}{\partial t} = \varphi(x), -\infty < x < +\infty \end{cases} \tag{7.96}$$

其中$a > 0$, 令

$$\begin{cases} \xi = x + at \\ \eta = x - at \end{cases}$$

则(7.96)式成为

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = 0$$

解此方程, 可得通解为

$$u(x,t) = F_1(x+at) + F_2(x-at)$$

其中F_1, F_2为两个任意函数, 代入初始条件, 就得初值问题(7.96)的解

$$u(x,t) = \frac{1}{2}[\varphi(x+at) - \varphi(x-at)] + \frac{1}{2a}\int_{x-at}^{x+at} \varphi(\xi)\mathrm{d}\xi \tag{7.97}$$

称(7.97)为达朗贝尔(d'Alembert)公式.

由达朗贝尔不难看出, 初值问题(7.96)的解$u(x,t)$在点(x_0,t_0)的值只依赖于x轴从$x_0 - at_0$ 到$x_0 + at_0$之间的初值, 称区间$[x_0 - at_0, x_0 + at_0]$为解在点(x_0,t_0)的依赖区间, 过点(x_0,t_0) 做两条特征线

$$\begin{cases} x + at = x_0 - at_0 \\ x + at = x_0 + at_0 \end{cases}$$

则依赖区间就是它们在x轴上截得的闭区间, 图7.13所示.

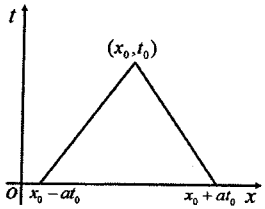


图 7.13: 波动方程特征线及依赖区间

7.3.3.2 对流方程的差分格式

网格划分如同抛物型方程, 记h为x轴方向步长, τ为t 轴方向步长, 以u_i^k代替$u(x_i,t_k)$, 可得如下差分格式.

（一）迎风格式

迎风格式的基本思想是关于x的导数用偏在特征线一侧的差商来代替, 即有

$$\begin{cases} \dfrac{u_i^{k+1}-u_i^k}{\tau}+a\dfrac{u_i^k-u_{i-1}^k}{h}=0,\ a>0 \\ \dfrac{u_i^{k+1}-u_i^k}{\tau}+a\dfrac{u_{i+1}^k-u_i^k}{h}=0,\ a<0 \end{cases} \tag{7.98}$$

方向及节点见图7.14.

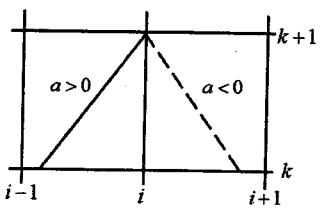


图 7.14: 迎风格式节点示意图

令$r=\dfrac{|a|\tau}{h}$, 称网格比, 整理成如下形式

$$u_i^{k+1}=u_i^k-r\begin{cases}(u_i^k-u_{i-1}^k),\ a>0\\(u_i^k-u_{i+1}^k),\ a<0\end{cases} \tag{7.99}$$

显然, (7.99)式为两层显格式, 其截断误差为$O(\tau+h)$。

用傅里叶方法判别差分格式 (7.99) 的稳定性, 先考察 $a>0$ 情形, 令 $u^k(x)=v^k(\sigma)\mathrm{e}^{-j\sigma x}$ 代入 (7.99) 的上式, 可得传播因子

$$G(\sigma,\tau)=1-r(1-\mathrm{e}^{-j\sigma h})=1-r(1-\cos\sigma h)-rj\sin\sigma h$$

因此得到

$$|G(\sigma,\tau)|^2=\left(1-2r\sin^2\frac{\sigma h}{2}\right)^2+(r\sin\sigma h)^2=1-4r(1-r)\sin^2\frac{\sigma h}{2}$$

知 $r\le 1$ 时, 有 $|G(\sigma,\tau)|\le 1$, 满足 V-N 条件, 差分格式稳定。对于 $a<0$ 的情形可做同样讨论, 稳定性条件也是 $r\le 1$。

由上述分析可见, 差分格式 (7.99) 的稳定性与特征线方程中 a 符号有关。实际上, 如果将 (7.99) 式中 a 分别变号, 差分格式则是不稳定的。对此, 下面从几何直观给予进一步说明。

我们已经知道, 对流方程 (7.94) 的解 $u(x,t)$ 在点 (x_i,t_k) 的依赖域是过该点特征线与 x 轴的交点 $(x_i-at_k,0)$, 考察差分方程 (7.99) 在 (x_i,t_k) 点有依赖域, 当$a<0$ 时, 由 (7.99) 式计算 u_i^k, 需要乃至 $k-1$ 层 u_i^{k-1},u_{i+1}^{k-1}, 计算u_i^{k-1},u_{i+1}^{k-1}, 又要用到用到 $k-2$ 层的 $u_i^{k-2},u_{i+1}^{k-2},u_{i+2}^{k-2}$, 如此递推, 用到 $k=0$ 的值 $u_i^0,\cdots,u_{i+k}^0$, 因此差分方程解在点 (x_i,t_k) 的依赖域为$[x_i,t_{i+k}]$。显然, 当 $a<0,r=\dfrac{|a|t}{h}\le 1$, 即网格对角线斜率 $\dfrac{r}{h}\le\dfrac{1}{|a|}$ 时, 微分方程的依赖域一定落在差分方程 (7.99) 的依赖域内, 见图 7.15(a). 同样, 当 $a>0$, $r\le 1$ 时, x_i-at_k 落在差分方程 (7.99) 的依赖域$[x_{i-k},x_i]$ 内, 见图 7.15(b) 所示。

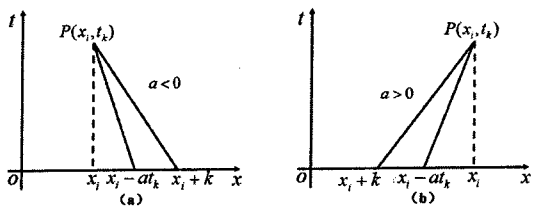


图 7.15: 迎风格式信赖域示意图

由此得到差分方程解稳定的一个必要条件, 即差分方程的依赖域必须包含微分方程的依赖域, 此条件称 柯朗(Courant) 条件, 柯朗条件实际是指出, 差分格式如与微分方程特征线走向不一致, 对任何网格比都不稳定的, 当差分格式 (7.99) 时 a 分别变号时, 就是这种情形。

注意: 柯朗条件是稳定的必要条件, 不是充分条件.

（二）蛙跳格式

对 x 和 t 导数都用中心差商来代替, 所得差分格式称为蛙跳格式

$$\frac{u_i^{k+1}-u_i^{k-1}}{\tau}+a\frac{u_{i+1}^k-u_{i-1}^k}{2h}=0 \tag{7.100}$$

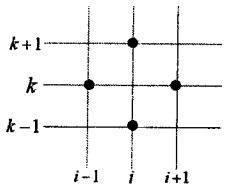


图 7.16: 蛙跳格式节点示意图

节点如图7.16所示。

容易看出, 这一差分格式对τ和h具有二阶精度, 化简整理得

$$u_i^{k+1} = u_i^{k-1} - r(u_{i+1}^k - u_{i-1}^k), \qquad a > 0 \tag{7.101}$$

是一个三层显格式。 在实际计算中, 除要使用初始条件外, 还必须用一个两层格式计算出第一层的值 u_i^1, 才能应用(7.101)式进行计算。

讨论其稳定性, 要先将三层格式化为等价的二层差分格式方程组

$$\begin{cases} u_i^{k+1} = v_i^k - r(u_{i+1}^k - u_{i-1}^k) \\ v_i^{k+1} = u_i^k \end{cases}$$

令 $\boldsymbol{u} = [u, v]^T$, 就可以求得传播矩阵为

$$G(\sigma, \tau) = \begin{pmatrix} -2r\sin\sigma h & 1 \\ 1 & 0 \end{pmatrix}$$

两个特征值为

$$\lambda_{1,2} = rj\sin\sigma h \pm \sqrt{1 - r^2\sin^2\sigma h}$$

其中$j = \sqrt{-1}$。

当$r = \dfrac{|a|\tau}{h}$时, 有

$$|\lambda_{1,2}|^2 = r^2\sin^2\sigma h + 1 - r^2\sin^2\sigma h = 1$$

满足V-N条件。需要指出的是，由于 $G(\sigma, \tau)$ 此时是一个矩阵, 所以是一个矩阵, 所以V-N条件只是稳定的必要条件, 但可以证明, 当 $r \leq 1$ 时, 这一条件也是稳定的充分条件。

（三）拉格斯-温德罗夫格式

拉格斯-温德罗夫格式是一个具有二阶精度的两层格式, 其构造方法与前面直接用差商代替导数方法有所不同, 它是利用泰勒级数展开, 并借助方程本身特点构造的。

将 $u(x_i,t_{k+1})$ 在点 (x_i,t_k) 作泰勒级数展开

$$u(x_i,t_{k+1})=u(x_i,t_k)+\tau\left[\frac{\partial u}{\partial t}\right]_i^k+\frac{\tau^2}{2!}\left[\frac{\partial^2 u}{\partial t^2}\right]_i^k+O(\tau^3)$$

利用微分方程 (7.94), 有

$$\frac{\partial u}{\partial t}=-a\frac{\partial u}{\partial x}$$

$$\frac{\partial^2 u}{\partial t^2}=\frac{\partial}{\partial t}\left(-a\frac{\partial u}{\partial x}\right)=-a\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial t}\right)=a^2\frac{\partial^2 u}{\partial x^2}$$

将上述两式代入前式, 得

$$u(x_i,t_{k+1})=u(x_i,t_k)-a\tau\left[\frac{\partial u}{\partial x}\right]_i^k+\frac{a^2\tau^2}{2}\left[\frac{\partial^2 u}{\partial x^2}\right]_i^k+O(t^3)$$

对 x 的一阶、二阶导数用中心差商代替

$$\left[\frac{\partial u}{\partial x}\right]_i^k=\frac{1}{2h}[u(x_{i+1},t_k)-u(x_{i-1},t_k)]+O(h^2)$$

$$\frac{\partial^2 u}{\partial x^2}_i^k=\frac{1}{h^2}[u(x_{i+1},t_k)-2u(x_i,t_k)+u(x_{i-1},t_k)]+O(h^2)$$

代入整理后得到

$$\begin{aligned}u(x_i,t_{k+1})=&u(x_i,t_k)-\frac{a\tau}{2h}[u(x_{i+1},t_k)-u(x_{i-1},t_k)]+\\&\frac{a^2\tau^2}{2h^2}[u(x_{i+1},t_k)-2u(x_i,t_k)+u(x_{i-1},t_k)]+O(\tau h^2)+O(\tau^2h^2)+O(\tau^3)\end{aligned}$$

略去误差项, u_i^k 代替 $u(x_i,t_k)$, 得到如下差分格式

$$u_i^{k+1}=u_i^k-\frac{a\tau}{2h}(u_{i+1}^k-u_{i-1}^k)+\frac{a^2\tau^2}{2h^2}(u_{i+1}^k-2u_i^k+u_{i-1}^k)\tag{7.102}$$

(7.102) 式称为拉格斯-温德罗夫格式, 其截断误差为 $O(\tau^2+h^2)$, 节点如图7.17 所示。

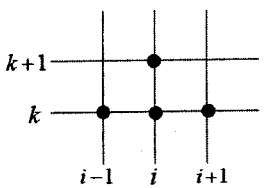


图 7.17: 拉格斯-温德罗夫格式节点示意图

令 $r=\dfrac{|a|\tau}{h}$, 就得到 $a>0$ 时的公式

$$u_i^{k+1}=u_i^k-\frac{r}{2}(u_{i+1}^k-u_{i-1}^k)+\frac{r^2}{2}(u_{i+1}^k-2u_i^k+u_{i-1}^k)\tag{7.103}$$

当 $a<0$ 时同样可得。

用傅里叶方法分析这一格式的稳定性, 容易见 (7.103) 式的传播因子为

$$G(\sigma,\tau)=1-2r^2\sin^2\frac{\sigma h}{2}-jr\sin\sigma h$$

从而得到

$$\begin{aligned}|G(\sigma,\tau)|^2&=\left(1-2r^2\sin^2\tfrac{\sigma h}{2}\right)^2+r^2\sin^2\sigma h\\&=1-4r^2(1-r^2)\sin^4\tfrac{\sigma h}{2}\end{aligned}$$

于是当 $r\le 1$ 时有 $|G(\sigma,\tau)|^2\le 1$, 知差分格式 (7.103) 在条件 $r\le 1$ 下是稳定的。

由于拉格斯-温德罗夫格式是二阶精度的两层显格式, 具有计算简单、精度高的特点， 在实用中受到广泛重视。

上述建立的差分格式都是显格式, 且一般为条件稳定, 如果寻求稳定性好的差分格式, 需建立相应隐格式, 例如迎风隐格式

$$\frac{u_i^{k+1}-u_i^k}{\tau}+a\frac{u_i^{k+1}-u_{i-1}^{k+1}}{h}=0,\qquad a>0 \tag{7.104}$$

中心隐格式

$$\frac{u_i^{k+1}-u_i^k}{\tau}+a\frac{u_{i+1}^{k+1}-u_{i-1}^{k+1}}{2h}=0,\qquad a>0 \tag{7.105}$$

它们都是无条件稳定的.

7.3.3.3 波动方程的差分格式

波动方程(7.96)的差分格式可直接由中心差商代替导数得到

$$\frac{u_i^{k+1}-2u_i^k+u_i^{k-1}}{\tau^2}-a^2\frac{u_{i+1}^k-2u_i^k+u_{i-1}^k}{2h^2}=0 \tag{7.106}$$

其截断误差为$O(\tau^2+h^2)$。令$r=\dfrac{a\tau}{h}$, 由上式可写成

$$u_i^{k+1}=r^2u_{i+1}^k+2(1-r^2)u_i^k+r^2u_{i-1}^k-u_i^{k-1} \tag{7.107}$$

这是一个三层五点显格式, 节点如图7.18所示。

初始条件用下列差分方程代替

$$u_i^0=\varphi(x_i)=\varphi_i \tag{7.108}$$

$$\frac{u_i^1-u_i^0}{\tau}=\varphi(x_i)=\varphi_i \tag{7.109}$$

容易看出, (7.109) 式的截断误差为 $O(\tau)$, 如果采用差分格式 (7.107) 计算, 两者精度不一致, 为使 (7.109) 式截断误差为 $O(\tau^2)$, 可采用中心差商代替导数

$$\frac{u_i^1-u_i^{-1}}{2\tau}=-\varphi_i \tag{7.110}$$

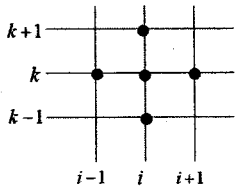


图 7.18: 波动方程三层五点显格式节点示意图

对引进的新未知数 u_i^{-1}, 在 (7.107) 式中令 $k=0$, 得

$$u_i^1 = r^2 u_{i+1}^0 + 2(1-r^2)u_i^0 + r^2 u_{i-1}^0 - u_i^{-1}$$

与(7.110)式联立消去, 得到

$$u_i^1 = \frac{r^2}{2}(\varphi_{i-1} + \varphi_{i+1}) + (1-r^2)\varphi_i + \tau\varphi_i \tag{7.111}$$

于是利用(7.108)式和(7.111)式可算出初始层($k=0$)和第一层($k=1$)网格点上的值, 再应用差分格式(7.107)就能逐层算出任意网格点的值.

关于差分格式(7.107)的稳定性讨论, 通常做法是将三层格式化成等价的两层格式方程组, 然后判断传播矩阵族$\{G(\sigma,\tau)\}$的一致有界性, 这是比较繁琐，这里不再赘述，留作习题请读者自行完成。

7.4 偏微分方程的有限元方法

偏微分方程的有限元法也是数值求解偏微分方程的重要方法之一，其基本思想也是将连续问题离散化，化成有限形式的线性代数方程组进行求解。但与有限差分法离散是从定解问题的微分或积分形式出发，用数值微分或数值积分公式导出相应的线性代数方程组的过程不同，有限元法则从定解问题的变分形式出发，用里茨-伽辽金(Ritz-Galerkin) 方法导出相应的线性代数方程组。

7.4.1 变分方法

变分方法在微分方程边值问题的近似计算中有着广泛应用, 变分方法实际上是微分学处理函数极值问题的扩展形式, 所不同的是, 变分问题是处理广泛函. 即自变量是函数的函数的极值问题, 其解是一个函数或函数组, 而函数极值问题的解为单个或有限个数值变量。在本书前面章节中介绍求解方程组的极小化方法时，已经介绍了针对方程组求解的变分原理，在本节中，我们重点介绍针对偏微分方程的变分方法。

7.4.1.1 变分方法的基本引理

定理 7.8 设$f(x)$在$[a,b]$上连续，若对任意在$[a,b]$上连续且满足条件$\eta(a)=\eta(b)=0$的函数$\eta(x)$，都有

$$\int_a^b f(x)\eta(x)\mathrm{d}x=0$$

则在$[a,b]$上$f(x)=0$。

证明：(反证法)设有$x_0\in[a,b]$，使$f(x_0)\neq 0$，不妨设$f(x_0)>0$. 由于$f(x)$的连续性，存在x_0的一个领域$(x_0-\delta,x_0+\delta)$，使在其上$f(x)>0$，取函数

$$\eta(x)=\begin{cases}[(x-x_0)^2-\delta^2]^2, & x\in(x_0-\delta,x_0+\delta)\\ 0, \text{其他}\end{cases}$$

显然$\eta(x)$在$[a,b]$连续，而

$$\int_a^b f(x)\eta(x)\mathrm{d}x\int_{x_0-\delta}^{x_0+\delta} f(x)[(x-x_0)^2-\delta^2]^2\mathrm{d}x>0$$

与已知条件矛盾.

在二维情形，同样可证下述原理:

定理 7.9 设$f(x,y)$在闭区域Ω上连续，$\partial\Omega$为Ω的边界，如果对每一个在Ω上连续且满足$\eta\Big|_{\partial\Omega}=0$的函数$\eta(x,y)$，都成立

$$\iint_\Omega f(x,y)\eta(x,y)\mathrm{d}x\mathrm{d}y=0$$

则$f(x,y)$在Ω上恒为零。

7.4.1.2 椭圆型方程边值问题的变分原理

(一)里茨形式变分原理

考察椭圆型方程第一边值问题

$$\begin{cases}Lu=-\left[\dfrac{\partial}{\partial x}\left(p\dfrac{\partial u}{\partial x}\right)+\dfrac{\partial}{\partial y}\left(p\dfrac{\partial u}{\partial y}\right)\right]+qu=f(x,y),\ (x,y)\in\Omega\\ u\Big|_{\partial\Omega}=0\end{cases}\tag{7.112}$$

其中 $p=p(x,y)>0, q=q(x,y)\geq 0$。

首先，建立泛函如下

$$J(u)=\frac{1}{2}(Lu,u)-(f,u)=\frac{1}{2}\iint_\Omega Lu\cdot u\mathrm{d}x\mathrm{d}y-\iint_\Omega fu\mathrm{d}x\mathrm{d}y\tag{7.113}$$

利用格林公式, 得

$$(Lu,v)=\iint_{\Omega}\left\{-\left[\frac{\partial}{\partial x}\left(p\frac{\partial u}{\partial x}\right)+\frac{\partial}{\partial y}\left(p\frac{\partial u}{\partial y}\right)\right]+qu\right\}v\mathrm{d}x\mathrm{d}y$$
$$=\iint_{\Omega}\left[p\left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right)+quv\right]\mathrm{d}x\mathrm{d}y-\int_{\Omega}p\frac{\partial u}{\partial n}v\mathrm{d}s$$

其中 $\frac{\partial u}{\partial n}$ 是 u 沿 $\partial\Omega$ 外法线的方向导数, 如果 u,v 都满足边界条件 (7.112), 则上式第二项积分为零, 即有

$$J(u)=\frac{1}{2}\iint_{\Omega}\left\{p\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2\right]+qu^2\right\}\mathrm{d}x\mathrm{d}y-\iint_{\Omega}fu\mathrm{d}x\mathrm{d}y \tag{7.114}$$

记

$$a(u,v)=\iint_{\Omega}\left[p\left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right)+quv\right]\mathrm{d}x\mathrm{d}y \tag{7.115}$$

则(7.114)式可写成

$$J(u)=\frac{1}{2}a(u,u)-(f,u) \tag{7.116}$$

于是, 得边值问题 (7.112) 里茨定义下的变分问题: 求函数 $u^*\in C_E^2$, 使得 $J(u^*)=\min J(u)$, 其中

$$C_E^2=\{u|u\in C^2(\Omega),u|_{\partial\Omega}=0\} \tag{7.117}$$

由 (7.115) 式定义的 $a(u,v)$ 十分重要, 称双线性泛函, 它具有如下性质:

① 对称性: $a(u,v)=a(v,u)$, 对任意 $u,v\in C_E^2$;

② 双线性: $a(u,v)$ 分别对 u,v 都是线性泛函, 即

$$a(c_1u_1+c_2u_2,v)=c_1a(u_1,v)+c_2a(u_2,v)$$

$$a(u,c_1v_1+c_2v_2)=c_1a(u,v_1)+c_2a(u,v_2)$$

③ 正定性: $a(u,u)\le\gamma||u||$, 其中 γ 是与 u 无关的正常数。

下面给出里茨形式的变分原理:

定理 7.10 设 $u\in C^2(\Omega)$, 则 u 是边值问题 (7.112) 的解的充要条件是在该边值下, u 使泛函 (7.116) 取极小值。

证明: 充分性:设 $u(x,y)\in C_E^2=\left\{u|u\in C^2(\Omega),u\Big|_{\partial\Omega}=0\right\}$, 使 $J(u)$ 取极小值, 即对任给的 $u\in C_E^2$ 及任意实数 t, 有

$$J(u+tv)\ge J(u)$$

把 $J(u+tv)$ 展开, 并应用 $a(u,v)$ 的性质, 得

$$\begin{aligned}J(u+tv)=&\frac{1}{2}(u+tv,u+tv)-(u+tv,f)\\=&J(u)+t[a(u,v)-(f,v)]+\frac{t^2}{2}a(v,v)\end{aligned}$$

$J(u+tv)$ 为 t 的一元函数, 从 $t=0$ 取得极值的必要条件

$$\frac{\partial}{\partial t}J(u+tv)\Big|_{t=0}=0$$

得

$$a(u,v)-(f,v)=0,\qquad \text{对任给}v\in C_E^2\text{成立} \tag{7.118}$$

应用格林公式, 得

$$a(u,v)-(f,v)=(Lu,v)-(f,v)=(Lu-f,v)=0$$

上式对任给的 $v\in C_E^2$ 成立, 由变分法基本引理, $Lu-f=0,(x,y)\in\Omega$. 注意到$u\in C_E^2$, 因此 $u\Big|_{\partial\Omega}=0$, 由此 $u(x,y)$ 是边值问题 (7.112) 的解.

必要性:设 $u(x,y)$ 是边值问题 (7.112) 的解即有 $a(u,v)-(f,v)=0$, 同样考察

$$\begin{aligned}J(u+tv)=&\frac{1}{2}(u+tv,u+tv)-(u+tv,f)\\=&J(u)+t[a(u,v)-(f,v)]+\frac{t^2}{2}a(v,v)=J(u)+\frac{t^2}{2}a(v,v)\end{aligned}$$

由于 $a(v,v)$ 正定, 于是 $J(u+tv)\geq J(u)$, 对任何实数 t 及 $v\in C_E^2$ 成立, 即 u 是变分问题 (7.117)的解。

这一定理也称 极小值原理, 其解称边值问题在里茨定义下的广义解。

（二）伽辽金形式变分原理

由上面证明看出, (7.118) 式

$$a(u,v)-(f,v)=0$$

是一关键式子, 它可由方程 $Lu-f=0$ 两边乘以 v, 并在 Ω 上积分直接得到, 即

$$(Lu-f,v)=a(u,v)-(f,v)=0$$

其中

$$a(u,v)=\iint_\Omega\left[\left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right)+quv\right]\mathrm{d}x\mathrm{d}y$$

$$(f,v)=\iint_\Omega fv\mathrm{d}x\mathrm{d}y$$

由此有伽辽金形式变分问题, 求 $u \in H_E^1$, 使得

$$a(u,v)-(f,v)=0 \tag{7.119}$$

对一切 $v \in H_E^1$ 成立, $H_E^1=\left\{u \left| u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \in L_2(\Omega), u\right|_{\partial\Omega}=0\right\}$。

与定理7.10证法类似, 可得伽辽金形式变分原理:

定理 7.11 设 $u \in C^2(\Omega)$, 则 u 为边值问题 (7.112) 的解的充要条件是 u 为变分问题 (7.119) 的解。

该定理也叫 虚功原理, 其解为边界问题 (7.112)在伽辽金意义下的广义解。

例 7.13 求椭圆型方程

$$\begin{cases} -\Delta u=f(x,y), & (x,y)\in\Omega \\ \left(\dfrac{\partial u}{\partial n}+\alpha u\right)\Big|_{\partial\Omega}=0, & \alpha\geq 0 \end{cases}$$

对应的变分问题。

解: 由格林公式, 得

$$\begin{aligned}(-\Delta u-f,v)&=\iint_\Omega -\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)v\mathrm{d}x\mathrm{d}y-\iint_\Omega fv\mathrm{d}x\mathrm{d}y\\ &=\iint_\Omega\left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right)\mathrm{d}x\mathrm{d}y-\int_{\partial\Omega}\frac{\partial u}{\partial n}v\mathrm{d}s-\iint_\Omega fv\mathrm{d}x\mathrm{d}y\\ &=\iint_\Omega\left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right)\mathrm{d}x\mathrm{d}y+\int_{\partial\Omega}auv\mathrm{d}s-\iint_\Omega fv\mathrm{d}x\mathrm{d}y\end{aligned}$$

定义

$$a(u,v)=\iint_\Omega\left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right)\mathrm{d}x\mathrm{d}y+\int_{\partial\Omega}\alpha yv\mathrm{d}s$$

$$(f,v)=\iint_\Omega fv\mathrm{d}x\mathrm{d}y$$

于是变分问题提法是: 求$u \in H_E^1$, 使对任给$v \in H_E^1$ 满足伽辽金方程

$$a(u,v)-(f,v)=0$$

注意: 利用格林公式

$$a(u,v)-(f,v)=\iint_\Omega(-\Delta u-f)v\mathrm{d}x\mathrm{d}y+\int_{\partial\Omega}\left(\frac{\partial u}{\partial n}+\alpha u\right)v\mathrm{d}s$$

边界条件在这里被自然满足, 所以第二、第三边界条件称自然边界条件, 而称第一边界条件为本质边界条件.

7.4.1.3 里茨–伽辽金方法在微分方程近似解法中的应用

前面讨论了如何化边值问题为等价的变分问题, 本节给出变分问题的近似解法——里茨–伽辽金方法, 这一方法在解微分方程边值问题中有着广泛应用, 同以前的函数极值问题相比, 变分问题的难点是在无穷维空间上求泛函数的极值, 因此解决问题的关键要选取有穷维空间近似代替无穷维空间.

以v表示C_E^2, H_E^1等无穷维函数空间, V_n是V的n维子空间, $\varphi_1, \varphi_2, \cdots, \varphi_n$为$V_n$的一组基函数, 于是对$V_n$的任一函数$u_n$, 有

$$u_n = \sum_{i=1}^{n} c_i \varphi_i \tag{7.120}$$

里茨方法的要点是:

(1) 将u_n代入$J(u)$

$$J(u_n) = \frac{1}{2} a(u_n, u_n) - (f, u_n) \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a(\varphi_i, \varphi_j) c_i c_j - \sum_{j=1}^{n} c_j (f, \varphi)$$

得$c_1, c_2, \cdots, c_n$的二次函数;

(2) 令$J(u_n)$取极小值, 由

$$\frac{\partial J(u_n)}{\partial c_j} = 0, \qquad j = 1, 2, \cdots, n$$

得$c_1, c_2, \cdots, c_n$的线性方程组

$$\sum_{i=1}^{n} c_i a(\varphi_i, \varphi_j) = (f, \varphi_j), \qquad j = 1, 2, \cdots, n \tag{7.121}$$

(3) 求出$c_i (i = 1, 2, \cdots, n)$代入(7.122)式, 就得到近似解$u_n$.

伽辽金方法的步骤是:

(1) 将(7.122)式代入伽辽金方程

$$a(u_n, v_n) = (f, v_n), \qquad \text{对任意} v_n \in V_n$$

得方程组

$$\sum_{i=1}^{n} c_i a(\varphi_i, \varphi_j) = (f, \varphi_j), \qquad j = 1, 2, \cdots, n \tag{7.122}$$

(2) 解出$c_1, c_2, \cdots, c_n$, 得到近似解u_n.

上述方程组与里茨方法导出的方程组(7.121)完全相同, 可见在处理边值问题的变公问题量, 里茨与伽辽金方法是一样的, 但里茨法要求$a(u,v)$对称正定, 伽辽金方法则没有这些限制, 所以应用更为广泛.

无论是用里茨法还是伽辽金法求近似解, 基函数的选取都是非常重要的, 在有限元方法出现以前, 通常选取代数或三角多项式作为基函数, 并要求所选基函数满足本质边界条件, 这就使得实际应用中里茨-伽辽金方法会遇到较多困难.

7.4.2 偏微分方程的有限元方法

有限元方法是针对里茨-伽辽金方法的缺陷, 在变分原理基础上, 通过对区域的剖分和插值建立起来的方法, 具有基函数选取容易, 边界条件处理简单, 便于上机计算等优点, 下面以二阶椭圆方程为例介绍有限元法解边界问题的基本思想. 设

$$\begin{cases} -\Delta u = f(x,y), & (x,y) \in \Omega \\ \dfrac{\partial u}{\partial n} + \alpha u = g(x,y), & (x,y) \in \partial\Omega \end{cases} \tag{7.123}$$

对应变分问题的伽辽金方程方程为

$$(-\Delta u - f, v) = \iint_{\Omega} \left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} \right) \mathrm{d}x\mathrm{d}y - \iint_{\Omega} f v \mathrm{d}x\mathrm{d}y + \int_{\partial\Omega} (\alpha u - g) v \mathrm{d}s = 0 \tag{7.124}$$

7.4.2.1 进行区域的剖分

常用方法是把区域Ω剖分成三角形组合, 如图Fig:7-19, 称三角形的顶点为节点, 每一个三角形为单元。

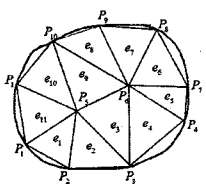


图 7.19: 有限元三角网格节点示意图

设节点为 $P(x_i,y_i), i = 1,2,\cdots,n$; 单元为 $e_k, k = 1,2,\cdots,m$ 对区域 Ω 进行三角剖分及节点编号时, 应注意以下几点:

(1) 每个单元顶点一定是相邻单元顶点, 不能是相邻单元边上的内点;

(2) 三角形单元的大小不一定相等, 但尽量避免出现大的钝角;

(3) 在$u(x,y)$梯度变化较大的地方, 网格要适当加密, 梯度变化小的地方可相对稀一些;

(4) 单元的编号可以任意, 但节点编号会直接影响总刚度矩阵元素的排列, 故要求相邻节点编号差的绝对值尽可能小, 以减小计算量.

7.4.2.2 构造插值函数

所谓构造插值函数, 就是用分片光滑曲面来近似代替 Ω 上的光滑曲面 $u(x,y)$, 为了简单, 通常采用线性插值, 此时分片光滑曲面是三角形单元上的空间平面, 设单元 e_n 的三顶点分别为 $P_i(x_i,y_i)$, $P_j(x_j,y_j)$, $P_m(x_m,y_m)$, 对应函数 $u(x,y)$ 的值为 u_i,u_j,u_m, 做线性插值函数

$$u_n(x,y)=ax+by+c \tag{7.125}$$

使得

$$\begin{cases} u_n(x_i,y_i)=ax_i+by_i+c=u_i \\ u_n(x_j,y_j)=ax_j+by_j+c=u_j \\ u_n(x_m,y_m)=ax_m+by_m+c=u_m \end{cases} \tag{7.126}$$

求解此方程组, 可得

$$a=\frac{1}{2\Delta e}\begin{vmatrix} u_i & y_i & 1 \\ u_j & y_j & 1 \\ u_m & y_m & 1 \end{vmatrix},\quad b=\frac{1}{2\Delta e}\begin{vmatrix} x_i & u_i & 1 \\ x_j & u_j & 1 \\ x_m & u_m & 1 \end{vmatrix},\quad c=\frac{1}{2\Delta e}\begin{vmatrix} x_i & y_i & u_i \\ x_j & y_j & u_j \\ x_m & y_m & u_m \end{vmatrix}$$

其中

$$\Delta e=\frac{1}{2}\begin{vmatrix} x_i & y_i & 1 \\ x_j & y_j & 1 \\ x_m & y_m & 1 \end{vmatrix}$$

为单元 e_n 的面积.

把上述 a,b,c 代入 (7.125) 式, 并对 u_i,u_j,u_m 进行同类项合并, 得到

$$u_n(x,y)=N_i(x,y)u_i+N_j(x,y)u_j+N_m(x,y)u_m \tag{7.127}$$

其中

$$N_i(x,y)=\frac{1}{2\Delta e}\begin{vmatrix} x & y & 1 \\ x_j & y_j & 1 \\ x_m & y_m & 1 \end{vmatrix},\quad N_j(x,y)=\frac{1}{2\Delta e}\begin{vmatrix} x & y & 1 \\ x_m & y_m & 1 \\ x_i & y_i & 1 \end{vmatrix},$$

$$N_m(x,y)=\frac{1}{2\Delta e}\begin{vmatrix} x & y & 1\\ x_i & y_i & 1\\ x_j & y_j & 1\end{vmatrix} \tag{7.128}$$

函数 $N_i(x,y),N_j(x,y),N_m(x,y)$ 称单元 e_n 上的基函数. 满足

$$N_s(x_t,y_t)=\delta_{st}=\begin{cases}1, & s=t\\ 0, & s\neq t\end{cases}\quad s,t=i,j,m$$

且有

$$N_i+N_j+N_m=1$$

引入记号

$$N=[N_i(x,y),N_j(x,y),N_m(x,y)],\quad u_e=[u_i,u_j,u_m]^T$$

则在单元 e_n 上, 有

$$u_n=Nu_e^T \tag{7.129}$$

u_n 的梯度向量可表为

$$\nabla u_n=\begin{pmatrix}\dfrac{\partial u_n}{\partial x}\\ \dfrac{\partial u_n}{\partial y}\end{pmatrix}=Bu_e \tag{7.130}$$

其中

$$B=\begin{pmatrix}\dfrac{\partial N_i}{\partial x} & \dfrac{\partial N_j}{\partial x} & \dfrac{\partial N_m}{\partial x}\\ \dfrac{\partial N_i}{\partial y} & \dfrac{\partial N_j}{\partial y} & \dfrac{\partial N_m}{\partial y}\end{pmatrix}$$

由 (7.128) 式可看出, N_i,N_j,N_m 是两三角形面积之比, 也称点 (x,y) 的面积坐标, 如图 7.20 所示。由此知, (7.127) 式或 (7.129) 式就是单元 e_n 上的线性插值函数。

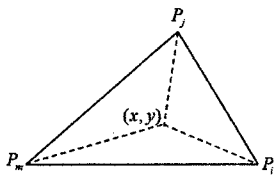


图 7.20: 面积坐标示意图

当点 (x,y) 在 $\Delta P_iP_jP_m$ 的某一条边上时, 例如在 $\overline{p_iP_j}$, 若 $l=|\overline{p_iP_j}|$, t 为点 (x,y) 与 P_i 的距离, 这时 $N_m(x,y)=0$, 成为两点插值函数

$$u_n(x,y)=\frac{l-t}{l}u_i+\frac{t}{l}u_j$$

7.4.2.3 形成有限元方程

由对区域 Ω 有三角划分, (7.124) 式就可写成

$$\sum_n \iint_{e_n} \nabla u_n \cdot \nabla v_n \mathrm{d}x\mathrm{d}y + \sum_n \int_{\gamma_n} \alpha u_n v_n \mathrm{d}s = \sum_n \iint_{e_n} f v_n \mathrm{d}x\mathrm{d}y + \sum_n \int_{\gamma_n} g v_n \mathrm{d}s \quad (7.131)$$

其中 $\gamma_n = \partial e_n \cap \partial\Omega$. 当 e_n 的边至少有一条是在 $\partial\Omega$ 上, 称这样的 e_n 为边界元.

(1) 单元刚度矩阵的计算

设 u_n, v_n 在单元 e_n 为 $\Delta P_i P_j P_m$ 三顶点的值分别是 u_i, u_j, u_m 与 v_i, v_j, v_m。 记

$$u_e = \begin{pmatrix} u_i \\ u_j \\ u_m \end{pmatrix}, \qquad v_e = \begin{pmatrix} v_i \\ v_j \\ v_m \end{pmatrix}$$

当 e_n 是内部元时, 由(7.130)式

$$\iint_{e_n} \nabla u_n \cdot \nabla v_n \mathrm{d}x\mathrm{d}y = \iint_{e_n} (\nabla v_n)^T (\nabla u_n) \mathrm{d}x\mathrm{d}y = \iint_{e_n} [Bv_e]^T [Bu_e] \mathrm{d}x\mathrm{d}y = v_e^T [K]_e u_e$$

其中

$$[K]_e = \iint_{e_n} B^T B \mathrm{d}x\mathrm{d}y = \Delta e B^T B = \begin{pmatrix} k_{ii} & k_{ij} & k_{im} \\ k_{ji} & k_{jj} & k_{jm} \\ k_{mi} & k_{mj} & k_{mm} \end{pmatrix} \quad (7.132)$$

$[K]_e$ 称单元刚度矩阵, 式中 k_{st} 计算公式为

$$k_{st} = \Delta e \left(\frac{\partial N_s}{\partial x} \frac{\partial N_t}{\partial x} + \frac{\partial N_s}{\partial y} \frac{\partial N_t}{\partial y} \right), \qquad s, t = i, j, m$$

如果 e_n 是边界元, 设 γ_n 是 $\overline{P_i P_j}, |\overline{P_i P_j}| = l$, 由 $N = \left[1 - \frac{t}{l}, \frac{t}{l}, 0\right]$, 还需计算

$$\int_{\gamma_n} \alpha u_n v_n \mathrm{d}s = \int_0^l \alpha [Nv_e]^T [Nu_e] \mathrm{d}t = v_e^T [\bar{K}]_e u_e$$

其中

$$[\bar{K}]_e = \int_0^l \alpha N^T N \mathrm{d}t = \begin{pmatrix} \bar{k}_{ii} & \bar{k}_{ij} & \bar{k}_{im} \\ \bar{k}_{ji} & \bar{k}_{jj} & \bar{k}_{jm} \\ \bar{k}_{mi} & \bar{k}_{mj} & \bar{k}_{mm} \end{pmatrix} \quad (7.133)$$

式中

$$\bar{k}_{ii} = \int_0^l \alpha N_i^2 \mathrm{d}t = \int_0^l \alpha \left(1 - \frac{t}{l}\right)^2 \mathrm{d}t$$

$$\bar{k}_{ij}=\bar{k}_{ji}=\int_0^l \alpha N_i N_j \mathrm{d}t=\int_0^l \alpha\left(1-\frac{t}{l}\right)\frac{t}{l}\mathrm{d}t$$

$$\bar{k}_{jj}=\int_0^l \alpha\left(\frac{t}{l}\right)^2 \mathrm{d}t,\quad \bar{k}_{sm}=\bar{k}_{ms}=0,\quad s=i,j,m$$

因此, 对 e_n 是边界元, 它的单元刚度矩阵应是 $[K]_e+[\bar{K}]_e$, 其元素是$k_{st}+\bar{k}_{st}, s,t=i,j,m$, 仍简记为 $[K]_e$ 及 k_{st}。

用同样方法可计算 (7.131) 式的右端积分项. 当 e_n 为内部元时, 只有一项

$$\iint_{e_n} f v_n \mathrm{d}x\mathrm{d}y=\iint_{e_n}[Nv_e]^T f\mathrm{d}x\mathrm{d}y=v_e^T F_e$$

其中

$$F_e=\iint_{e_n} N^T f\mathrm{d}x\mathrm{d}y=\begin{bmatrix}F_i\\F_j\\F_m\end{bmatrix},\quad F_s=\iint_{e_n} N_s f\mathrm{d}x\mathrm{d}y,\quad s=i,j,m \tag{7.134}$$

这里 F_e 称单元荷载向量, 当 e_n 为边界元时, 仍设 $\overline{P_iP_j}$ 为 γ_n, 此时$N=\left[1-\frac{t}{l},\frac{t}{l},0\right]$, 还应计算

$$\int_{\gamma_n} g v_n \mathrm{d}s=\int_{\gamma_n}[Nv_e]^T g\mathrm{d}s=v_e^T\bar{F}_e$$

其中

$$\bar{F}_e=\int_{\gamma_n} N^T g\mathrm{d}s=\begin{pmatrix}\int_0^l\left(1-\frac{t}{l}\right)g\mathrm{d}t\\ \int_0^l\frac{t}{l}g\mathrm{d}t\\ 0\end{pmatrix} \tag{7.135}$$

对边界元, 单元荷载向量为 $F_e+\bar{F}_e$, 仍简记 F_e。

(2) 总刚度矩阵的计算

将各单元刚度矩阵及荷载向量分别扩展成 N 阶矩阵及 N 维向量, 单元节点的序号 i,j,m 按实际大小放在扩展后的矩阵和向量的相应行列上, 记号设为 $[K]_{e_n}$ 及 F_{e_n} 有

$$[K]_{e_n}=\begin{pmatrix}
 & \vdots & & \vdots & & \vdots & \\
\cdots & k_{ii} & \cdots & k_{ij} & \cdots & k_{im} & \cdots\\
 & \vdots & & \vdots & & \vdots & \\
\cdots & k_{ji} & \cdots & k_{jj} & \cdots & k_{jm} & \cdots\\
 & \vdots & & \vdots & & \vdots & \\
\cdots & k_{mi} & \cdots & k_{mj} & \cdots & k_{mm} & \cdots\\
 & \vdots & & \vdots & & \vdots &
\end{pmatrix},\quad F_{e_n}=\begin{pmatrix}\vdots\\F_i\\\vdots\\F_j\\\vdots\\F_m\\\vdots\end{pmatrix}$$

矩阵及向量中其他元素皆为零.

设 $\boldsymbol{u}=[u_1,u_2,\cdots,u_N]^T$, $\boldsymbol{v}=[v_1,v_2,\cdots,v_N]^T$, 则 (7.131) 式成为

$$\begin{aligned}\sum_n v_e^T[K]_e u_e-\sum_n v_e^T F_e&=\sum_n \boldsymbol{v}[K]_{e_n}\boldsymbol{u}-\sum_n \boldsymbol{v}^T F_{e_n}\\&=\boldsymbol{v}^T\sum_n[K]_{e_n}\boldsymbol{u}-\boldsymbol{v}^T\sum_n F_{e_n}\\&=\boldsymbol{v}^T[K]\boldsymbol{u}-\boldsymbol{v}^T F\\&=0\end{aligned}\tag{7.136}$$

式中 $[K]=\sum_n[K]_{e_n}$ 称总刚度矩阵, 是各单元刚度矩阵叠加组成, $F=\sum_n F_{e_n}$ 提前总荷载向量, 是各单元荷载向量叠加组成。

由(7.136)式, 得

$$\boldsymbol{v}^T([K]\boldsymbol{u}-F)=0\tag{7.137}$$

对一切向量 $\boldsymbol{v}$ 都成立, 就得到 $\boldsymbol{u}$ 满足的方程组

$$[K]\boldsymbol{u}-F=0\tag{7.138}$$

该方程组称为有限元方程, 可以证明, 矩阵 $[K]$ 对称、正定, 所以 (7.138) 式有唯一解 $\boldsymbol{u}=[u_1,u_2,\cdots,u_N]^T$.

7.4.2.4 约束条件的处理

上面讨论的是第三边值问题, 其边界条件为自然边界条件, 不必进行处理, 若是第一边值问题, 则需进行约束处理, 设 $P_1,P_2,\cdots,P_l$ 是边界节点, $P_{l+1},\cdots,P_N$ 是其他节点, 考虑在 $\partial\Omega$ 上零的条件, 取

$$\boldsymbol{v}=[0,\cdots,0,v_{l+1},\cdots,v_N]^T$$

(7.137)式可写成

$$[0,(\boldsymbol{v}^2)^T]\left(\begin{pmatrix}\boldsymbol{k}_{11}&\boldsymbol{k}_{12}\\\boldsymbol{k}_{21}&\boldsymbol{k}_{22}\end{pmatrix}\begin{pmatrix}\boldsymbol{u}^1\\\boldsymbol{u}^2\end{pmatrix}-\begin{pmatrix}F^1\\F^2\end{pmatrix}\right)=0$$

其中 $\boldsymbol{v}^2=[v_{l+1},\cdots,v_N]^T$, $\boldsymbol{u}^1$, F^1 是1至l行元素组成向量, $k_{11},k_{12},k_{21},k_{22}$ 为矩阵 $[K]$ 的分块矩阵.

$$\boldsymbol{k}_{11}=\begin{pmatrix}k_{11}&\cdots&k_{1l}\\\vdots&&\vdots\\k_{l1}&\cdots&k_{ll}\end{pmatrix},\qquad \boldsymbol{k}_{22}=\begin{pmatrix}k_{l+1\ l+1}&\cdots&k_{l+1\ N}\\\vdots&&\vdots\\k_{N\ l+1}&\cdots&k_{N\ N}\end{pmatrix}$$

上式展开为

$$(\boldsymbol{v}^2)^T(\boldsymbol{k}_{22}\boldsymbol{u}^2+\boldsymbol{k}_{21}\boldsymbol{u}^1-F^2)=0$$

它对一切 v^2 都成立, 所以得到

$$\boldsymbol{k}_{22}\boldsymbol{u}^2 = F^2 - \boldsymbol{k}_{21}\boldsymbol{u}^1 \tag{7.139}$$

这是一个 N-l 阶方程方程组, 其系数矩阵对称正定, 实际 计算中, 也可保留 $[K]$ 的阶数, 将方程组改成

$$\begin{pmatrix} E_l & 0 \\ 0 & \boldsymbol{k}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{u}^1 \\ \boldsymbol{u}^2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{u}_0^1 \\ F^2 - \boldsymbol{k}_{21}\boldsymbol{u}_0^1 \end{pmatrix}$$

即

$$\begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 1 & \cdots & 0 \\ \vdots & & \vdots & & \\ & & & \boldsymbol{k}_{22} & \\ 0 & \cdots & 0 & & \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_l \\ u_{l+1} \\ \vdots \\ u_N \end{pmatrix} = \begin{pmatrix} u_1^0 \\ \vdots \\ u_l^0 \\ \vdots \\ F_i - \sum\limits_{j=1}^{l} k_{ij}u_j^0 \\ \vdots \end{pmatrix}$$

其中 $u_j^0 (j = 1, 2, \cdots, l)$ 是 $\boldsymbol{u}$ 在约束边界点 $P_1, \cdots, P_l$ 的值, 在齐次边界条件时, $u_j^0 = 0, j = 1, 2, \cdots, l$。

综上, 可看出用有限元方法解边值问题的步骤:

(1) 写出微分方程对应的变分形式;

(2) 选定单元形状, 对求解区域进行剖分, 确定各单元编号及节点编号, 求出节点坐标;

(3) 利用 (7.132) 式计算单元刚度矩阵, 用 (7.134) 式计算单元荷载向量, 注意对边界元还要到用(7.133) 式及 (7.135)式;

(4) 合成总刚度矩阵及总荷载向量;

(5) 处理边界条件;

(6) 解有限元方程组, 求得节点处解的近似值 $u_1, u_2, \cdots, u_N$。

目前工程技术领域，已经开发了许多著名的有限元分析软件，比较常用的有ANSYS，Patran/Nastran，Abaqus，Hypermesh等等，这几种软件基本功能都类似，只是界面操作方式不同而已。虽然已经有了这么多的有限元分析软件，但在本节中，我们还是介绍了有限元方法的基本数学原理以及计算流程，使读者能够了解到软件背后蕴含的数学建模的思想和方法。

第七章习题

1. 设有一根长为 的均匀柔软细弦，当它作微小横振动时,除受内部张力外，还受到周围介质所产生的阻尼力的作用，阻尼力与速度的平方成正比(比例系数为b)，试写出带有阻尼力的弦振动方程。

2. 选已知函数 $w(x,t)$，并作函数代换 $v=u-w$ 将以下边界条件齐次化

   (1) $u(0,t)=t,u(2,t)=\sin t$;

   (2) $u(0,t)=1,\ u_x(l,t)=1+t^2$;

   (3) $u_x(0,t)=\varphi(t),u_x(3,t)=\psi(t)$;

   (4) $u_x(0,t)=t^2,u_x(2,t)+u(2,t)=t$.

3. 考虑如下有界弦振动方程定解问题

$$\begin{cases} u_{tt}=a^2u_{xx},0<x<l,t>0 \\ u(0,t)=0,u(l,t)=0,t\geq 0 \\ u(x,0)=0,u_t(x,0)=\psi(x),0\leq x\leq l. \end{cases}$$

   (1)将 $\psi(x)$ 在 $[0,l]$ 按正交函数系 $\{\sin\frac{n\pi}{l}x\}_{n\geq 1}$ 展成Fourier 级数,并求Fouier系数;

   (2) 对于任意的整数$n\geq 1$，验证$u_n(x,t)=\frac{l}{n\pi a}\psi_n\sin\frac{n\pi a}{l}t\sin\frac{n\pi}{l}x$是如下问题的解

$$\begin{cases} u_{tt}=a^2u_{xx},0<x<l,t>0 \\ u(0,t)=0,u(l,t)=0,t\geq 0 \\ u(x,0)=0,u_t(x,0)=\psi_n\sin\frac{n\pi}{l}x,0\leq x\leq l. \end{cases}$$

   (3) 利用(2)中的结果试写出问题(1)中的定解问题的解。

4. 用古典显格式在$h=0.2,r=\dfrac{1}{6}$时, 计算

$$\begin{cases} \dfrac{\partial u}{\partial t}=\dfrac{\partial^2 u}{\partial t^2}, & 0<x<1,\ \ t>0 \\ u(x,0)=4x(1-x), & 0\leq x\leq 1 \\ u(0,t)=u(1,t)=0, & t\geq 0 \end{cases}$$

   前两层的差分解。

5. 用显格式解 取$r=1, h=0.2$, 求$k=1,2$层上的差分解.

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - \dfrac{\partial^2 u}{\partial x^2} = 0, & 0 < x < 1, \quad t > 0 \\ u(x,0) = \sin \pi x, & \dfrac{\partial u(x,0)}{\partial t} = x(1-x), \quad 0 \le x \le 1 \\ u(0,t) = u(1,t) = 0, & t \ge 0 \end{cases}$$

5. 用有限元法解下列边值问题

$$\begin{cases} -\Delta u + 12u = 18xy + 2, & (x,y) \in \Omega \\ \dfrac{\partial u}{\partial n} + 3u\Big|_{\partial\Omega} = 0 \end{cases}$$

其中Ω是图7.21表示的正方形, $\partial\Omega$是其边界, 设Ω剖分成两个三角形.

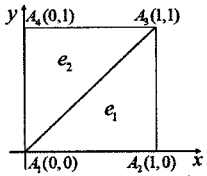


图 7.21: 三角剖分示意图

第八章　统计分析

本章主要介绍一元与多元线性回归模型与方法，单因素方差分析，然后介绍已经取得蓬勃发展和广泛应用的贝叶斯统计方法，最后介绍有着重要应用的多元正态总体的统计分析方法。

8.1　一元线性回归

现实世界中，普遍存在着变量之间的各种关系。这些变量之间的关系常见的主要包括两类：一类是确定性关系，如圆的面积公式、勾股定理等。这样的确定关系可用变量之间的函数关系明确表示；另一类是相关关系，比如人的身高与体重的关系，作物收成与降雨、施肥的关系等，这类关系是非确定的。研究变量间相关关系的统计分析方法称为回归分析。可通过回归分析研究一个变量对另一个或多个变量的依赖关系，且可应用这种方法解决估计和预测问题。这节我们先介绍一元线性回归，在下节我们再介绍多元线性回归。

8.1.1　一元线性回归模型

考虑两个变量x与y之间的关系，假设x表示人的年龄，y表示人的血压。显然，变量x的值影响y的值，一般随着人的年龄增加，血压要增高。但是，同龄人（即x相同）的血压并不一定相同。这个例子中，变量x与y是有联系的，但是当x的值确定时，y的值却是不确定的。这种变量间的关系就是相关关系。研究变量间相关关系的统计分析方法称为回归分析。这里x称为自变量，y称为因变量。在回归分析中，因变量y是随机变量，自变量x可以是随机变量，也可以是非随机的确定变量。在实际应用中，自变量x通常看成是可控变量或是可精确测量的普通变量。

对于自变量x的每一个确定值，都有一个随机变量y与之相对应。因为y是随机变量，所以在不同的试验中的y观察值是不同的。如果当x的值确定时，y与之对应的条件的数学期望$E(y|x)$存在，那么$E(y|x)$称为y关于x的回归函数。于是自变量与因变量之间的关系可以用如下的模型来描述

$$\begin{cases} y = E(y|x) + \varepsilon \\ E(\varepsilon) = 0,\ \mathrm{var}(\varepsilon) = \sigma^2 \end{cases} \tag{8.1}$$

这里ε是不可观测的随机误差，它是一个随机变量。随机误差ε应理解为除x外对y有

影响但未加考虑的诸多因素（包括随机因素）所产生的影响总和。(8.1)称为一元回归模型，基于(8.1)的的统计分析称为一元回归分析。

如果回归函数$E(y|x)$是x的线性函数，即$E(y|x)=\beta_0+\beta_1 x$，则模型(8.1)可化为

$$\begin{cases} y=\beta_0+\beta_1 x+\varepsilon \\ E(\varepsilon)=0,\ \mathrm{var}(\varepsilon)=\sigma^2 \end{cases} \tag{8.2}$$

其中β_0和β_1是未知参数，称β_0为回归常数，称β_1 称为回归系数。(8.2)称为一元线性回归模型，基于(8.2) 的的统计分析称为一元线性回归分析。若进一步要求$\varepsilon\sim N(0,\sigma^2)$，即

$$\begin{cases} y=\beta_0+\beta_1 x+\varepsilon \\ \varepsilon\sim N(0,\sigma^2) \end{cases} \tag{8.3}$$

则称为一元正态线性回归模型。

如果回归函数$E(y|x)$不是x的线性函数，则对应的(8.1)称为一元非线性回归模型，此时基于(8.1)的的统计分析称为一元非线性回归分析。本节只考虑一元线性回归。

对(x,y)进行观察，得到n个观察值

$$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$$

如果符合模型式(8.1)，则

$$\begin{cases} y_i=\beta_0+\beta_1 x_i+\varepsilon_i \quad (i=1,2,\cdots,n) \\ \varepsilon_i\text{相互独立},\ E(\varepsilon_i)=0,\ \mathrm{var}(\varepsilon_i)=\sigma^2 \end{cases} \tag{8.4}$$

(8.4)称为一元线性回归的样本模型。

显然，y_i相互独立且

$$E(y_i)=\beta_0+\beta_1 x_i,\ \mathrm{var}(y_i)=\sigma^2$$

如果符合模型式(8.2)，则

$$\begin{cases} y_i=\beta_0+\beta_1 x_i+\varepsilon_i \quad (i=1,2,\cdots,n) \\ \varepsilon_i\text{相互独立},\ (\varepsilon_i)\sim N(0,\sigma^2) \end{cases} \tag{8.5}$$

(8.5)称为一元正态线性回归的样本模型。

显然，y_i相互独立且

$$E(y_i)=\beta_0+\beta_1 x_i,\ \mathrm{var}\ (y_i)=\sigma^2,\ y_i\sim N\left(\beta_0+\beta_1 x_i,\sigma^2\right).$$

说明: 样本模型中的y_i是随机变量，(x_i,y_i)中的y_i 是随机变量y_i的观察值。

8.1.2 参数的最小二乘估计

本节考虑一元线性回归，其模型为

$$y = \beta_0 + \beta_1 x + \varepsilon$$

其中β_0与β_1都是未知参数，分别表示直线$E(y|x) = \beta_0 + \beta_1 x$的截距和斜率。一元线性回归分析的主要任务就是用适当的统计方法获得β_0与β_1的估计值$\hat{\beta}_0$与$\hat{\beta}_1$。称

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{8.6}$$

为y关于x的一元线性经验回归方程或简称回归方程，方程对应的直线称为回归直线。$\hat{\beta}_0$与$\hat{\beta}_1$分别是回归直线的截距与斜率。

记

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - E(y_i)]^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2 \tag{8.7}$$

所谓的最小二乘法，就是寻找参数β_0与β_1的估计值$\hat{\beta}_0$与$\hat{\beta}_1$，使上式达到最小，即寻找$\hat{\beta}_0$与$\hat{\beta}_1$，满足

$$\begin{aligned} Q(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned} \tag{8.8}$$

由此求出的$\hat{\beta}_0$与$\hat{\beta}_1$称为参数β_0与β_1的最小二乘估计。称

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{8.9}$$

为$y_i (i = 1, 2, \cdots, n)$的回归拟合值，简称回归值或拟合值。为求最小二乘估计，由微积分学中的极值原理，就是要解下面的方程组

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0}\Big|_{\beta_0 = \hat{\beta}_0} = -2\sum\limits_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \dfrac{\partial Q}{\partial \beta_1}\Big|_{\beta_1 = \hat{\beta}_1} = -2\sum\limits_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \end{cases} \tag{8.10}$$

经整理后，得到正规方程组

$$\begin{cases} n\hat{\beta}_0 + (\sum\limits_{i=1}^{n} x_i)\hat{\beta}_1 = \sum\limits_{i=1}^{n} y_i \\ (\sum\limits_{i=1}^{n} x_i)\hat{\beta}_0 + (\sum\limits_{i=1}^{n} x_i^2)\hat{\beta}_1 = \sum\limits_{i=1}^{n} x_i y_i \end{cases} \tag{8.11}$$

解正规方程组得β_0与β_1的最小二乘估计为：

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\sum\limits_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \end{cases} \tag{8.12}$$

式中$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$，记

$$l_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n(\bar{x})^2=\sum_{i=1}^{n}(x_i-\bar{x})x_i \tag{8.13}$$

$$l_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}=\sum_{i=1}^{n}(x_i-\bar{x})\,y_i \tag{8.14}$$

于是(8.12)式可简写为

$$\begin{cases}\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}\\ \hat{\beta}_1=l_{xy}/l_{xx}\end{cases} \tag{8.15}$$

定理 8.1 对于一元线性回归的样本模型(8.4)

$$\begin{cases}y_i=\beta_0+\beta_1x_i+\varepsilon_i\quad(i=1,2,\cdots,n)\\ \varepsilon_i\text{相互独立},\ E(\varepsilon_i)=0,\ \mathrm{var}(\varepsilon_i)=\sigma^2\end{cases}$$

β_0与β_1的最小二乘估计$\hat{\beta}_0$与$\hat{\beta}_1$有如下性质：

(1) $\hat{\beta}_0$与$\hat{\beta}_1$是$y_i(i=1,2,\cdots,n)$的线性函数；

(2) $\hat{\beta}_0$与$\hat{\beta}_1$是β_0与β_1的无偏估计；

(3) $\mathrm{var}(\hat{\beta}_1)=\frac{\sigma^2}{l_{xx}}$，$var(\hat{\beta}_0)=\left[\frac{1}{n}+\frac{(\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]\sigma^2=\frac{\sum_{i=1}^{n}x_i^2}{nl_{xx}}\sigma^2$；

(4) $\mathrm{cov}(\bar{y},\hat{\beta}_1)=0$，$\mathrm{cov}(\hat{\beta}_0,\hat{\beta}_1)=-\frac{\bar{x}}{l_{xx}}\sigma^2$。

定理 8.2 对于一元正态线性回归的样本模型(8.5)

$$\begin{cases}y_i=\beta_0+\beta_1x_i+\varepsilon_i\quad(i=1,2,\cdots,n)\\ \varepsilon_i\text{相互独立},\ \varepsilon_i\sim N(0,\sigma^2)\end{cases}$$

β_0与β_1的最小二乘估计$\hat{\beta}_0$与$\hat{\beta}_1$有如下性质：

(1) $y_i\sim N(\beta_0+\beta_1x_i,\sigma^2)$；

(2) $\hat{\beta}_0\sim N\left(\beta_0,\frac{\sum_{i=1}^{n}x_i^2}{nl_{xx}}\sigma^2\right)$；

(3) $\hat{\beta}_1\sim N\left(\beta_1,\frac{\sigma^2}{l_{xx}}\right)$；

(4) $\hat{y}\sim N\left(\beta_0+\beta_1x,\left[\frac{1}{n}+\frac{(x-\bar{x})^2}{l_{xx}}\right]\sigma^2\right)$。

为了便于后面的讨论，引入下面的概念与记号。

称$\hat{\varepsilon}_i=y_i-\hat{y}_i$为残差。

称$SSE=(\hat{\varepsilon}_i)^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\hat{\beta}_0\text{-}\hat{\beta}_1x_i)^2$为残差平方和。

$$\hat{\sigma}^2=\frac{SSE}{n-2}=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2. \tag{8.16}$$

定理 8.3 对于一元线性回归的样本模型(8.4)

$$\begin{cases} y_i=\beta_0+\beta_1x_i+\varepsilon_i \quad (i=1,2,\cdots,n) \\ \varepsilon_i\text{相互独立},\ E(\varepsilon_i)=0,\ \mathrm{var}(\varepsilon_i)=\sigma^2 \end{cases}$$

则$\hat{\sigma}^2$是σ^2的无偏估计。

定理 8.4 对于一元正态线性回归的样本模型(8.5)

$$\begin{cases} y_i=\beta_0+\beta_1x_i+\varepsilon_i \quad (i=1,2,\cdots,n) \\ \varepsilon_i\text{相互独立},\ \varepsilon_i\sim N(0,\sigma^2) \end{cases}$$

有

(1) $\frac{SSE}{\sigma^2}=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sigma^2}=\frac{(n-2)\hat{\sigma}^2}{\sigma^2}\sim\chi^2(n-2)$;

(2) $\frac{SSE}{\sigma^2}=\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$与$\hat{\beta}_1$相互独立。

下面的内容如果没有特殊说明都是在一元正态线性回归的样本模型(8.5)中进行。

8.1.3 线性假设的显著性检验

当我们得到了一个实际问题的回归方程$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x$后，我们需要对回归方程进行检验。需要检验的假设为

$$\mathrm{H}_0:\beta_1=0,\ \mathrm{H}_1:\beta_1\neq0$$

1、t检验法

t检验法是统计推断中一种常用的检验方法。在回归分析中，t检验法用于检验回归分析中回归系数β_1的显著性。

当$\mathrm{H}_0:\beta_1=0$为真时，由定理(8.2)知

$$\hat{\beta}_1\sim N\left(0,\frac{\sigma^2}{l_{xx}}\right)$$

由定理(8.4)知

$$\frac{(n-2)\,\hat{\sigma}^2}{\sigma^2}\sim\chi^2(n-2)$$

且$\frac{SSE}{\sigma^2}=\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$与$\hat{\beta}_1$相互独立，于是构造$t$统计量

$$t=\frac{\hat{\beta}_1\sqrt{l_{xx}}}{\hat{\sigma}}=\frac{\hat{\beta}_1\sqrt{l_{xx}}\Big/\sigma}{\sqrt{(n-2)\,\hat{\sigma}^2/[(n-2)\,\sigma^2]}}\sim t\,(n-2) \tag{8.17}$$

式中$\hat{\sigma}^2=\frac{1}{n-2}\sum\limits_{i=1}^{n}(y_i-\hat{y}_i)^2$。给定显著性水平$\alpha$，拒绝域为

$$\left\{|t|\geq t_{\alpha/2}(n-2)\right\} \tag{8.18}$$

2、F检验法

回归分析中回归系数β_1的显著性的另一种检验是F检验法。F检验法是根据平方和分解式，直接从回归效果来检验回归方程的显著性。

观察值 $y_1,y_2,\cdots,y_n$ 之间存在差异，这些差异来自两个方面：自变量x 的取值不同和随机误差的影响。如果观察值 $y_1,y_2,\cdots,y_n$ 之间的差异主要来自于自变量x的取值不同，则应有$\beta_1\neq 0$；相反若观察值 $y_1,y_2,\cdots,y_n$ 之间的差异主要来自于随机误差的影响，则应有$\beta_1=0$。

平方和分解式

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

其中

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 \tag{8.19}$$

称为总离差平方和或总偏差平方和，简记为SST或$\mathrm{S_T}$或$\mathrm{S_{总}}$；

$$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \tag{8.20}$$

称为回归平方和，简记为SSR或$\mathrm{S_{回}}$；

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{8.21}$$

称为残差平方和或剩余平方和，简记为SSE或$\mathrm{S_{残}}$或$\mathrm{S_{剩}}$。

定理 8.5 对于一元正态线性回归的样本模型(8.5)

$$\begin{cases} y_i=\beta_0+\beta_1x_i+\varepsilon_i \quad (i=1,2,\cdots,n) \\ \varepsilon_i\text{相互独立},\ \varepsilon_i\sim N\,(0,\sigma^2) \end{cases}$$

当$\mathrm{H_0}:\beta_1=0$为真时，有

(1) $SSR/\sigma^2 \sim \chi^2(1)$;

(2) SSR与SSE独立;

(3) $F = \frac{SSR}{SSE/(n-2)} \sim F(1, n-2)$。

由定理8.5，当$H_0 : \beta_1 = 0$ 为真时，构造统计量

$$F = \frac{SSR}{SSE/(n-2)} \tag{8.22}$$

给定显著性水平α，拒绝域为

$$\{F \geq F_\alpha(1, n-2)\} \tag{8.23}$$

3、相关系数的显著性检验

称

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{l_{xy}}{\sqrt{l_{xx} l_{yy}}} = \hat{\beta}_1 \sqrt{\frac{l_{xx}}{l_{yy}}} \tag{8.24}$$

为x与y的简单相关系数，简称相关系数， 其中

$$l_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{8.25}$$

可以验证相关系数的取值范围为$|r| \leq 1$。

当$H_0 : \beta_1 = 0$ 为真时，给定显著性水平α，拒绝域为

$$\{|r| \geq r_\alpha(n-2)\} \tag{8.26}$$

其中，$r_\alpha(n-2)$是检验的临界值，可在附表中查找。

对于以上三种检验法，可以验证

$$F = t^2,\ t = \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}},\ \frac{1}{r^2} = 1 + \frac{n-2}{F}$$

三种检验法实际上是等价的，但导出导出种检验法的统计思想是不同的。 其中导出F 检验法的统计思想是方差分析，而导出t检验法的统计思想是正态总体的假设检验。

8.1.4 回归系数β_1的区间估计

当线性回归效果显著，即$H_0 : \beta_1 = 0$ 时，需要对回归系数β_1作区间估计。 可以取枢轴量

$$t = \frac{\left(\hat{\beta}_1 - \beta_1\right)\sqrt{l_{xx}}}{\hat{\sigma}} \tag{8.27}$$

这里

$$\hat{\sigma}^2=\frac{SSE}{n-2}=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

因为

$$\hat{\beta}_1\sim N(\beta_1,\frac{\sigma^2}{l_{xx}})$$

从而

$$\left(\hat{\beta}_1-\beta_1\right)\sqrt{l_{xx}}\Big/\sigma\sim N(0,1)$$

又由于

$$\frac{(n-2)\,\hat{\sigma}^2}{\sigma^2}\sim\chi^2(n-2)$$

所以

$$t=\frac{\left(\hat{\beta}_1-\beta_1\right)\sqrt{l_{xx}}}{\hat{\sigma}}=\frac{\left(\hat{\beta}_1-\beta_1\right)\sqrt{l_{xx}}\Big/\sigma}{\sqrt{(n-2)\,\hat{\sigma}^2/[(n-2)\,\sigma^2]}}\sim t(n-2)$$

于是

$$P\left\{|t|<t_{\alpha/2}(n-2)\right\}=1-\alpha$$

即

$$P\left\{\hat{\beta}_1-t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{l_{xx}}}<t<\hat{\beta}_1+t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{l_{xx}}}\right\}=1-\alpha$$

利用正态总体置信区间的知识，可以得到β_1的置信水平为$1-\alpha$的置信区间为

$$\left(\hat{\beta}_1-t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{l_{xx}}},\ \hat{\beta}_1+t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{l_{xx}}}\right)\tag{8.28}$$

可以简写为

$$\left(\hat{\beta}_1\pm t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{l_{xx}}}\right).$$

8.1.5 因变量的预测

建立回归方程的目的是应用，而预测是回归方程的最重要的应用。

给定$x=x_0$，线性模型为

$$y_0=a+bx_0+\varepsilon_0,\ \varepsilon_0\sim N(0,\sigma^2)$$

1、单值预测

当给定$x=x_0$时，利用$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x$，取

$$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_0\tag{8.29}$$

作为预测值。 这里$E(\hat{y}_0)=E(y_0)=\beta_0+\beta_1x_0$。

2、区间预测

由定理8.2，知

$$\hat{y}_0 \sim N\left(\beta_0+\beta_1 x_0, \left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right)\sigma^2\right)$$

进而

$$y_0-\hat{y}_0 \sim N\left(0, \left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right)\sigma^2\right)$$

于是

$$\frac{y_0-\hat{y}_0}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}} \sim N(0,1)$$

又由定理8.4，知

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

所以

$$t=\frac{y_0-\hat{y}_0}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}}=\frac{\frac{y_0-\hat{y}_0}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}}}{\sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2(n-2)}}} \sim t(n-2)$$

于是

$$P\left\{|t|<t_{\alpha/2}(n-2)\right\}=1-\alpha$$

即

$$P\left\{\hat{y}_0-t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}\right.$$
$$\left.<y_0<\hat{y}_0+t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}\right\}=1-\alpha.$$

利用正态总体置信区间的知识，可以得到y_0的置信水平为$1-\alpha$的预测区间为

$$\left(\hat{y}_0-t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}\hat{y}_0+t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}\right) \quad (8.30)$$

可以简写为

$$\left(\hat{y}_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}\right)$$

例 8.1 保险公司希望确定居民住宅区火灾造成的损失数额与该住宅区到消防站的最近距离之间的相关关系，以便准确地确定保险金额。经调查得15起火灾事故的损失及火灾发生地与消防站的最近距离，数据如下表。

距消防站距离x(km)	3.4	1.8	4.6	2.3	3.1	5.5	0.7	3.0
火灾损失y(千元)	26.2	17.8	31.3	23.1	27.5	36.0	14.1	22.3
距消防站距离x(km)	2.6	4.3	2.1	1.1	6.1	4.8	3.8	
火灾损失y(千元)	19.6	31.3	24.0	17.3	43.2	36.4	26.1	

试进行线性回归分析：(1)求一元线性回归方程; (2)σ^2的无偏估计值$\hat{\sigma}^2$; (3) 相关系数的显著性检验; (4)回归系数β_1的置信水平95%的置信区间; (5)对每一个y求置信水平为95%的预测区间。

解：因为 $n=15$，$\bar{x}=3.28$，$\bar{y}=26.413$，

$$l_{xx}=\sum_{i=1}^{n}x_i^2-n\left(\bar{x}\right)^2=34.784,$$

$$l_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})y_i=171.114,$$

$$\hat{\beta}_1=l_{xy}/l_{xx}=4.919,$$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=10.278$$

于是， 所求的回归方程为

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x=10.278+4.919x$$

σ^2的无偏估计值为

$$\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{\beta}_0+\hat{\beta}_1x_i)^2=5.366$$

由于

$$r=\frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}=0.961>r_\alpha(n-2)=r_{0.05}(13)=0.514$$

因此， 在显著性水平为0.05条件下，认为火灾损失对消防站距离的一元线性回归效果显著。

参数β_1的置信水平为95%的置信区间为

$$\left(\hat{\beta}_1-t_{\alpha/2}\left(n-2\right)\frac{\hat{\sigma}}{\sqrt{l_{xx}}},\ \hat{\beta}_1+t_{\alpha/2}\left(n-2\right)\frac{\hat{\sigma}}{\sqrt{l_{xx}}}\right)=(4.071,\ 5.768)$$

给定$x_0=x_i$，$i=1,2,\cdots,15$， 根据

$$\left(\hat{y}_0\pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}\right)$$

对应的y_0的95%的预测区间为

序号	y_0	y_0的下限	y_0的上限	序号	y_0	y_0的下限	y_0的上限
1	26.2	25.94	28.07	9	19.6	21.91	24.23
2	17.8	17.66	20.61	10	31.3	30.16	32.71
3	31.3	31.51	34.31	11	24.0	19.27	21.95
4	23.1	20.33	22.85	12	17.3	13.84	17.54
5	27.5	24.46	26.59	13	43.2	38.06	42.51
6	36.0	35.46	39.21	14	36.4	32.39	35.39
7	14.1	11.64	15.81	15	26.1	27.85	30.09
8	22.3	23.96	26.11				

8.1.6 可线性化的一元非线性回归

在许多实际问题中，变量之间的关系并不是线性的，而是非线性关系。对于这些非线性关系我们不能照搬之前的线性回归方程来处理。但是某些非线性的关系可以通过变量代换转化为线性关系，从而可以用线性回归方程的方式来加以解决。

例如，y和x之间的关系为

$$y = \alpha x^{\beta} e^{\varepsilon}, \varepsilon \sim N(0, \sigma^2) \tag{8.31}$$

其中α, β, σ^2为与x无关的未知参数。将上式两边取对数，得

$$\ln y = \ln \alpha + \beta \ln x + \varepsilon$$

令$U = \ln Y$，$a = \ln \alpha$，$b = \beta$，$v = \ln x$，则$y = \alpha x^{\beta} \mathrm{e}^{\varepsilon}$化为一元回归模型

$$U = a + bv + \varepsilon, \varepsilon \sim N(0, \sigma^2) \tag{8.32}$$

下面是几种常见的可线性化的非线性函数，我们先讲它们线性化，再用最小二乘法求最小二乘估计，进而求出回归方程。

1、双曲线 $y = \frac{x}{ax+b}$

首先进行变量代换，令$x' = 1/x, y' = 1/y$，则有线性关系$y' = a + bx'$。令

$$x_i' = \frac{1}{x_i}, \quad y_i' = \frac{1}{y_i}, \quad i = 1, ..., n$$

$$l_{x'y'} = \sum_i (x_i' - \overline{x'}) y_i', l_{x'x'} = \sum_{i=1}^{n} (x_i')^2 - n \left(\overline{x'}\right)^2$$

利用最小二乘法，可以得到最小二乘估计

$$\begin{cases} \hat{b} = \frac{l_{x'y'}}{l_{x'x'}} \\ \hat{a} = \overline{y'} - \hat{b}\overline{x'} \end{cases}$$

于是回归方程为$\hat{y}=\frac{x}{\hat{a}x+\hat{b}}$

2、幂函数曲线$y=ax^b$

首先进行变量代换，令$x'=\ln x, y'=\ln y$， 则有线性关系 $y'=\ln a+bx'$。 令

$$x_i'=\ln x_i,\quad y_i'=\ln y_i,\quad i=1,...,n$$

$$l_{x'y'}=\sum_i (x_i'-\overline{x'})y_i', l_{x'x'}=\sum_{i=1}^{n}(x_i')^2-n\left(\overline{x'}\right)^2$$

利用最小二乘法，可以得到最小二乘估计

$$\begin{cases}\hat{b}=\frac{l_{x'y'}}{l_{x'x'}}\\ \hat{a}=\overline{y'}-\hat{b}\overline{x'}\end{cases}$$

于是回归方程为$\hat{y}=\hat{a}x^{\hat{b}}$。

3、对数曲线 $y=a+b\ln x$

首先进行变量代换，令$x'=\ln x, y'=y$， 则有线性关系 $y'=a+bx'$。 令

$$x_i'=\ln x_i,\quad y_i'=y_i,\quad i=1,...,n$$

$$l_{x'y'}=\sum_i (x_i'-\overline{x'})y_i', l_{x'x'}=\sum_{i=1}^{n}(x_i')^2-n\left(\overline{x'}\right)^2$$

利用最小二乘法，可以得到最小二乘估计

$$\begin{cases}\hat{b}=\frac{l_{x'y'}}{l_{x'x'}}\\ \hat{a}=\overline{y'}-\hat{b}\overline{x'}\end{cases}\tag{8.33}$$

于是回归方程为

$$\hat{y}=\hat{a}+\hat{b}\ln x$$

4、S型曲线

$$y=\frac{1}{a+b\mathrm{e}^{-x}}\,(a>0,b>0)$$

首先进行变量代换，令

$$x'=e^{-x}, y'=1/y$$

则有$y'=a+bx'$。 令

$$x_i'=e^{-x_i},\quad y_i'=1/y_i,\quad i=1,...,n$$

$$l_{x'y'}=\sum_i (x_i'-\overline{x'})y_i', l_{x'x'}=\sum_{i=1}^{n}(x_i')^2-n\left(\overline{x'}\right)^2$$

利用最小二乘法，可以得到最小二乘估计

$$\begin{cases} \hat{b} = \frac{l_{x'y'}}{l_{x'x'}} \\ \hat{a} = \overline{y'} - \hat{b}\overline{x'} \end{cases}$$

于是回归方程为

$$\hat{y} = \frac{1}{\hat{a} + \hat{b}\mathrm{e}^{-x}}$$

例 8.2 某工业产品单位时间产量y和催化剂使用量x有关系，现在做了7次试验，采集数据如下，

催化剂使用量 x	e	e^5	e^7	e^9	e^{12}	e^{16}	e^{18}
单位时间产量 y	6	20	32	41	49	63	77

设 x 和 y 存在对数曲线关系 $y = a + b\ln x$，试求y关于x的非线性回归方程，并给出催化剂使用量为14时单位时间产量的预测值。

解：首先进行变量代换，令$x' = \ln x, y' = y$, 于是根据原数据表转化成新的数据表

催化剂使用量 x'	1	5	7	9	12	16	18
单位时间产量 y'	6	20	32	41	49	63	77

因为 $n = 7$，$\overline{x'} = 9.7143, \overline{y'} = 41.143$，$\sum\limits_{i=1}^{n} (x_i')_i^2 = 880$，

$$l_{x'x'} = \sum_{i=1}^{n} (x_i')^2 - n\left(\overline{x'}\right)^2 = 219.429,$$

$$l_{x'y'} = \sum_{i} (x_i' - \overline{x'})y_i' = 883.286,$$

$$\hat{b} = \frac{l_{x'y'}}{l_{x'x'}} = 4.025,\ \hat{a} = \overline{y'} - \hat{b}\overline{x'} = 2.039,$$

因此回归方程为：$\hat{y} = \hat{a} + \hat{b}\ln x = 2.039 + 4.025\ln x$。 当$x = 14$时，$\hat{y} = 12.662$。

8.2 多元线性回归

8.2.1 多元线性回归模型

在实际问题中，影响因变量的自变量往往不止一个。 这是就需要考虑含多个自变量的的回归分析问题。

设有p个自变量$x_1, x_2, \cdots, x_p$ $(p>1)$ 对因变量y有影响。对于自变量 $x_1, x_2, \cdots, x_p$ 的每一组确定值，都有一个随机变量 y 与之相对应。因为y 是随机变量，所以在不同的试验中的 y 观察值是不同的。如果当 $x_1, x_2, \cdots, x_p$ 的值确定时，y与之对应的条件数学期望 $E(y|x_1, x_2, \cdots, x_p)$ 存在，那么 $E(y|x_1, x_2, \cdots, x_p)$ 称为 y 关于 $x_1, x_2, \cdots, x_p$ 的回归函数。于是自变量与因变量之间的关系可以用如下的模型来描述

$$\begin{cases} y = E(y|x_1, x_2, \cdots, x_p) + \varepsilon \\ E(\varepsilon) = 0 \\ \mathrm{var}(\varepsilon) = \sigma^2 \end{cases} \tag{8.34}$$

这里ε是不可观测的随机误差，它是一个随机变量。随机误差 ε 应理解为除 $x_1, x_2, \cdots, x_p$ 外对 y 有影响但未加考虑的诸多因素(包括随机因素)所产生的影响总和。

(8.34)称为多元回归模型，基于(8.34)的的统计分析称为多元回归分析。

如果回归函数$E(y|x_1, x_2, \cdots, x_p)$是$x_1, x_2, \cdots, x_p$的线性函数，即

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

则模型(8.34)可化为

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \\ E(\varepsilon) = 0 \\ \mathrm{var}(\varepsilon) = \sigma^2 \end{cases} \tag{8.35}$$

其中$\beta_0, \beta_1, \cdots, \beta_p$是$p+1$个未知参数。称$\beta_0$为回归常数，称$\beta_1, \cdots, \beta_p$称为回归系数。(8.35) 称为多元线性回归模型，基于(8.35)的的统计分析称为多元线性回归分析。若进一步要求

$$\varepsilon \sim N\left(0, \sigma^2\right)$$

即

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \\ \varepsilon \sim N\left(0, \sigma^2\right) \end{cases} \tag{8.36}$$

则称为多元正态线性回归模型。

如果回归函数$E(y|x_1, x_2, \cdots, x_p)$不是$x_1, x_2, \cdots, x_p$的线性函数，则对应的(8.34)称为多元非线性回归模型，此时基于(8.34) 的的统计分析称为多元非线性回归分析。

对$(x_1, x_2, \cdots, x_p, y)$进行观察，得到$n$ 个观察值

$$\left(x_{i1}, ..., x_{ip}, y_i\right), i = 1, ..., n$$

如果符合模型式(8.35)，则

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (i = 1, 2, \cdots, n) \\ \varepsilon_i \text{相互独立}, \ E(\varepsilon_i) = 0, \ \mathrm{var}(\varepsilon_i) = \sigma^2 \end{cases} \tag{8.37}$$

(8.37)称为多元线性回归的样本模型。

显然

$$E(y_i)=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip},\ \mathrm{var}(y_i)=\sigma^2$$

如果符合模型式(8.36)，则

$$\begin{cases} y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}+\varepsilon_i \quad (i=1,2,\cdots,n) \\ \varepsilon_i\text{相互独立},\ \varepsilon_i\sim N(0,\sigma^2) \end{cases} \tag{8.38}$$

(8.38)称为多元正态线性回归的样本模型。

在上述两个样本模型中

$$\begin{cases} y_1=\beta_0+\beta_1x_{11}+\beta_2x_{12}+\cdots+\beta_px_{1p}+\varepsilon_1 \\ y_2=\beta_0+\beta_1x_{21}+\beta_2x_{22}+\cdots+\beta_px_{2p}+\varepsilon_2 \\ \cdots\cdots \\ y_n=\beta_0+\beta_1x_{n1}+\beta_2x_{n2}+\cdots+\beta_px_{np}+\varepsilon_n \end{cases} \tag{8.39}$$

即

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}=\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}+\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{8.40}$$

写成矩阵形式就是

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \tag{8.41}$$

这里

$$\boldsymbol{Y}=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\ \boldsymbol{X}=\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},\ \boldsymbol{\beta}=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},\ \boldsymbol{\varepsilon}=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

在多元正态线性回归的样本模型中

$$\boldsymbol{\varepsilon}\sim\boldsymbol{N}(0,\mathrm{I_n}\boldsymbol{\sigma}^2),\ \boldsymbol{Y}\sim\boldsymbol{N}(\boldsymbol{X}\boldsymbol{\beta},\mathrm{I_n}\boldsymbol{\sigma}^2). \tag{8.42}$$

说明: 样本模型中的y_i是随机变量，$(x_{i1},...,x_{ip},y_i)$ 中的 y_i 是随机变量y_i的观察值。

8.2.2 参数的最小二乘估计

本节考虑多元线性回归，其模型为

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon$$

多元线性回归分析的主要任务就是用适当的统计方法获得未知参数$\beta_0,\beta_1,\cdots,\beta_p$的估计值$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_p$。称

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_px_p \tag{8.43}$$

为多元线性经验回归方程或回归方程。

称

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_{i1}+\hat{\beta}_2x_{i2}+\cdots+\hat{\beta}_px_{ip} \tag{8.44}$$

为$y_i(i=1,2,\cdots,n)$的回归拟合值，简称回归值或拟合值。

记

$$Q(\beta_0,\beta_1,\cdots,\beta_p)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip})^2 \tag{8.45}$$

所谓的最小二乘法，就是寻找参数$\beta_0,\beta_1,\cdots,\beta_p$的估计值$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_p$，使上式达到最小，即寻找$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_p$，满足

$$\begin{aligned}&Q\left(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_p\right)=\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_px_{ip}\right)^2\\&=\min_{\beta_0,\beta_1,\cdots,\beta_p}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip})^2\end{aligned} \tag{8.46}$$

由上式求出的$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_p$称为参数$\beta_0,\beta_1,\cdots,\beta_p$的最小二乘估计。

由微积分学中的极值原理，令

$$\left.\frac{\partial Q}{\partial\beta_i}\right|_{\beta_i=\hat{\beta}_i}=0,\quad i=0,1,\cdots,p$$

即要解下面的方程组

$$\begin{cases}\sum\limits_{i=1}^{n}(y_i-\beta_0-\sum\limits_{j=1}^{p}\beta_jx_{ij})=0\\\sum\limits_{i=1}^{n}(y_i-\beta_0-\sum\limits_{j=1}^{p}\beta_jx_{ij})x_{ik}=0\quad k=1,...,p\end{cases} \tag{8.47}$$

经整理后，得到正规方程组

$$\begin{cases}n\beta_0+\beta_1\sum\limits_{i=1}^{n}x_{i1}+\beta_2\sum\limits_{i=1}^{n}x_{i2}+...+\beta_p\sum\limits_{i=1}^{n}x_{ip}=\sum\limits_{i=1}^{n}y_i\\\beta_0\sum\limits_{i=1}^{n}x_{i1}+\beta_1\sum\limits_{i=1}^{n}x_{i1}^2+\beta_2\sum\limits_{i=1}^{n}x_{i1}x_{i2}+...+\beta_p\sum\limits_{i=1}^{n}x_{i1}x_{ip}=\sum\limits_{i=1}^{n}x_{i1}y_i\\\vdots\\\beta_0\sum\limits_{i=1}^{n}x_{ip}+\beta_1\sum\limits_{i=1}^{n}x_{ip}x_{i1}+\beta_2\sum\limits_{i=1}^{n}x_{ip}x_{i2}+...+\beta_p\sum\limits_{i=1}^{n}x_{ip}^2=\sum\limits_{i=1}^{n}x_{ip}y_i\end{cases} \tag{8.48}$$

写成矩阵形式为

$$\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)\boldsymbol{\beta}=\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y} \tag{8.49}$$

当$\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$可逆时，方程有唯一解

$$\hat{\boldsymbol{\beta}}=\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y} \tag{8.50}$$

式中$\hat{\boldsymbol{\beta}}$是$\boldsymbol{\beta}$的最小二乘估计，这里

$$\hat{\beta}=\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\\ \vdots\\ \hat{\beta}_p\end{pmatrix}$$

定理 8.6 对于多元线性回归的样本模型(8.37)

$$\begin{cases} y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}+\varepsilon_i \quad (i=1,2,\cdots,n)\\ \varepsilon_i\text{相互独立},\ E(\varepsilon_i)=0,\mathrm{var}(\varepsilon_i)=\sigma^2 \end{cases}$$

$\boldsymbol{\beta}$的最小二乘估计$\hat{\boldsymbol{\beta}}$有如下性质：

(1) $\hat{\boldsymbol{\beta}}$是随机向量$\boldsymbol{Y}$的一个线性变换;

(2) $\hat{\boldsymbol{\beta}}$是$\boldsymbol{\beta}$的无偏估计;

(3) $D\left(\hat{\boldsymbol{\beta}}\right)=\boldsymbol{\sigma}^2\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}$。

为了便于后面的讨论，引入下面的概念与记号：

称 $\hat{\varepsilon}_i=y_i-\hat{y}_i$ 为残差。 称 $SSE=\left(\hat{\varepsilon}_i\right)^2=\sum\limits_{i=1}^{n}(y_i-\hat{y}_i)^2$ 为残差平方和。 称 $SSR=\sum\limits_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ 为回归平方和。 称 $SST=\sum\limits_{i=1}^{n}(y_i-\bar{y})^2$ 为总偏差平方和。

$$\hat{\sigma}^2=\frac{SSE}{n-p-1}=\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2.$$

定理 8.7 对于多元正态线性回归的样本模型(8.38)

$$\begin{cases} y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}+\varepsilon_i \quad (i=1,2,\cdots,n)\\ \varepsilon_i\text{相互独立},\ \varepsilon_i\sim N\left(0,\sigma^2\right) \end{cases}$$

有如下性质：

(1) $\hat{\boldsymbol{\beta}}\sim\boldsymbol{N}\left(\boldsymbol{\beta},\boldsymbol{\sigma}^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)$;

(2) $\frac{SSE}{\sigma^2}\sim\chi^2\left(n-p-1\right)$;

(3) 当$\beta_1=\beta_2=...=\beta_p=0$时，

$$F=\frac{SSR/p}{SSE/(n-p-1)}\sim F(p,n-p-1)$$

(4) 当$\beta_j=0$时，

$$t_j=\frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}}\sim t(n-p-1)$$

这里c_{jj}是矩阵$\boldsymbol{C}=\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}$的第$j+1$行$j+1$列元素。

下面的内容如果没有特殊说明都是在多元正态线性回归的样本模型中进行。

8.2.3 线性假设的显著性检验

当我们得到了一个实际问题的回归方程 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_px_p$ 后，我们需要对回归方程进行检验。

1、t 检验法

在多元线性回归中，回归方程显著并不意味着每个自变量对因变量的影响偶显著。 如果某个自变量x_j对因变量y的影响不显著，那么在回归模型中，它的系数就取值为零。 因此，检验自变量x_j对y的作用是否显著，等价于检验假设

$$\mathrm{H}_{0j}:\beta_j=0,\ j=1,2,\cdots,p \tag{8.51}$$

根据定理8.7，知

$$t_j=\frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}}\sim t(n-p-1) \tag{8.52}$$

给定显著性水平α，拒绝域为

$$\left\{|t|\geq t_{\alpha/2}(n-p-1)\right\} \tag{8.53}$$

2、F 检验法

F 检验法是根据平方和分解式，直接从回归效果来检验回归方程的显著性。

观察值$y_1,y_2,\cdots,y_n$之间存在差异，这些差异来自两个方面：自变量 $x_1,x_2,\cdots,x_p$ 的取值不同和随机误差的影响。 如果观察值 $y_1,y_2,\cdots,y_n$ 之间的差异主要来自于自变量的取值不同，则应有 $\beta_1,\beta_2,\cdots,\beta_p$ 不全为零; 相反，若观察值$y_1,y_2,\cdots,y_n$之间的差异主要来自于随机误差的影响，则应有$\beta_1=\beta_2=\cdots=\beta_p=0$。 为此提出与原假设

$$\mathrm{H}_0:\beta_1=\beta_2=\cdots=\beta_p=0 \tag{8.54}$$

类似一元正态线性回归检验，我们仍然利用平方和分解式

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

即$SST=SSR+SSE.$

由定理8.7，知

$$F=\frac{SSR/p}{SSE/(n-p-1)}\sim F(p,n-p-1) \tag{8.55}$$

给定显著性水平α，拒绝域为

$$\{F\geq F_\alpha(p,n-p-1)\} \tag{8.56}$$

8.2.4 回归系数β_j的区间估计

当线性回归效果显著，即$H_{0j}:\ \beta_j=0$时，需要对回归系数β_j作区间估计。由定理8.7，知

$$t_j=\frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}}\sim t(n-p-1)$$

于是

$$P\{|t|<t_{\alpha/2}(n-p-1)\}=1-\alpha$$

即

$$P\left\{\hat{\beta}_j-t_{\alpha/2}(n-p-1)\sqrt{c_{jj}}\hat{\sigma}<t<\hat{\beta}_j+t_{\alpha/2}(n-p-1)\sqrt{c_{jj}}\hat{\sigma}\right\}=1-\alpha$$

利用正态总体置信区间的知识，可以得到$\hat{\beta}_j$的置信水平为$1-\alpha$的置信区间为

$$\left(\hat{\beta}_j-t_{\alpha/2}(n-p-1)\sqrt{c_{jj}}\hat{\sigma},\hat{\beta}_j+t_{\alpha/2}(n-p-1)\sqrt{c_{jj}}\hat{\sigma}\right) \tag{8.57}$$

可以简写为

$$\left(\hat{\beta}_j\pm t_{\alpha/2}(n-p-1)\sqrt{c_{jj}}\hat{\sigma}\right)$$

这里

$$\hat{\sigma}^2=\frac{SSE}{n-p-1}=\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2.$$

8.2.5 因变量的预测

建立回归方程的目的是应用，而预测是回归方程的最重要的应用。给定$(x_1,x_2,\cdots,x_p)=(x_{01},x_{02},\cdots,x_{0p})$，线性模型为

$$\begin{cases}y_0=\beta_0+\beta_1x_{01}+\beta_2x_{02}+\cdots+\beta_px_{0p}+\varepsilon_0\\ \varepsilon_0\sim N(0,\sigma^2)\end{cases}$$

单值预测

当给定$(x_1, x_2, \cdots, x_p) = (x_{01}, x_{02}, \cdots, x_{0p})$时，利用

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p,$$

取

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_p x_{0p} \tag{8.58}$$

作为y_0的预测值。

8.3 单因素方差分析

在实际中，影响试验结果的因素有很多。要判断因素对试验结果的影响就要经常面临判断两个以上正态总体的样本均值的差异是来自抽样的随机性，还是源于被抽样的总体均值之间的差异。例如，对同样的商品，采用三种不同的包装是否会导致明显不同的销售量; 三个居民区之间是否收入水平会有差异等等。用来解决上述问题的方法称为方差分析法，是由英国统计学家费歇尔(R.A.Fisher)最早提出。在前面所提及的例子中，包装、居民区称为因素，因素所处的状态称为水平。如三种不同的包装就是包装这一因素的三个水平、三个居民区就是居民区这一因素的三个水平。如果在试验中，只有一个因素取不同的水平，其他因素保持不变，那么这种试验称为单因素试验，如果有两个或两个以上的因素取不同的水平，称为多因素试验，本节讨论单因素试验。

8.3.1 单因素方差分析模型

例 8.3 为检验甲、乙、丙三种不同配料方案所制成水泥的凝固时间是否相同，某实验员试验后得到了下表实验数据。其中甲、乙、丙的配料方案分别用A_1, A_2, A_3表示。

配料方案	凝固时间						$\bar{x}$
A_1	12	10	13	15	12		12.4
A_2	17	16	19	20	13	15	16.7
A_3	15	19	12	14			15

试问这三种配料方案制成的水泥在凝固时间上是否有显著性差异?

解：这里配料方案为因素，三种配料方案为三个水平，记为A_1, A_2, A_3，这就是单因素三水平的试验。从每一种水平A_i的数据可发现，在水平A_i下，凝固时间存在差异，对于每个水平，如果不存在各种不可控制的随机因素，一般认为在水平A_i下凝固时间是常数μ_i。但由于不可控因素是客观存在的，因此在选择配料方案A_i时，凝固时间是随机变量，即认为每个水平A_i对应一个总体X_i，设ε_i是各种随机因素的综合效应，根据中心极限定理，认为$\varepsilon_i \sim N(0, \sigma_i^2)$，$i = 1, 2, 3$，是合理的，这样有$X_i = \mu_i + \varepsilon_i$且$X_i \sim N(\mu_i, \sigma_i^2)$，$i = 1, 2, 3$。另一方面，从不同的水平$A_i$之间的均值来看也存在着差异，样本均值反映了总体的情况，因此样本均值存在差异的事实说明，把三个水平之下的总体X_i看成三个不同的正态总体更为合理。

所谓判断水泥凝固时间是否相同的问题，就是要判断凝固时间的差异是受随机因素的影响，还是受配料方案的不同所致。一般地，在安排单因素试验时，除所要考虑的因素外，其余的条件尽可能做到相同，因此我们认为不同水平下所对应的正态总体的方差应是相等的。因此，要判断几个正态总体是否来自同一分布的问题就是判断几个具有相同方差的正态总体的均值是否相等的问题。

下面考虑单因素方差分析的数学模型。

设因素 A 有 k 个不同的水平 $A_1, A_2, \cdots, A_k$，在水平 A_i 下的总体分布为 $N(\mu_i, \sigma^2)$，从水平 A_i 的总体中抽取一个容量为 n_i 的样本 $X_{i1}, X_{i2}, \cdots, X_{in_i}$ $(i = 1, 2, \cdots, k)$，这k个样本相互独立，我们要检验假设

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k,\ H_1 : \mu_1, \mu_2, \cdots, \mu_k \text{ 不全相等.} \tag{8.59}$$

另设$\varepsilon_{ij} = X_{ij} - \mu_i$，$\varepsilon_{ij} \sim N(0, \sigma^2)$。$\varepsilon_{ij}$所反映的便是不可控因素对试验指标的综合影响。因此，由X_{ij}的数据结构得到单因素方差分析模型

$$\begin{cases} X_{ij} = \mu_i + \varepsilon_{ij} \quad i = 1, 2, \cdots k \quad j = 1, 2, \cdots, n_i \\ \varepsilon_{ij} \text{相互独立},\ \varepsilon_{ij} \sim N(0, \sigma^2) \end{cases} \tag{8.60}$$

当$k = 2$时，检验$H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$，可采用t检验法; 当k >2时，t检验已不再适用。

为了便于后面的讨论，引入下面的概念。记

$$n = \sum_{i=1}^{k} n_i,\ \mu = \frac{1}{n} \sum_{i=1}^{k} n_i \mu_i$$

$$\delta_i = \mu_i - \mu,\ (i = 1, 2, \cdots, k)$$

称μ为总平均或一般平均，δ_i为因素A的第i水平A_i的效应。δ_i表示水平A_i下总体均值μ_i 与总平均μ的差异，δ_i 满足满足如下关系式

$$\sum_{i=1}^{k} n_i \delta_i = 0,\ \delta_i - \delta_j = \mu_i - \mu_j.$$

由以上假定，单因素方差分析模型可改写为

$$\begin{cases} X_{ij}=\mu+\delta_i+\varepsilon_{ij} \quad i=1,2,\cdots k \quad j=1,2,\cdots,n_i \\ \varepsilon_{ij}\text{相互独立, } \varepsilon_{ij}\sim N(0,\sigma^2) \\ \sum\limits_{i=1}^{k} n_i\delta_i=0 \end{cases} \tag{8.61}$$

相应地可将假设检验(8.59)等价地表示为

$$H_0:\delta_1=\delta_2=\cdots=\delta_k=0,\ H_1:\delta_1,\delta_2\cdots,\delta_k \text{ 不全为0} \tag{8.62}$$

8.3.2 单因素方差分析的统计分析

为了更好的进行分析，引入一些概念。

称$\bar{X}_i=\frac{1}{n_i}\sum\limits_{j=1}^{n_i}X_{ij}$为水平$A_i$的均值。

称$\bar{X}=\frac{1}{n}\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_j}X_{ij}=\frac{1}{n}\sum\limits_{i=1}^{k}n_i\bar{X}_i$为总均值。

称$S_T=\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i}\left(X_{ij}-\bar{X}\right)^2$为总离差平方和。

称$S_E=\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^2$为误差平方和或组内平方和。

称$S_A=\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i}\left(\bar{X}_i-\bar{X}\right)^2$为效应平方和或组间平方和。

其中

$$S_T=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[\left(X_{ij}-\bar{X}_i\right)+\left(\bar{X}_i-\bar{X}\right)\right]^2$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{X}_i-\bar{X}\right)^2+2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)\left(\bar{X}_i-\bar{X}\right)$$

注意到

$$2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)\left(\bar{X}_i-\bar{X}\right)=2\sum_{i=1}^{k}\left(\bar{X}_i-\bar{X}\right)\left(\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)\right)$$

$$=2\sum_{i=1}^{k}\left(\bar{X}_i-\bar{X}\right)\left(\sum_{j=1}^{n_i}X_{ij}-n_i\bar{X}_i\right)=0$$

于是我们就有S_T的分解: $S_T=S_E+S_A$。

下面讨论方差分析的基本原理。

如果H_0成立，那么k个总体间无显著性差异，也即认为因素对试验结果影响不显著，可以把X_{ij}看作来自同一正态总体$N(\mu,\sigma^2)$。反之，若H_0不成立，则各X_{ij}的差异除随机波动引起的差异外，还应包括因素的不同水平作用的差异。

S_T反映了全体样本的波动程度，S_E反映了随机误差的大小。而S_A除了反映了随机误差之外，还反映了各个水平之间的差异。

S_A与S_E比较，可考察比值S_A/S_E，若比值不大，则说明水平差异的影响不大，就不应该拒绝H_0。反之，若比值较大，则说明水平差异的影响较大，这时应拒绝H_0。因此，原假设$H_0:\mu_1=\mu_2=\cdots=\mu_k$的拒绝域为$S_A/S_E>c$，$c$为一正常数，其值与指定的检验水平有关。

定理 8.8 *在单因素方差分析模型中，如果H_0成立时，则*

(1) $S_E/\sigma^2\sim\chi^2(n-k)$;

(2) $S_A/\sigma^2\sim\chi^2(k-1)$，*且$S_A$与$S_E$相互独立*;

(3) $F=\frac{S_A/(k-1)}{S_E/(n-k)}\sim F(k-1,n-k)$.

因此，构造统计量

$$F=\frac{S_A/(k-1)}{S_E/(n-k)} \tag{8.63}$$

在显著性水平为α的条件下，检验问题的拒绝域为

$$\left\{F=\frac{S_A/(k-1)}{S_E/(n-k)}\geq F_\alpha(k-1,n-k)\right\} \tag{8.64}$$

通常，可将上述检验过程写成如下的方差分析表的形式

表 8.1: 单因素试验方差分析表

来源	平方和	自由度	均方	F比
因素A	S_A	$k-1$	$\bar{S}_A=\frac{S_A}{k-1}$	$F=\frac{\bar{S}_A}{\bar{S}_E}$
误差	S_E	$n-k$	$\bar{S}_E=\frac{S_E}{n-k}$	
总和	S_T	$n-1$		

表中$\bar{S}_A=\frac{S_A}{k-1}$，$\bar{S}_E=\frac{S_E}{n-k}$分别称为$S_A,S_E$的均方。在实际中，我们按以下公式和记号来计算$S_T,S_A,S_E$:

$$S_T=\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}^2-n\bar{X}^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}^2-\frac{T_{\cdot\cdot}^2}{n} \tag{8.65}$$

$$S_A=\sum_{i=1}^{k}n_i\bar{X}_i^2-n\bar{X}^2=\sum_{i=1}^{k}\frac{T_{i\cdot}^2}{n_i}-\frac{T_{\cdot\cdot}^2}{n} \tag{8.66}$$

$$S_E = S_T - S_A \tag{8.67}$$

其中$T_{i\cdot} = \sum_{j=1}^{n_i} X_{ij}, i = 1, 2, \cdots, k$，$T_{\cdot\cdot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}$。

在例 8.3的检验问题中，$k = 3$，$n_1 = 5, n_2 = 6, n_3 = 4$，$n = 15$，

$$S_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{T_{\cdot\cdot}^2}{n} = 3408 - \frac{222^2}{15} = 122.4$$

$$S_A = \sum_{i=1}^{k}\frac{T_{i\cdot}^2}{n_i} - \frac{T_{\cdot\cdot}^2}{n} = 3335.5 - \frac{222^2}{15} = 49.9$$

$$S_E = S_T - S_A = 72.5$$

S_T, S_A, S_E的自由度依次分别为$n-1$=14，$k-1$=2，$n-k$=12，得

表 8.2: 例 8.3的方差分析表

来源	平方和	自由度	均方	F比
因素	49.9	2	24.95	4.13
误差	72.5	12	6.04	
总和	122.4	14		

因为$F_{0.05}(2, 12) = 3.89$，故在水平0.05的条件下拒绝H_0，即认为不同配料方案对水泥的凝固时间有显著性差异。

例 8.4 现有四组鼠脾的DNA含量数据(mg/g)，试由以下数据分析这四组DNA含量是否存在显著性差异。

组序	DNA含量数据(mg/g)								
A_1	12.3	13.2	13.7	15.2	15.4	15.8	16.9	17.3	
A_2	10.8	11.6	12.3	12.7	13.5	13.5	14.8		
A_3	9.3	10.3	11.1	11.7	11.7	12.0	12.3	12.4	13.6
A_4	9.5	10.3	10.5	10.5	10.5	10.9	11.0	11.5	

解：由题意有 $k = 4$，$n_1 = 8$，$n_2 = 7$，$n_3 = 9$，$n_4 = 8$，$n = 32$，

$$S_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{T_{\cdot\cdot}^2}{n} = 5086.01 - 4952.61 = 133.39,$$

$$S_A = \sum_{i=1}^{k}\frac{T_{i\cdot}^2}{n_i} - \frac{T_{\cdot\cdot}^2}{n} = 5038.469 - 4952.61 = 85.856,$$

$$S_E = S_T - S_A = 47.541,$$

S_T, S_A, S_E 的自由度依次分别为 $n-1$=31，$k-1$=3，$n-k$=28，得方差分析表如下。

表 8.3: 例8.4的方差分析表

来源	平方和	自由度	均方	F比
因素	85.856	3	28.619	16.855
误差	47.541	28	1.698	
总和	133.397	31		

因为$F_{0.05}(3,28) = 2.95$，故在水平0.05的条件下拒绝H_0，认为不同组间的DNA含量是有显著性差异。

8.3.3 未知参数的估计

无论H_0是否为真，$E\left(\frac{S_E}{n-k}\right) = \sigma^2$，故$\hat{\sigma}^2 = \frac{S_E}{n-k}$是$\sigma^2$的无偏估计。因为$E\left(\bar{X}\right) = \mu$，$E\left(\bar{X}_i\right) = \mu_i$，所以$\hat{\mu} = \bar{X}, \hat{\mu}_i = \bar{X}_i$分别是$\mu$ 与μ_i的无偏估计。

下面考虑δ_i的估计问题。若拒绝H_0，则意味着$\delta_1, \delta_2, \cdots, \delta_k$不全为零。由$\delta_i = \mu_i - \mu$可得$\delta_i$的无偏估计为$\hat{\delta}_i = \bar{X}_i - \bar{X}$。另一方面，

$$E\left(\bar{X}_i - \bar{X}_j\right) = (\mu + \delta_i) - (\mu + \delta_j) = \delta_i - \delta_j,\ i \neq j,\ i, j = 1, 2, \cdots, k$$

$$D\left(\bar{X}_i - \bar{X}_j\right) = \sigma^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)$$

所以

$$\bar{X}_i - \bar{X}_j \sim N\left(\delta_i - \delta_j, \sigma^2(\tfrac{1}{n_i} + \tfrac{1}{n_j})\right),\ i \neq j,\ i, j = 1, 2, \cdots, k$$

利用$\frac{\bar{X}_i - \bar{X}_j - (\delta_i - \delta_j)}{\sqrt{\bar{S}_E(1/n_i + 1/n_j)}} \sim t(n-k)$，可得$\delta_i - \delta_j$即$\mu_i - \mu_j$的置信度为$1-\alpha$的置信区间为

$$\left(\bar{X}_i - \bar{X}_j \pm t_{\alpha/2}(n-k)\sqrt{\bar{S}_E(1/n_i + 1/n_j)}\right) \tag{8.68}$$

例 8.5 *求例3中的未知参数$\sigma^2, \mu, \mu_i, \delta_i (i = 1, 2, 3)$的点估计及均值差的置信水平为0.95的置信区间。*

解：(1) 未知参数$\sigma^2, \mu, \mu_i, \delta_i$ $(i = 1, 2, 3)$的点估计为

$\hat{\sigma}^2 = S_E/(n-k) = 72.5/12 = 6.04$，$\hat{\mu} = \bar{x} = 14.8$，$\hat{\mu}_1 = \bar{x}_1 = 12.4$，$\hat{\mu}_2 = \bar{x}_2 = 16.7$，$\hat{\mu}_3 = \bar{x}_3 = 15$，$\hat{\delta}_1 = \bar{x}_1 - \bar{x} = -2.4$，$\hat{\delta}_2 = \bar{x}_2 - \bar{x} = 1.9$，$\hat{\delta}_3 = \bar{x}_3 - \bar{x} = 0.2$。

(2) 均值差的置信区间。 查表得$t_{0.025}(12) = 2.1788$，由(8.68)式得，$\mu_1 - \mu_2$，$\mu_1 - \mu_3$和$\mu_2 - \mu_3$的置信水平为0.95的置信区间分别为

$$\left(12.4 - 16.7 \pm 2.1788\sqrt{6.04\left(\frac{1}{5} + \frac{1}{6}\right)}\right) = (-7.542, 1.058)$$

$$\left(12.4 - 15 \pm 2.1788\sqrt{6.04\left(\frac{1}{5} + \frac{1}{4}\right)}\right) = (-6.193, 0.992)$$

$$\left(16.7 - 15 \pm 2.1788\sqrt{6.04\left(\frac{1}{6} + \frac{1}{4}\right)}\right) = (-1.756, 5.156)$$

8.4 贝叶斯(Bayes)统计分析

贝叶斯(Thomas Bayes，1701—1761)是英国著名的数学家。 贝叶斯统计学派的奠基性工作是他的一篇论文，或者是他自己也感觉到他的学说尚有不完善之处，所以在他死后，此文由他的朋友代为发表。 数学家拉普拉斯(Laplace. P. S)利用贝叶斯在此文中提出的方法，推导出了极其重要的“相继律”。 此后，历经重重困难和波折，经过许多统计学家的不懈努力，譬如意大利的菲纳特(B. de Finetti)、英国的杰弗莱(Jeffreys. H)、瓦尔德(Wald. A)、罗宾斯(Robbins. H)等人，才使得贝叶斯的方法和理论逐渐被人们理解和重视起来。 在今天，贝叶斯思想和方法对数理统计学的发展产生了极其深远的影响，已经取得了和经典统计分析同等重要的学术地位，并且在理论和应用中都获得了极大的发展和广泛的应用。

贝叶斯统计的内容极其丰富，但限于篇幅，本节重点介绍贝叶斯统计的基本思想方法，以及在应用中比较常见的相关问题，诸如先验分布的选取、参数估计和假设检验等等。

8.4.1 贝叶斯统计的基本观点

著名统计学家耐曼(Neyman 1894~1981)指出，在统计推断问题中有三种重要的信息：

(1) 总体信息，即总体分布或所属分布族提供的信息;

(2) 样本信息，即从总体中抽取的样本观测值提供的信息;

(3) 先验信息，即在抽样之前有关统计推断的一些信息，包括相关的历史数据或对事物的经验认识等。

经典统计学解决问题的出发点是样本，它只利用总体信息和样本信息。而贝叶斯统计则认为，在统计推断中不应该忽视对研究对象的经验认知，即先验信息。例如对参数统计模型中的参数θ，其先验信息以先验分布的形式表达。在得到样本观测值$x=(x_1\cdots x_n)^T$之后，由x与先验分布提供的信息，利用贝叶斯公式得到后验分布，于是后验分布融合了样本与先验信息，形成了信息量更丰富的后验信息。以此为基础进行相关的统计分析。大量的理论和实践都表明，贝叶斯统计分析由于利用了先验知识，因而对小样本情况下效果明显优于经典统计分析。下面给出贝叶斯公式的两种常见形式：

事件形式的贝叶斯公式为：

设事件$A_1,\cdots,A_n$构成完备事件组，B为概率大于0的事件，则

$$P(A_k\,|B)=\frac{P(A_k)P(B\,|A_k)}{\sum\limits_{i=1}^{n}P(A_i)P(B\,|A_i)} \tag{8.69}$$

其中$k=1,2,\cdots,n$，这里的序列$\{P(A_i),i=1,\cdots,n\}$则体现了对导致结果B发生的各种原因的经验认知，即先验分布。在事件B发生后，序列$\{P(A_k\,|B),k=1,\cdots,n\}$则体现了对各种原因的一种新的认知，也就是说对$A_1,\cdots,A_n$发生的概率进行了重新的估计，此即为后验分布。由此可看出，贝叶斯公式反映了先验分布向后验分布的转化，是认识的一种深化。

随机变量形式的贝叶斯公式为：

设二维连续型随机变量(X,Y)的联合概率密度函数为$f(x,y)$，条件概率密度函数为$f_{X|Y}(x|y),f_{Y|X}(y|x)$，则

$$f_{X|Y}(x|y)=\frac{f_X(x)f_{Y|X}(y|x)}{\int_{-\infty}^{+\infty}f_X(x)f_{Y|X}(y|x)dx} \tag{8.70}$$

这两个公式就是贝叶斯统计分析的基础。贝叶斯统计的特点是将未知参数θ看做随机变量，其先验分布记为$\pi(\theta)$。然而从贝叶斯统计诞生的时刻起，直至今天，这一点都在受到反对者的种种批评，这些批评主要集中于两点：

(1) 参数θ看成随机变量是否合理？

(2) 如何选取恰当的先验分布？

关于问题(1)，参数θ看成随机变量在某些情况下是合理的，其先验分布也比较容易确定。例如，某厂生产的一大批产品中，次品率为0.01，将这批产品整箱包装，每箱装100个，此时每箱中的次品数θ就是随机变量，其先验分布为$B(100,0.01)$，即一箱中含有θ个次品的概率为

$$C_{100}^{\theta}0.01^{\theta}0.99^{100-\theta}$$

此时，如果从一箱中抽检50个产品，则次品数Y的分布为

$$P(Y=k)=\frac{C_{\theta}^{k}C_{100-\theta}^{50-k}}{C_{100}^{50}}$$

这种情况下，对于问题(1)，(2)，是可以得到满意回答的。但是某些问题将参数看做随机变量并不合理或不容易被人们接受。例如，如果需要估计某射击运动员的命中率θ，此时将θ看做随机变量就不太合理，但如果进一步考虑，当人们对此人毫无认知时，如果认为θ等可能的取$(0,1)$中的数值，也是可以理解的，即可以将其作为区间$(0,1)$中的均匀分布。

需要指出的是，对于问题(2)，却有很多现实的困难，近200多年以来，贝叶斯统计学派的学者付出了很多艰辛的努力，并取得了很大的发展，在实用的范围内，解决了很多先验分布的选取的问题。然而，遗憾的是，时至今日，仍然没有一个完整统一的方法一举解决先验分布的选取问题。

下面给出在统计分析中适用的贝叶斯公式。设未知参数θ是连续型随机变量，先验分布为$\pi(\theta)$，总体X的密度函数为$p(x;\theta)$，样本$X=(X_1,...,X_n)$的联合密度即似然函数表示为$L(\theta|x)=\prod_{i=1}^{n}p(x_i|\theta)$，则已知样本观测值$x=(x_1\cdots x_n)^T$的条件下，$\theta$的条件密度$h(\theta|x)$为

$$h(\theta|x)=\frac{L(\theta|x)\pi(\theta)}{\int\limits_{\Theta}L(\theta|x)\pi(\theta)d\theta}$$

下面通过一个简单的例子说明贝叶斯统计分析的基本观点和处理问题的过程。

例 8.6 设某人朝一目标射击n次，命中r次，讨论此人的命中率θ的估计值。

解：设θ为随机变量，其先验分布为$(0,1)$中的均匀分布，当命中率为θ 时，射击n 次，命中r次的概率为

$$g(r|\theta)=C_n^r\theta^r(1-\theta)^{n-r}$$

则利用贝叶斯公式可得后验分布$f(\theta|r)$为

$$f(\theta|r)=\frac{\pi(\theta)g(r|\theta)}{\int_0^1\pi(\theta)g(r|\theta)d\theta}=\frac{\theta^r(1-\theta)^{n-r}}{\int_0^1\theta^r(1-\theta)^{n-r}d\theta}=\frac{\theta^r(1-\theta)^{n-r}}{B(r+1,n-r+1)}$$

此即为参数θ的β分布。自然可以想到，应用后验分布的数学期望$E(\theta|r)$作为参数θ的估计，即

$$\begin{aligned}\hat{\theta} &= E(\theta|r)\\ &= \frac{1}{B(r+1,n-r+1)}\int_0^1\theta\cdot\theta^r(1-\theta)^{n-r}d\theta=\frac{B(r+2,n-r+1)}{B(r+1,n-r+1)}=\frac{r+1}{n+2}.\end{aligned}$$

一般而言，在经典统计分析中，θ的极大似然估计值为$\hat{\theta}=\frac{r}{n}$，下面和贝叶斯统计的结果做比较：

试验数据	经典统计$\hat{\theta}=\frac{r}{n}$	贝叶斯统计$\hat{\theta}=\frac{r+1}{n+2}$
(1) $n=r=1$	$\hat{\theta}=1$	$\hat{\theta}=\frac{2}{3}$
(2) $n=r=100$	$\hat{\theta}=1$	$\hat{\theta}=\frac{101}{102}$
(3) $n=1,\ r=0$	$\hat{\theta}=0$	$\hat{\theta}=\frac{1}{3}$
(4) $n=100,\ r=0$	$\hat{\theta}=0$	$\hat{\theta}=\frac{1}{102}$

显而易见，此时情形(1)和(3)，由于试验次数人太少，经典统计分析的结论并没有说服力，与此形成鲜明对比的是，贝叶斯统计分析的结果却比较“柔和”，更容易让人信服。容易想到的是，如果对该运动员有了进一步的了解，将θ的先验分布修正为(0. 5，1)的均匀分布，那么贝叶斯统计分析的结果则会发生改变，以对其先验分布做出进一步的修订。此外，在这一分析中，容易看到，不同的先验分布会对计算结果产生比较大的影响，并且后验分布的计算过程也并不容易。

上面例子的分析表明，在贝叶斯统计分析中，先验信息包含在$\pi(\theta)$中，样本信息包含在$L(\theta|x)$中，两者的融合所得到的后验信息，由后验分布$h(\theta|x)$得以体现。如果可以得到恰当的θ的先验分布，则利用后验分布$h(\theta|x)$对θ的推断理所当然比仅利用样本信息的经典统计推断法得到的结论更令人信服。但应该注意到，这种优越性是以高质量的先验分布为前提的，如果先验分布与实际情况偏差太大，这种优势将不复存在。因此，现在统计学界比较统一的看法是，当先验信息比较确信时，建议采用贝叶斯的统计分析方法，否则，当然建议使用经典的统计分析。本节之后的讨论和分析，都是建立在先验分布恰当合理的基础之上，不再一一说明。

为了方便后继的讨论，引进记号“$\propto$”。若随机变量X的密度函数为$p(x)=cg(x)$，其中c是与x无关的数，则记为$p(x)\propto g(x)$，称$g(x)$为$p(x)$的核。例如

正态分布$N(\mu,\sigma^2)$的密度函数$p(x)\propto\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$;

二项分布$B(n,p)$的分布$p(x)\propto\binom{n}{x}\left(\frac{p}{1-p}\right)^x$;

β分布$\beta(a,\sigma)$的密度函数$p(x)\propto x^{a-1}(1-x)^{b-1}$。

符号“$\propto$”可以更清楚的描述公式变化的本质，而不必纠结于常数的变化，如当多个观察值相继得到时，关于参数θ的信息也不断得到更新：

$$h(\theta|x)\propto L(\theta|x)\pi(\theta)\propto\prod_{i=1}^{n}p(x_i|\theta)\pi(\theta)$$

即样本的作用使对θ的认识不断深化，由先验分布转化为后验分布。由此引出下面重要的推断原则。

Bayes统计推断原则：对参数θ所作任何推断必须且只能基于θ的后验分布。

充分统计量是描述了估计量的效率，在经典统计中是一个非常重要的概念，在贝叶斯统计中，也有类似的定义。

定义 8.1 设$T=T(X)$为一统计量，若不论θ的先验分布如何，θ的后验密度$h(\theta\,|x)$总是θ和$T(x)$的函数，则$T=T(x)$称为θ的充分统计量。

定义表明，由样本观测值x提供的有关θ的信息，完全包含在充分统计量$T=T(x)$中。

定理 8.9 (因子分离定理)

$T=T(\boldsymbol{X})$是θ充分统计量的充要条件是存在函数$g(\theta,t)$与非负函数$h(x)$，使得$L(\theta\,|\boldsymbol{x})=g(\theta,T(\boldsymbol{x}))h(\boldsymbol{x})$。

证明：充分性，因为

$$h(\theta\,|x)=\frac{\pi(\theta)f(x\,|\theta)}{\int_\Theta\pi(\theta)f(x\,|\theta)d\theta}=\frac{\pi(\theta)g(\theta,T(x))h(x)}{\int_\Theta\pi(\theta)g(\theta,T(x))h(x)d\theta}=\frac{\pi(\theta)g(\theta,T(x))}{\int_\Theta\pi(\theta)g(\theta,T(x))d\theta},$$

上式右端仅与θ和$T(x)$有关，故$T=T(x)$为θ的充分统计量。

必要性，若$N(0,\theta)$是θ的充分统计量，由定义：$h(\theta\,|x)=g_0(\theta,T(x))$。故

$$h(\theta\,|x)=\frac{\pi(\theta)f(x\,|\theta)}{\int_\Theta\pi(\theta)f(x\,|\theta)d\theta}=g_0(\theta,T(x))$$

记$h(x)=\int_\Theta\pi(\theta)f(x\,|\theta)d\theta$，$g(\theta,T(x))=g_0(\theta,T(x))(\pi(\theta))^{-1}$。则$f(x\,|\theta)=g(\theta,T(x))h(x)$。

8.4.2 先验分布的选取

在贝叶斯统计分析中，先验分布的选取是重大的问题，下面给出几种常用的先验分布选取的方法。

1、贝叶斯假设

当对于未知参数的先验信息“一无所知”时，则可以认为参数θ在其取值范围内是等可能取值的，也就是说θ在服从取值范围内的“均匀分布”，即假定：

$$\pi(\theta)=C\text{或}\pi(\theta)\propto 1\quad \text{当}\theta\in\Theta$$

前面的例1即属于这种情形。但这在参数θ的取值范围是无穷区间时会与传统意义上的概率密度相矛盾，因此需引进广义先验分布的概念。

定义 8.2 若$\pi(\theta)$满足 (i) $\int_\Theta\pi(\theta)\mathrm{d}\theta=\infty$，(ii) $\int_\Theta p(x\,|\theta)\pi(\theta)\mathrm{d}\theta<\infty$，则称为广义先验分布。

需要指出的是，满足条件(i)和(ii)时，虽然$\pi(\theta)$不是通常意义下的概率分布，但其后验概率密度$h(\theta\,|x)$是存在的，因此$h(\theta\,|x)$仍可作为贝叶斯统计推断的依据。从

而，当$\pi(\theta)$为广义均匀分布时，有

$$h(\theta\,|x)\propto 1\cdot L(\theta\,|x)=L(\theta\,|x)$$

即似然函数就是后验密度的核，当θ有充分统计量$T=T(X)$时，记$h(\theta\,|t)$为当$T=t$时，θ的后验密度，由定理8. 1，$L(\theta\,|x)=g(\theta,T(x))h(x)\propto g(\theta,T)\triangleq L(\theta\,|t)$，则上式变为

$$h(\theta|x)\propto L(\theta\,|t)$$

例 8.7 设$X=(X_1,\cdots,X_n)$是来自正态总体$N(\mu,\sigma^2)$的样本，其中σ^2已知，在贝叶斯假设下，求μ的后验密度。

解：因为

$$\begin{aligned}L(\mu\,|x)&=\prod_{i=1}^{n}\tfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\\&=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\tfrac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}\\&=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\tfrac{1}{2\sigma^2}\left[\sum_{i=1}^{n}x_i^2-2\mu\sum_{i=1}^{n}x_i+n\mu^2\right]\right\}\\&=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\tfrac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2\right\}\cdot\exp\left\{\tfrac{-n}{2\sigma^2}\left(2\mu\bar{x}-\mu^2\right)\right\},\end{aligned}$$

由因子分解定理，$\bar{X}$是μ的充分统计量。

设$\pi(\theta)\propto 1$，$\theta\in(-\infty,+\infty)$，则

$$h(\mu\,|\bar{x})\propto L(\mu\,|\bar{x})\propto\exp\left[-\frac{n}{2\sigma^2}(\mu-\bar{x})^2\right]$$

故

$$\mu\,|_{\bar{x}}\sim N\left(\bar{x},\frac{\sigma^2}{n}\right).$$

例 8.8 设$X=(X_1,\cdots,X_n)$是来自正态分布$N(0,\sigma^2)$的样本，在贝叶斯假设下，求σ^2的后验密度。

解：似然函数$L\left(\frac{\sigma^2}{x}\right)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2\right)$。

记$\theta=\sigma^2$，在Bayes假设下

$$h(\theta\,|x)\propto L(\theta\,|x)\propto\left(\frac{1}{\theta}\right)^{\frac{n}{2}}\cdot e^{-\frac{t}{2\theta}}$$

其中$t=\sum_{i=1}^{n}x_i^2$，故$T=\sum_{i=1}^{n}X_i^2$是θ的充分统计量，故后验密度

$$h(\theta\,|x)\propto\left(\frac{1}{\theta}\right)^{\frac{n}{2}}\cdot e^{-\frac{t}{2\theta}}\ (\theta>0)$$

上式右端为逆Γ分布$I\Gamma\left(\frac{n}{2}-1,\frac{t}{2}\right)$的核，故

$$\theta\,|_t \sim /\Gamma\left(\frac{n}{2}-1,\frac{t}{2}\right)$$

其中$t=\sum_{i=1}^{n}x_i^2$。

需要注意的是，在没有任何关于参数θ的信息时，贝叶斯假设是可以理解的，也是一个合理假设。然而，这其中也有一个内在的矛盾，若参数θ的先验分布选用均匀分布，θ的函数也是未知的，也应该选择均匀分布，这当然是矛盾的。此外，这时候的后验分布往往不再服从贝叶斯假设。H. Raiffa, R. Schlaifer所提出的共轭先验分布可以较好的解决上述矛盾。

2、共轭分布法

定义 8.3 *设总体X的分布族为$\{p(x|\theta),\theta\in\Theta\}$，若先验分布$\pi(\theta)$与后验分布$h(\theta\,|x)$属于同一分布族，则称$\pi(\theta)$为$p(x\,|\theta)$的共轭先验分布，该分布族称之为参数$\theta$的共轭先验分布族。*

例 8.9 *设总体X服从0-1分布$B(1,p)$，参数p的先验分布$\pi(p)$选为β分布$\beta(a,b)$，则$\{\beta(a,b),a>0,b>0\}$为0-1分布族的共轭先验分布族。*

证明：因为$h(p\,|x)\propto\pi(\theta)L(\theta\,|x)\propto p^{a-1}(1-p)^{b-1}\cdot p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i}=p^{a+t-1}(1-p)^{n+b-t-1}$，其中$t=\sum_{i=1}^{n}x_i$，右端是$\beta(a+t,n+\sigma-t)$分布的核，故$p\,|_t\sim\beta(a+t,n+b-t)$。

共轭分布法的统计意义在于要求经验知识和样本信息有一定的共性，可以将这两种信息转化为同一类先验知识，也就是以后验分布作为进一步试验的先验分布。因此，利用共轭分布可以将历史上的各次试验信息进行有机融合，为今后的试验结果提供一个合理的前提。此外，这种先验分布还具有计算简单的优点，因此受到人们的欢迎。然而，尽管如此，在选用先验分布时，仍然必须以"准确适当"作为选用的原则，不能一味的追求计算的方便而忽视了准确性。

3、杰弗莱原则

在先验分布选取工作中，杰弗莱(Jeffreys)做出了重大的贡献，他提出了著名的Jeffreys原则，给出了求先验分布的方法，较好地解决θ与其函数$g(\theta)$不同分布的矛盾。杰弗莱原则的内容包括给出了先验分布应该满足的一个合理的要求，并给出了求解合理先验分布的具体方法。

设θ的先验分布为$\pi(\theta)$，对于θ的函数$\eta=g(\theta)$，根据概率论的知识，按同一原则决定的$\eta=g(\theta)$的先验分布是$\pi_g(\eta)$应满足

$$\pi(\theta)=\pi_g[g(\theta)]\,|g'(\theta)| \tag{8.71}$$

而，当$\pi(\theta)$为广义均匀分布时，有

$$h(\theta\,|x)\propto 1\cdot L(\theta\,|x)=L(\theta\,|x)$$

即似然函数就是后验密度的核，当θ有充分统计量$T=T(X)$时，记$h(\theta\,|t)$为当$T=t$时，θ的后验密度，由定理8. 1，$L(\theta\,|x)=g(\theta,T(x))h(x)\propto g(\theta,T)\triangleq L(\theta\,|t)$，则上式变为

$$h(\theta|x)\propto L(\theta\,|t)$$

例 8.7 设$X=(X_1,\cdots,X_n)$是来自正态总体$N(\mu,\sigma^2)$的样本，其中σ^2已知，在贝叶斯假设下，求μ的后验密度。

解：因为

$$\begin{aligned}L(\mu\,|x)&=\prod_{i=1}^{n}\tfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\\&=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\tfrac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}\\&=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\tfrac{1}{2\sigma^2}\left[\sum_{i=1}^{n}x_i^2-2\mu\sum_{i=1}^{n}x_i+n\mu^2\right]\right\}\\&=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\tfrac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2\right\}\cdot\exp\left\{\tfrac{-n}{2\sigma^2}\left(2\mu\bar{x}-\mu^2\right)\right\},\end{aligned}$$

由因子分解定理，$\bar{X}$是μ的充分统计量。

设$\pi(\theta)\propto 1$，$\theta\in(-\infty,+\infty)$，则

$$h(\mu\,|\bar{x})\propto L(\mu\,|\bar{x})\propto\exp\left[-\frac{n}{2\sigma^2}(\mu-\bar{x})^2\right]$$

故

$$\mu\,|_{\bar{x}}\sim N\left(\bar{x},\frac{\sigma^2}{n}\right).$$

例 8.8 设$X=(X_1,\cdots,X_n)$是来自正态分布$N(0,\sigma^2)$的样本，在贝叶斯假设下，求σ^2的后验密度。

解：似然函数$L\left(\frac{\sigma^2}{x}\right)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2\right)$。

记$\theta=\sigma^2$，在Bayes假设下

$$h(\theta\,|x)\propto L(\theta\,|x)\propto\left(\frac{1}{\theta}\right)^{\frac{n}{2}}\cdot e^{-\frac{t}{2\theta}}$$

其中$t=\sum_{i=1}^{n}x_i^2$，故$T=\sum_{i=1}^{n}X_i^2$是θ的充分统计量，故后验密度

$$h(\theta\,|x)\propto\left(\frac{1}{\theta}\right)^{\frac{n}{2}}\cdot e^{-\frac{t}{2\theta}}\ (\theta>0)$$

上式右端为逆Γ分布$I\Gamma\left(\frac{n}{2}-1,\frac{t}{2}\right)$的核，故

$$\theta|_t \sim /\Gamma\left(\frac{n}{2}-1,\frac{t}{2}\right)$$

其中$t=\sum_{i=1}^{n}x_i^2$。

需要注意的是，在没有任何关于参数θ的信息时，贝叶斯假设是可以理解的，也是一个合理假设。然而，这其中也有一个内在的矛盾，若参数θ的先验分布选用均匀分布，θ的函数也是未知的，也应该选择均匀分布，这当然是矛盾的。此外，这时候的后验分布往往不再服从贝叶斯假设。H. Raiffa, R. Schlaifer所提出的共轭先验分布可以较好的解决上述矛盾。

2、共轭分布法

定义 8.3 设总体X的分布族为$\{p(x|\theta),\theta\in\Theta\}$，若先验分布$\pi(\theta)$与后验分布$h(\theta|x)$属于同一分布族，则称$\pi(\theta)$为$p(x|\theta)$的共轭先验分布，该分布族称之为参数$\theta$的共轭先验分布族。

例 8.9 设总体X服从0-1分布$B(1,p)$，参数p的先验分布$\pi(p)$选为β分布$\beta(a,b)$，则$\{\beta(a,b),a>0,b>0\}$为0-1分布族的共轭先验分布族。

证明：因为$h(p|x)\propto\pi(\theta)L(\theta|x)\propto p^{a-1}(1-p)^{b-1}\cdot p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i}=p^{a+t-1}(1-p)^{n+b-t-1}$，其中$t=\sum_{i=1}^{n}x_i$，右端是$\beta(a+t,n+\sigma-t)$分布的核，故$p|_t\sim\beta(a+t,n+b-t)$。

共轭分布法的统计意义在于要求经验知识和样本信息有一定的共性，可以将这两种信息转化为同一类先验知识，也就是以后验分布作为进一步试验的先验分布。因此，利用共轭分布可以将历史上的各次试验信息进行有机融合，为今后的试验结果提供一个合理的前提。此外，这种先验分布还具有计算简单的优点，因此受到人们的欢迎。然而，尽管如此，在选用先验分布时，仍然必须以"准确适当"作为选用的原则，不能一味的追求计算的方便而忽视了准确性。

3、杰弗莱原则

在先验分布选取工作中，杰弗莱(Jeffreys)做出了重大的贡献，他提出了著名的Jeffreys原则，给出了求先验分布的方法，较好地解决θ与其函数$g(\theta)$不同分布的矛盾。杰弗莱原则的内容包括给出了先验分布应该满足的一个合理的要求，并给出了求解合理先验分布的具体方法。

设θ的先验分布为$\pi(\theta)$，对于θ的函数$\eta=g(\theta)$，根据概率论的知识，按同一原则决定的$\eta=g(\theta)$的先验分布是$\pi_g(\eta)$应满足

$$\pi(\theta)=\pi_g[g(\theta)]\,|g'(\theta)| \tag{8.71}$$

如果选出θ符合该条件，则用θ的函数$g(\theta)$导出的先验分布总是一致的，不会互相矛盾。杰弗莱利用Fisher信息量的不变性找到了符合该条件的$\pi(\theta)$。

当$X_1,\cdots,X_n$的联合概率密度为$L(\theta|x)$时，参数θ的Fisher信息量为

$$I(\theta)=E\left(\frac{\partial \ln L(\theta|x)}{\partial\theta}\right)^2=-E\left(\frac{\partial^2 \ln L(\theta|x)}{\partial\theta^2}\right) \tag{8.72}$$

如果$X_i\sim p(x|\theta)$，则$L(\theta|x)=\prod_{i=1}^{n}p(x_i|\theta)$。于是

$$I(\theta)=-E\left(\sum_{i=1}^{n}\frac{\partial^2 \ln p(x_i|\theta)}{\partial\theta^2}\right)=-\sum_{i=1}^{n}E\frac{\partial^2 \ln p(x_i|\theta)}{\partial\theta^2}$$

$$=-nE\left(\frac{\partial^2 \ln p(x_i|\theta)}{\partial\theta^2}\right)=nE\left(\frac{\partial \ln p(x_i|\theta)}{\partial\theta^2}\right)^2$$

即n个独立样本提供的关于参数θ的信息量是一个样品的n倍。

如果参数$L(\theta,a)=C(\theta)|\theta-a|$是多维向量，那么Fisher信息量记为$\theta=(\theta_1,\cdots,\theta_k)^{\mathrm{T}}$，则

$$\left(\frac{\partial \ln L(\theta|x)}{\partial\theta}\right)=\left(\frac{\partial \ln L(\theta|x)}{\partial\theta_1},\cdots,\frac{\partial \ln L(\theta|x)}{\partial\theta_k}\right)^T, \tag{8.73}$$

此时，$I(\theta)=E\left[\left(\frac{\partial \ln L(\theta|x)}{\partial\theta}\right)\left(\frac{\partial \ln L(\theta|x)}{\partial\theta}\right)^T\right]$。

杰弗莱认为θ的先验分布应以信息阵$I(\theta)$的行列式的平方根为核，即

$$\pi(\theta)\propto|I(\theta)|^{\frac{1}{2}} \tag{8.74}$$

由于$I(\theta)$是非负定阵，于是$|I(\theta)|\geq 0$，下面证明(8.72)所确定的先验分布具有(8.71)的不变性。

定理 8.10 设$g(\theta|x)=\frac{C(\theta)h(\theta|x)}{\int_\Theta C(\theta)h(\theta|x)d\theta}$是$\theta$的函数，$\eta=g(\theta)$与$\theta$具有相同的维数$k$，则有$|I(\theta)|^{\frac{1}{2}}=\left|\frac{\partial g(\theta)}{\partial\theta}\right||I(\eta)|^{\frac{1}{2}}$。

证明：记$\ln L=\ln L(\theta|x)$

$$\left(\frac{\partial \ln L}{\partial\theta}\right)=\left(\frac{\partial g(\theta)}{\partial\theta}\right)'\left(\frac{\partial \ln L}{\partial\eta}\right)\to k\times k\text{矩阵}.$$

则

$$\begin{aligned}I(\theta)&=E\left[\left(\frac{\partial \ln L}{\partial\theta}\right)\left(\frac{\partial \ln L}{\partial\theta}\right)'\right]\\&=E\left[\left(\frac{\partial g(\theta)}{\partial\theta}\right)'\cdot\left(\frac{\partial \ln L}{\partial\eta}\right)\left(\frac{\partial \ln L}{\partial\eta}\right)'\left(\frac{\partial g(\theta)}{\partial\theta}\right)\right]\\&=\left(\frac{\partial g(\theta)}{\partial\theta}\right)'E\left[\left(\frac{\partial \ln L}{\partial y}\right)\left(\frac{\partial \ln L}{\partial y}\right)'\right]\left(\frac{\partial g(\theta)}{\partial\theta}\right).\end{aligned}$$

所以$|I(\theta)|^{\frac{1}{2}}=\left|\left(\frac{\partial g}{\partial \theta}\right)' I(\eta)\cdot\left(\frac{\partial g}{\partial \theta}\right)\right|^{\frac{1}{2}}=|I(\eta)|^{\frac{1}{2}}\cdot\left|\frac{\partial g(\theta)}{\partial \theta}\right|$。 证毕。

定理8.10表明杰弗莱原则的合理性。只要取$|I(\theta)|^{\frac{1}{2}}$作为$\pi(\theta)$的核，即$\pi(\theta)\propto|I(\theta)|^{\frac{1}{2}}$，则$\eta=g(\theta)$的分布$\pi(\eta)$满足$\pi(\theta)=\left|\frac{\partial g(\theta)}{\partial \theta}\right|\pi(\eta)$，即(8.71)的不变性成立。

例 8.10 设$X_1,\cdots,X_n\overset{i.i.d}{\sim}N(\mu,\sigma^2)$，求$(\mu,\sigma)$的符合杰弗莱原则的先验分布$\pi(\mu,\sigma)$

解：此时，样本联合密度为：

$$L(\mu,\sigma^2|x_1,\cdots,x_n)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2\right\}$$

$$\ln L=-\frac{1}{2}\ln 2\pi-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2$$

$$\frac{\partial \ln L}{\partial \mu}=\frac{1}{\sigma^2}\sum_{i=1}^n(x_i-\mu)$$

$$\frac{\partial \ln L}{\partial \sigma}=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^n(x_i-\mu)^2$$

于是记$\theta=\begin{pmatrix}\mu\\ \sigma\end{pmatrix}$，由定义

$$I\begin{pmatrix}\mu\\ \sigma\end{pmatrix}=E_\theta\left[\begin{pmatrix}\frac{\partial \ln L}{\partial \mu}\\ \frac{\partial \ln L}{\partial \sigma}\end{pmatrix}\left(\frac{\partial \ln L}{\partial \mu}\ \frac{\partial \ln L}{\partial \sigma}\right)\right]$$

$$I\begin{pmatrix}\mu\\ \sigma\end{pmatrix}=E\begin{bmatrix}\left(\frac{\partial \ln L}{\partial \mu}\right)^2 & \left(\frac{\partial \ln L}{\partial \mu}\ \frac{\partial \ln L}{\partial \sigma}\right)\\ \left(\frac{\partial \ln L}{\partial \sigma}\ \frac{\partial \ln L}{\partial \mu}\right) & \left(\frac{\partial \ln L}{\partial \sigma}\right)^2\end{bmatrix}=\begin{pmatrix}\frac{n}{\sigma^2} & 0\\ 0 & \frac{2n}{\sigma^2}\end{pmatrix}$$

$$|I(\theta)|^{\frac{1}{2}}=\frac{\sqrt{2}n}{\sigma^2}$$

由杰弗莱原则可知，$\pi(\mu,\sigma)\propto\frac{1}{\sigma^2}$，且

$$\mu\in(-\infty,\infty),\ \ \sigma\in(0,\infty)$$

可进一步证明：当σ已知时，$\pi(\mu)\propto 1\ \mu\in(-\infty,\infty)$，当$\mu$已知时，$\pi(\sigma)\propto\frac{1}{\sigma}\ \sigma\in(0,\infty)$。

例 8.11 设$X_1,\cdots,X_n\overset{i.i.d}{\sim}B(1,p)$，$p$未知，根据杰弗莱原则确定$p$的先验分布$\pi(p)$.

解：我们有 $L(p|x)=\prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}=p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$，及

$$\ln L=\sum_{i=1}^n x_i\ln p+\left(n-\sum_{i=1}^n x_i\right)\ln(1-p)$$

$$\frac{\partial \ln L}{\partial p}=\frac{1}{p}\sum_{i=1}^n x_i-\frac{1}{1-p}\left(n-\sum_{i=1}^n x_i\right)=\frac{\sum_{i=1}^n x_i-np}{p(1-p)}$$

$$I(p)=E\left[\left(\frac{\partial \ln L}{\partial p}\right)^2\right]=E\left[\frac{\left(\sum_{i=1}^{n}X_i-np\right)^2}{p^2(1-p)^2}\right].$$

注意到

$$\sum_{i=1}^{n}X_i\sim B(n,p),\ E\left(\sum_{i=1}^{n}X_i\right)=np,$$

$$E\left[\left(\sum_{i=1}^{n}X_i-np\right)^2\right]=D\left(\sum_{i=1}^{n}X_i\right)=np(1-p).$$

所以

$$I(p)=\frac{np(1-p)}{p^2(1-p)^2}=\frac{n}{p(1-p)},$$

$$|I(p)|^{\frac{1}{2}}\propto p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}}.$$

由杰弗莱原则，

$$\pi(p)\propto p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}},\ 0<p<1.$$

即

$$\pi(p)\propto \beta\left(\frac{1}{2},\frac{1}{2}\right).$$

8.4.3 贝叶斯统计中的参数估计

与经典统计类似，贝叶斯参数估计也包括点估计与区间估计两种，但其理论却是以后验分布为基础，这与经典统计有着极大的不同。贝叶斯点估计又分为最大后验估计与条件期望估计。

1、最大后验估计

直观上，似然函数反映了试验结果发生的概率，因此，在一次试验中，随机变量θ取到它的最大可能取值的机会相对也较大。在贝叶斯统计分析中，参数θ的最大可能取值对应使后验分布$h(\theta|x)$达到最大的点，此即为最大后验估计。

定义 8.4 *若$\hat{\theta}=\hat{\theta}(x)$使得$h(\hat{\theta}|x)=\sup_{\theta\in\Theta}h(\theta|x)$，则称$\hat{\theta}$为$\theta$的最大后验估计。*

容易看出，当先验分布采用贝叶斯假设时，即$\pi(\theta)\propto 1$，则$h(\theta|x)\propto L(\theta|x)$，此时最大后验估计即为经典统计分析中的极大似然估计。由于后验密度中融合了未知参数的先验信息和样本信息，因此，基于后验密度得到的最大后验估计应比极大似然估计更加合理。

例 8.12 *设$X_1,\cdots,X_n\overset{i.i.d}{\sim}N(\mu,\sigma^2)$其中$\mu$未知，$\sigma^2$已知。$\mu$的先验分布为$N(\mu_0,\tau^2)$，求$\mu$的最大后验估计。*

解：μ的后验密度为

$$h(\mu|x) \propto L(\mu,\sigma^2|x_1,\cdots,x_n)\pi(\mu) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right\}, \tag{8.75}$$

$$\ln h(\mu|x) \propto -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{(\mu-\mu_0)^2}{2\tau^2}. \tag{8.76}$$

令

$$\frac{\partial \ln h}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) + \frac{1}{\tau^2}(\mu-\mu_0) = 0, \tag{8.77}$$

解得

$$\hat{\mu} = \left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{n}{\sigma^2}\bar{x}+\frac{1}{\tau^2}\mu_0\right). \tag{8.78}$$

记

$$p_1 = \left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{n}{\sigma^2}\right),\ p_2 = \left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{1}{\tau^2}\right), \tag{8.79}$$

则

$$\hat{\mu} = p_1\bar{x} + p_2\mu_0,\ \text{且}\ p_1+p_2=1. \tag{8.80}$$

即 $\hat{\mu}$ 是 $\bar{x}$ 与 μ_0 的加权平均。这一估计值反映了样本信息与先验信息的融合。

2、条件期望估计

在估计参数 $\boldsymbol{\theta}$ 时，使用基于后验概率密度对应的数学期望作为 θ 的估计，这也是一个很自然合理的想法。此即为条件期望估计。

定义 8.5 设θ的后验分布密度为$h(\theta|x)$，则后验分布的期望

$$\hat{\theta} = E(\theta|x) = \int_{\Theta}\theta h(\theta|x)d\theta \tag{8.81}$$

称为θ的 条件期望估计。

条件期望估计是Bayes点估计中最重要的一种，通常Bayes点估计指的就是条件期望估计。

例 8.13 设$X_1,\cdots,X_n$是来自poison分布总体$P(\lambda)$的样本，λ的先验分布为$\Gamma(\alpha,\mu)$，求λ的条件期望估计。

解：λ的后验密度为

$$h(\lambda|x) \propto L(\lambda|x_1,\cdots,x_n)\pi(\lambda) = \frac{\mathrm{e}^{-n\lambda}\lambda^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}\lambda^{\alpha-1}\mathrm{e}^{-\mu\lambda} \propto \lambda^{\alpha+n\bar{x}-1}\mathrm{e}^{-(n+\mu.)\lambda}$$

所以

$$\lambda|_x \sim \Gamma(\alpha + n\bar{x}, n + \mu).$$

于是

$$\hat{\lambda} = E[\lambda|x] = \frac{\alpha + n\bar{x}}{n + \mu} = \frac{\mu}{n+\mu}\frac{\alpha}{\mu} + \frac{n}{n+\mu}\bar{x}.$$

它是样本均值与先验分布$\Gamma(\alpha, \mu)$的均值$\frac{\alpha}{\mu}$ 的加权平均。

3、贝叶斯区间估计

在贝叶斯统计中，由于将未知参数θ看作随机变量，其后验分布为$h(\theta|x)$，此时易于确定θ落在某一区间中的后验概率，因此相对于经典统计，对区间估计的解释是非常自然和易于接受的。 下面对于给定的置信水平$1-\alpha$, 探讨求得贝叶斯统计分析中的最优置信区间的方法。

定义 8.6 已知参数θ的后验密度为$h(\theta|x)$, 对给定的置信概率$1-\alpha$，若存在区间I，满足：

(1) $P\{\theta \in I|x\} = \int_I h(\theta|x)d\theta = 1-\alpha$

(2) 任给$\theta_1 \in I, \theta_2 \notin I$，总有$h(\theta_1|x) \geq h(\theta_2|x)$。

则称I是参数θ的置信度为$1-\alpha$的 最大后验密度(HPD)区间估计， 简称$1-\alpha$ HPD区间。

条件(2)表明θ落在区间I内的概率不比落在区间I外的概率小，即区间I集中了参数θ取到可能性较大的点。

例 8.14 设$X_1, \cdots, X_n$是来自正态总体$N(\mu, \sigma^2)$的i.i.d. 样本, σ^2已知。 μ的先验分布为$N(\mu_0, \tau^2)$。 求μ的$1-\alpha$HPD区间估计。

解：μ的后验密度为

$$\begin{aligned} h(\mu|x) &\propto L(\mu, \sigma^2|x_1, \cdots, x_n)\pi(\mu) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right\} \\ &\propto \exp\left[\frac{-(\mu-\hat{\mu})^2}{2\gamma^2}\right] \end{aligned}$$

其中 $\hat{\mu} = \frac{\left(\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\tau^2}\right)}{\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)}$, $\gamma^2 = \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}$. 给定$1-\alpha$, 因$\frac{\mu-\hat{\mu}}{\gamma} \sim N(0,1)$, 故

$$P\left[\left|\frac{(\mu-\hat{\mu})}{\gamma}\right| \leq z_{\frac{\alpha}{2}}|x\right] = 1-\alpha$$

其中z_α是$N(0,1)$的上$\alpha/2$分位点。 故得μ的$1-\alpha$ HPD区间估计为

$$\left[\hat{\mu} - \gamma z_{\frac{\alpha}{2}}, \hat{\mu} + \gamma z_{\frac{\alpha}{2}}\right].$$

当 $\tau \to \infty$ 时,

$$\hat{\mu} \to \bar{x},\ \gamma^2 \to \frac{\sigma^2}{n},\ \left[\hat{\mu} - \gamma z_{\frac{\alpha}{2}},\ \hat{\mu} + \gamma z_{\frac{\alpha}{2}}\right] \to \left[\bar{x} - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}\right],$$

这与经典区间估计结论一致。

8.4.4 贝叶斯统计中的假设检验

根据贝叶斯统计推断原则，假设检验问题也易于处理。设假设检验问题为

$$H_0 : \theta \in \Theta_0; H_1 : \theta \in \Theta_1 \tag{8.82}$$

记α_0, α_1为下列后验概率:

$$\begin{cases} \alpha_0 = \alpha_0(x) = P\{\theta \in \Theta_0 | x\} = \int\limits_{\Theta_0} h(\theta|x) d\theta \\ \alpha_1 = \alpha_1(x) = P\{\theta \in \Theta_1 | x\} = \int\limits_{\Theta_1} h(\theta|x) d\theta \end{cases} \tag{8.83}$$

贝叶斯假设检验的推断原则是:

(1) 当$\alpha_0(x) \geq \alpha_1(x)$时，接受$H_0$;

(2) 当$\alpha_0(x) < \alpha_1(x)$ 时，拒绝H_0。

称比值$\frac{\alpha_0}{\alpha_1}$为后验概率比，其值为

$$\frac{\alpha_0}{\alpha_1} = \frac{\int_{\Theta_0} L(\theta|x)\pi(\theta)d\theta}{\int_{\Theta_1} L(\theta|x)\pi(\theta)d\theta} \tag{8.84}$$

后验概率比反映了两个假设成立的相对可能性。当$\frac{\alpha_0}{\alpha_1} \geq 1$时，接受$H_0$; 当$\frac{\alpha_0}{\alpha_1} < 1$时，拒绝$H_0$。

定理 8.11 设似然函数为$\{L(\theta|x); \theta \in \Theta, \Theta = \{\theta_0, \theta_1\}\}$, 假设检验问题为

$$H_0 : \theta = \theta_0; H_1 : \theta = \theta_1 \tag{8.85}$$

又设$\pi_0 = p\{\theta = \theta_0\}, \pi_1 = p\{\theta = \theta_1\}$, 则Bayes检验为:

当$\frac{L(\theta_1|x)}{L(\theta_0|x)} < \frac{\pi_0}{\pi_1}$ 时, 则接受H_0; 当$\frac{L(\theta_1|x)}{L(\theta_0|x)} > \frac{\pi_0}{\pi_1}$时, 则拒绝$H_0$。

证明：由Bayes公式, $\alpha_0 = h(\theta_0|x) = \frac{\pi_0 L(\theta_0|x)}{\pi_0 L(\theta_0|x) + \pi_1 L(\theta_1|x)}$,

$$\alpha_1 = h(\theta_1|x) = \frac{\pi_1 L(\theta_1|x)}{\pi_0 L(\theta_0|x) + \pi_1 L(\theta_1|x)}. \tag{8.86}$$

故当概率比$\frac{\alpha_0}{\alpha_1} = \frac{\pi_0 L(\theta_0|x)}{\pi_1 L(\theta_1|x)} > 1$, 即$\frac{L(\theta_1|x)}{L(\theta_0|x)} < \frac{\pi_0}{\pi_1}$ 时, 则接受H_0。

定理 8.12 设似然函数为$\{L(\theta|x);\theta\in\Theta\}$, $\Theta_0\cap\Theta_1=\varphi$。假设检验问题为

$$H_0:\theta\in\Theta_0;\ H_1:\theta\in\Theta_1 \tag{8.87}$$

设$\pi_0=p\{\theta\in\Theta_0\},\pi_1=p\{\theta\in\Theta_1\}$, 又$H_0$成立时，先验概率密度$\mu_0(\theta),\theta\in\Theta_0$; H_1成立时，先验概率密度$\mu_1(\theta),\theta\in\Theta_1$。即

$$\pi(\theta)=\begin{cases}\mu_0(\theta) & \theta\in\Theta_0\\ \mu_1(\theta) & \theta\in\Theta_1\end{cases} \tag{8.88}$$

则贝叶斯检验为:

当$\frac{\int_{\Theta_1}\mu_1(\theta)L(\theta|x)d\theta}{\int_{\Theta_0}\mu_0(\theta)L(\theta|x)d\theta}<\frac{\pi_0}{\pi_1}$时，则接受$H_0$; 否则拒绝$H_0$。

证明：由Bayes公式,

$$\alpha_0=P(\theta\in\Theta_0|x)=\frac{\pi_0\int_{\Theta_0}\mu_0(\theta)L(\theta|x)d\theta}{\pi_0\int_{\Theta_0}\mu_0(\theta)L(\theta|x)d\theta+\pi_1\int_{\Theta_1}\mu_1(\theta)L(\theta|x)d\theta}, \tag{8.89}$$

$$\alpha_1=P(\theta\in\Theta_1|x)=\frac{\pi_1\int_{\Theta_0}\mu_1(\theta)L(\theta|x)d\theta}{\pi_0\int_{\Theta_0}\mu_0(\theta)L(\theta|x)d\theta+\pi_1\int_{\Theta_1}\mu_1(\theta)L(\theta|x)d\theta}. \tag{8.90}$$

故当概率比

$\frac{\alpha_0}{\alpha_1}=\frac{\pi_0\int_{\Theta_0}\mu_0(\theta)L(\theta|x)d\theta}{\pi_1\int_{\Theta_0}\mu_1(\theta)L(\theta|x)d\theta}>1$, 即 $\frac{\int_{\Theta_1}\mu_1(\theta)L(\theta|x)d\theta}{\int_{\Theta_0}\mu_0(\theta)L(\theta|x)d\theta}<\frac{\pi_0}{\pi_1}$ 时, 则接受 H_0; 否则拒绝 H_0。

例 8.15 设$X_1,\ldots X_n$是来自正态总体$N(\mu,1)$的i.i.d.样本，要检验假设

$$H_0:\mu=\mu_0;\ H_1:\mu\neq\mu_0 \tag{8.91}$$

设H_0成立时先验密度$P\{\mu=\mu_0\}=1$; H_1成立时先验密度$\mu_1(\mu)\propto 1$; 且π_0,π_1已知。

解：此时，$\Theta_0=\{\mu_0\}$是单点集，$\mu_0(\theta)$为退化分布;

$$\frac{\int_{\Theta_1}\mu_1(\theta)L(\theta|x)d\theta}{\int_{\Theta_0}\mu_0(\theta)L(\theta|x)d\theta}=\frac{\int_{\mu\neq\mu_0}e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}d\mu}{e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_0)^2}}=\sqrt{\frac{2\pi}{n}}e^{\frac{n}{2}(\bar{x}-\mu_0)^2} \tag{8.92}$$

由定理当$\sqrt{\frac{2\pi}{n}}e^{\frac{n}{2}(\bar{x}-\mu_0)^2}>\frac{\pi_0}{\pi_1}$时，拒绝$H_0$。故当$|\bar{x}-\mu_0|$大于某个界限时拒绝$H_0$。

8.5 多元正态分布的参数估计与假设检验

总所周知，多元正态分布是概率统计中最重要的分布，在多元统计分析中也占有极其重要的地位。关于多元正态分布参数的估计和假设检验问题，在理论上和应用中也已经非常成熟，形成了一整套行之有效的方法。限于篇幅，本节择其要点对这一专题做一个简单介绍。

8.5.1 多元正态分布的定义和性质

定义 8.7 设p维随机向量$X=(X_1,\cdots,X_p)^T$的联合概率密度函数为

$$f(x)=\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\} \tag{8.93}$$

其中$x=(x_1,\cdots,x_n)^T,\mu=(\mu_1,\cdots,\mu_p)^T$是$p$维实向量，$\Sigma$是$p$维正定矩阵，则称$X$服从$p$维正态分布，记为$X\sim N_p(\mu,\Sigma)$。

当$p=1$时，即为一元正态分布密度函数。

当$p=2$时，设$\boldsymbol{X}=(X_1,X_2)^T$服从二元正态分布，协方差矩阵

$$\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}=\begin{pmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho\\ \sigma_2\sigma_1\rho & \sigma_2^2\end{pmatrix}, \tag{8.94}$$

$$\boldsymbol{\Sigma}^{-1}=\frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\begin{pmatrix}\sigma_2^2 & -\sigma_1\sigma_2\rho\\ -\sigma_2\sigma_1\rho & \sigma_1^2\end{pmatrix}. \tag{8.95}$$

则

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right\}. \tag{8.96}$$

下面给出多元正态分布的基本性质：

(1) 设$\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^T\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，且$\boldsymbol{\Sigma}$是对角矩阵，则$X_1$，$X_2,\cdots,X_p$相互独立且都服从正态分布；

(2) 若$\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^T\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，且$A$为$s\times p$阶常数阵，$d$为$s$维常数向量，则$\boldsymbol{AX}+\boldsymbol{d}\sim N_s(\boldsymbol{A\mu}+\boldsymbol{d},\ \boldsymbol{A\Sigma A'})$。该性质称之为正态分布的线性变换不变性，即正态随机向量的线性函数还是正态的。

(3) 设$\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$，$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}^{(1)} \\ \boldsymbol{X}^{(2)} \end{pmatrix}^{q}_{p-q}$，$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix}^{q}_{p-q}$，$\Sigma = \begin{matrix} q & p-q \\ \end{matrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{q}_{p-q}$，

则

$$\boldsymbol{X}^{(1)} \sim N_q(\boldsymbol{\mu}^{(1)}, \Sigma_{11}), \boldsymbol{X}^{(2)} \sim N_{p-q}(\boldsymbol{\mu}^{(2)}, \Sigma_{22}) \tag{8.97}$$

8.5.2 多元正态分布的参数估计

设样本资料矩阵为

$$\boldsymbol{X} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \cdots, \boldsymbol{X}_p) = \begin{pmatrix} \boldsymbol{X}^T_{(1)} \\ \boldsymbol{X}^T_{(2)} \\ \vdots \\ \boldsymbol{X}^T_{(n)} \end{pmatrix} \tag{8.98}$$

下面给出常用的多元正态总体的相关量：

(1) 样本均值向量

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{X}_{(k)} = (\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_p)^T \tag{8.99}$$

(2) 样本离差阵

$$\boldsymbol{S}_{p\times p} = \sum_{k=1}^{n} (\boldsymbol{X}_{(k)} - \bar{\boldsymbol{X}})(\boldsymbol{X}_{(k)} - \bar{\boldsymbol{X}})^T = (s_{ij})_{p\times p} \tag{8.100}$$

其中

$$\begin{aligned}
&\sum_{k=1}^{n}(\boldsymbol{X}_{(k)}-\bar{\boldsymbol{X}})(\boldsymbol{X}_{(k)}-\bar{\boldsymbol{X}})^{T}\\
&=\sum_{k=1}^{n}\left(\begin{pmatrix}X_{k1}-\bar{X}_1\\X_{k2}-\bar{X}_2\\\vdots\\X_{kp}-\bar{X}_p\end{pmatrix}(X_{k1}-\bar{X}_1,X_{k2}-\bar{X}_2,\cdots,X_{kp}-\bar{X}_p)\right)\\
&=\sum_{k=1}^{n}\begin{pmatrix}(X_{k1}-\bar{X}_1)^2 & (X_{k1}-\bar{X}_1)(X_{k2}-\bar{X}_2) & \cdots & (X_{k1}-\bar{X}_1)(X_{kp}-\bar{X}_p)\\(X_{k2}-\bar{X}_2)(X_{k1}-\bar{X}_1) & (X_{k2}-\bar{X}_2)^2 & \cdots & (X_{k2}-\bar{X}_2)(X_{kp}-\bar{X}_p)\\\vdots & \vdots & & \vdots\\(X_{kp}-\bar{X}_p)(X_{k1}-\bar{X}_1) & (X_{kp}-\bar{X}_p)(X_{k2}-\bar{X}_2) & \cdots & (X_{kp}-\bar{X}_p)^2\end{pmatrix}\\
&=\begin{pmatrix}s_{11} & s_{12} & \cdots & s_{1p}\\s_{21} & s_{22} & \cdots & s_{2p}\\\vdots & \vdots & & \vdots\\s_{p1} & s_{p2} & \cdots & s_{pp}\end{pmatrix}=(s_{ij})_{p\times p}
\end{aligned}\tag{8.101}$$

(3) 样本协方差阵

$$\boldsymbol{V}_{p\times p}=\frac{1}{n}\boldsymbol{S}=\frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{X}_{(k)}-\bar{\boldsymbol{X}})(\boldsymbol{X}_{(k)}-\bar{\boldsymbol{X}})'=(v_{ij})_{p\times p}\tag{8.102}$$

其中

$$\begin{aligned}
\frac{1}{n}\boldsymbol{S}&=\frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{X}_{(k)}-\bar{\boldsymbol{X}})(\boldsymbol{X}_{(k)}-\bar{\boldsymbol{X}})'\\
&=\left(\frac{1}{n}\sum_{k=1}^{n}(X_{ki}-\bar{X}_i)(X_{kj}-\bar{X}_j)\right)_{p\times p}=(v_{ij})_{p\times p}
\end{aligned}\tag{8.103}$$

(4) 样本相关阵

$$\hat{\boldsymbol{R}}_{p\times p}=(r_{ij})_{p\times p}\tag{8.104}$$

其中$r_{ij}=\frac{v_{ij}}{\sqrt{v_{ii}}\sqrt{v_{jj}}}=\frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}$。

下面讨论$\boldsymbol{\mu},\boldsymbol{\Sigma}$的极大似然估计。 极大似然法是通过似然函数即 样品的联合分布密度函数来求总体参数的估计量。

首先，构造似然函数，即容量为n的 独立随机样本的联合分布密度函数

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma})=\prod_{i=1}^{n}f(\boldsymbol{X}_{(i)},\boldsymbol{\mu},\boldsymbol{\Sigma})$$

$$= \frac{1}{(2\pi)^{pn/2}|\Sigma|^{n/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{X}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu})\right\}. \quad (8.105)$$

在上式中，样本一旦抽定，$(\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}, \cdots, \boldsymbol{X}_{(n)})$就是已知的常数向量，而只有总体均值向量$\boldsymbol{\mu}$和$\boldsymbol{\Sigma}$未知。因此，可将此式看作是$\boldsymbol{\mu}, \boldsymbol{\Sigma}$的似然函数。为了求出使此似然函数取极大值的$\boldsymbol{\mu}, \boldsymbol{\Sigma}$的值，将此似然函数的两边取对数：

$$\ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}pn\ln(2\pi) - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{X}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu}) \quad (8.106)$$

根据矩阵的微分理论，由于对实对称矩阵A，有

$$\frac{\partial(x'Ax)}{\partial x} = 2Ax, \frac{\partial(x'Ax)}{\partial A} = x'x, \frac{\partial \ln|A|}{\partial A} = A^{-1} \quad (8.107)$$

所以

$$\begin{cases} \frac{\partial \ln L(\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{n}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu}) = 0 \\ \frac{\partial \ln L(\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} = -\frac{n}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})'(\boldsymbol{\Sigma}^{-1})^2 = 0 \end{cases} \quad (8.108)$$

从而得到$\boldsymbol{\mu}$和$\boldsymbol{\Sigma}$的极大似然估计量为

$$\begin{cases} \hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i = \bar{\boldsymbol{X}} \\ \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})(\boldsymbol{X}_i-\bar{\boldsymbol{X}})' = \frac{1}{n}\boldsymbol{S} \end{cases} \quad (8.109)$$

下面讨论估计量的基本性质：

(1) $E(\bar{\boldsymbol{X}}) = \boldsymbol{\mu}$，即$\bar{\boldsymbol{X}}$是$\boldsymbol{\mu}$ 的无偏估计；

(2) $E\left(\frac{1}{n-1}\boldsymbol{S}\right) = \boldsymbol{\Sigma}$，即$\frac{1}{n-1}\boldsymbol{S}$是$\boldsymbol{\Sigma}$的无偏估计；

(3) $\bar{\boldsymbol{X}}, \frac{1}{n-1}\boldsymbol{S}$分别是$\boldsymbol{\mu}$和$\boldsymbol{\Sigma}$的“最小方差”无偏估计量，即是其有效估计；

此外，$\bar{\boldsymbol{X}}$和$\boldsymbol{S}$还满足下面重要的结论。

定理 8.13 设$\bar{\boldsymbol{X}}$和$\boldsymbol{S}$分别是正态总体$N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$的样本均值向量和离差阵，则

(1) $\bar{\boldsymbol{X}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$；

(2) 离差阵可以写为$\boldsymbol{S} = \sum_{k=1}^{n-1}\boldsymbol{Z}_k\boldsymbol{Z}_k^T$ 其中，$\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_{n-1}$相互独立且都服从$N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$；

(3) $\bar{\boldsymbol{X}}$和$\boldsymbol{S}$相互独立；

(4) $\boldsymbol{S}$为正定阵的充要条件是$n > p$。

例 8.16 已知$\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)^T$服从正态分布$\boldsymbol{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，从中抽取容量为20的一个样本，样本观察值见下表：

序号	1	2	3	4	5	6	7	8	9	10
x_1	63	63	70	6	65	9	10	12	20	30
x_2	971	892	1125	82	931	112	162	321	315	357

序号	11	12	13	14	15	16	17	18	19	20
x_1	33	27	21	5	14	27	17	53	62	65
x_2	462	352	305	34	229	332	185	703	872	740

求μ和Σ的最小方差无偏估计。

解：注意到，μ和Σ的最小方差无偏估计量为

$$\hat{\mu}=\bar{\boldsymbol{X}},\ \hat{\Sigma}=\frac{\boldsymbol{S}}{\boldsymbol{n}-1}=\frac{1}{\boldsymbol{n}-1}\sum_{k=1}^{n}(\boldsymbol{X}_k-\bar{\boldsymbol{X}})(\boldsymbol{X}_k-\bar{\boldsymbol{X}})^{\mathrm{T}}.$$

代入样本值可得

$$\hat{\mu}=\frac{1}{\boldsymbol{n}}\begin{pmatrix}\sum\limits_{i=1}^{20}x_{1i}\\ \sum\limits_{i=1}^{20}x_{2i}\end{pmatrix}=\begin{pmatrix}33.85\\ 477.50\end{pmatrix}=\begin{pmatrix}\bar{x}_1\\ \bar{x}_2\end{pmatrix},$$

$$\begin{aligned}\hat{\Sigma}&=\frac{\boldsymbol{S}}{\boldsymbol{n}-1}=\frac{1}{\boldsymbol{n}-1}\sum_{k=1}^{n}(\boldsymbol{X}_k-\bar{\boldsymbol{X}})(\boldsymbol{X}_k-\bar{\boldsymbol{X}})^{\mathrm{T}}\\&=\frac{1}{20-1}\begin{pmatrix}\sum\limits_{i=1}^{20}(x_{1i}-\bar{x}_1)^2 & \sum\limits_{i=1}^{20}(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)\\ \sum\limits_{i=1}^{20}(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2) & \sum\limits_{i=1}^{20}(x_{2i}-\bar{x}_2)^2\end{pmatrix}\\&=\begin{pmatrix}570.45 & 7845.08\\ 7845.08 & 112404.26\end{pmatrix}\end{aligned}$$

8.5.3 多元正态总体参数的假设检验

首先给出几个重要统计量的分布。

定义 8.8 设$\boldsymbol{X}_{(a)}=(X_{a1},X_{a2},\cdots,X_{ap})'\sim N_p(\boldsymbol{\mu}_a,\boldsymbol{\Sigma})$，其中$a=1,2,\cdots,n$且相互独立，则由$\boldsymbol{X}_{(a)}$组成的随机矩阵

$$\boldsymbol{W}_{p\times p}=\sum_{a=1}^{n}\boldsymbol{X}_{(a)}\boldsymbol{X}_{(a)}^{T}\tag{8.110}$$

的分布称为非中心Wishart分布，记为$W_p(n,\Sigma,Z)$，其中$Z=\sum_{a=1}^{n}\mu_a\mu_a^T$，$\mu_a$称为非中心参数；当$\mu_a=0$时称为中心Wishart分布，记为$W_p(n,\Sigma)$，当$n\geq p$，$\Sigma>0$，$W_p(n,\Sigma)$有密度存在，其表达式为

$$f(w)=\begin{cases}\dfrac{|w|^{\frac{1}{2}(n-p-1)}\exp\left\{-\frac{1}{2}tr\Sigma^{-1}w\right\}}{2^{np/2}\pi^{p(p-1)/4}|\Sigma|^{n/2}\prod_{i=1}^{p}\Gamma\left(\frac{n-i+1}{2}\right)}, & \text{当}w\text{为正定阵}\\ 0, & \text{其它.}\end{cases}\tag{8.111}$$

显然，当$p=1$，$\Sigma=\sigma^2$时，$f(w)$就是$\sigma^2\chi^2(n)$的分布密度，此时$W=\sum_{a=1}^{n}X_{(a)}^2$，有$\frac{1}{\sigma^2}\sum_{a=1}^{n}X_{(a)}^2\sim\chi^2(n)$。因此，Wishart分布是$\chi^2$分布在$p$维正态情况下的推广。

下面给出Wishart分布的基本性质：

(1) 若$X_{(a)}\sim N_p(\mu,\Sigma)$，$a=1,2,\cdots,n$且相互独立，则样本离差阵

$$S=\sum_{a=1}^{n}(\mathrm{X}_{(a)}-\bar{\mathrm{X}})(\mathrm{X}_{(a)}-\bar{\mathrm{X}})'\sim W_p(n-1,\Sigma)\tag{8.112}$$

其中$\bar{\mathrm{X}}=\frac{1}{n}\sum_{a=1}^{n}\mathrm{X}_{(a)}$;

(2) 若$S_i\sim W_p(n_i,\Sigma)$，$i=1,\cdots,k$，且相互独立，则$\sum_{i=1}^{k}S_i\sim W_p(\sum_{i=1}^{k}n_i,\Sigma)$;

(3) 若$X_{p\times p}\sim W_p(n,\Sigma)$，$C_{p\times p}$为非奇异阵，则$CXC'\sim W_p(n,C\Sigma C')$。

定义 8.9 设$X\sim N_p(\mu,\Sigma)$，$S\sim W_p(n,\Sigma)$，且相互独立，$n\geq p$，则称统计量$T^2=nX'S^{-1}X$的分布为非中心 Hotelling T^2分布，记为$T^2\sim T^2(p,n,\mu)$。当$\mu=0$时，称T^2服从(中心) Hotelling T^2分布，记为$T^2(p,n)$。

下面给出 Hotelling T^2 分布与 F 分布的关系。

定理 8.14 设$X\sim N_p(0,\Sigma)$，$S\sim W_p(n,\Sigma)$，且相互独立，令$T^2=nX'S^{-1}X$，则

$$\frac{n-p+1}{np}T^2\sim F(p,n-p+1)\tag{8.113}$$

定理表明，将 Hotelling T^2 统计量乘上一个适当的常数后，便成为F统计量，从而可利用F分布来进行推断。应用中常用上式给出的F统计量来代替 Hotelling T^2 统计量进行推断。

1、正态总体均值向量的检验

设$X_{(1)},X_{(2)},\cdots,X_{(n)}$来自$p$ 维正态总体$N_p(\mu,\Sigma)$，$\bar{X}$和S定义同前所述，

假设检验的问题为

$$H_0:\ \mu=\mu_0,\ H_1:\ \mu\neq\mu_0\tag{8.114}$$

情形1 协差阵Σ已知时均值向量的检验

当原假设H_0：$\mu=\mu_0$成立时，由于

$$T_0^2=n(\bar{X}-\mu_0)'\Sigma^{-1}(\bar{X}-\mu_0)\sim\chi^2(p) \tag{8.115}$$

因此，对于给定的检验水平α，查χ^2分布表找出临界值χ_α^2。若$T_0^2>\chi_\alpha^2$，则接受备择假设；否则接受原假设。

情形2 协差阵Σ未知时均值向量的检验

当原假设H_0：$\mu=\mu_0$成立时，由于$\frac{(n-1)-p+1}{(n-1)p}T^2\sim F(p,n-p)$，其中$T^2=(n-1)[\sqrt{n}\left(\bar{X}-\mu_0\right)^T S^{-1}\sqrt{n}(\bar{X}-\mu_0)]\sim T^2(p,n-1)$，此时对于给定的检验水平$\alpha$，查$F$分布表找出临界值$F_\alpha$。若$\frac{n-p}{(n-1)p}T^2>F_\alpha$，则接受备择假设；否则接受原假设。

例 8.17 人的出汗多少与人体内的钾和钠的含量有一定的关系，现在测量了20名健康成年女性的出汗量(X_1)、钠的含量(X_2)和钾的含量(X_3)，样本值见下表：

成年女性的出汗量和体内钾含量与钠含量的测试数据

序号	X_1	X_2	X_3	序号	X_1	X_2	X_3
1	3.7	48.5	9.3	11	4.7	65.1	8.0
2	3.8	47.2	10.9	12	3.2	53.2	12.0
3	3.1	55.5	9.7	13	4.6	36.1	7.9
4	2.4	24.8	14.0	14	7.2	33.1	7.6
5	6.7	47.4	8.5	15	5.4	54.1	11.3
6	3.9	36.9	12.7	16	4.5	58.8	12.3
7	3.5	27.8	9.8	17	4.5	40.2	8.4
8	1.5	13.5	10.1	18	8.5	56.4	7.1
9	4.5	71.6	8.2	19	6.5	52.8	10.9
10	4.1	44.1	11.2	20	5.5	40.9	9.4

取显著性水平为$\alpha=0.05$，试检验

$$H_0:\mu=\mu_0=(4,50,10)^T\quad H_1:\mu\neq\mu_0 \tag{8.116}$$

解：随机向量$X=(X_1,X_2,X_3)^T\sim N_3(\mu,\Sigma)$，检验

$$H_0:\mu=\mu_0=(4,50,10)^T\quad H_1:\mu\neq\mu_0.$$

取检验统计量为$F=\frac{n-p}{(n-1)p}T^2$，其中$p=3,n=20$，则有样本值可得

$$\bar{X}=(4.64,45.4,9.965)^T,$$

$$S=\begin{pmatrix}54.708 & 0 & 0\\ 190.190 & 3795.98 & 0\\ -34.372 & -107.16 & 68.926\end{pmatrix},$$

$$S^{-1}=\begin{pmatrix}0.0308503 & 0 & 0\\ -0.001162 & 0.0003193 & 0\\ 0.0135773 & 0.0000830 & 0.0211498\end{pmatrix}.$$

进一步可求得

$$D^2=(n-1)(\bar{X}-\mu_0)^TS^{-1}(\bar{X}-\mu_0)=19\times 0.02563=0.48694,$$

$$T^2=n(n-1)(\bar{X}-\mu_0)^TS^{-1}(\bar{X}-\mu_0)=9.7388,$$

$$F=\frac{n-p}{(n-1)p}T^2=2.9045.$$

根据F分布的上0.05分位点表可知，$F_{0.05}(3,17)=3.2$，注意到$2.9045<3.2$，所以接受原假设。

2、协方差矩阵的检验

设$\boldsymbol{X}_{(1)},\boldsymbol{X}_{(2)},\cdots,\boldsymbol{X}_{(n)}$来自$p$ 维正态总体$N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，其中协方差矩阵$\boldsymbol{\Sigma}$未知，假设检验问题为

$$H_0:\ \boldsymbol{\Sigma}=\boldsymbol{\Sigma}_0,\ H_1:\ \boldsymbol{\Sigma}\neq\Sigma_0 \tag{8.117}$$

其中$\boldsymbol{\Sigma}_0$为已知正定矩阵。

情形 1 $\boldsymbol{\Sigma}_0=I_p$.

利用似然比原理构造似然比统计量λ为

$$\lambda=|\boldsymbol{S}|^{\frac{n}{2}}\left(\frac{e}{n}\right)^{\frac{np}{2}}\exp\left\{-\frac{1}{2}tr\boldsymbol{S}\right\} \tag{8.118}$$

其中$\boldsymbol{S}=\sum\limits_{a=1}^{n}(\boldsymbol{X}_{(a)}-\bar{\boldsymbol{X}})(\boldsymbol{X}_{(a)}-\bar{\boldsymbol{X}})^T$。对于给定的检验水平$\alpha$，因为直接由$\lambda_1$的分布计算临界值$\lambda_0$ 很困难，所以通常采用λ的近似分布。

在原假设成立时，当n较大时，其对数$-2\ln\lambda$的极限分布是$\chi^2\left(\frac{1}{2}p(p+1)\right)$分布。因此当$n>>p$时，由样本值计算出$\lambda$值，若$-2\ln\lambda>\chi^2_\alpha$，即$\lambda<e^{-\chi^2_\alpha/2}$时，则拒绝原假设，否则接受原假设。

情形 2 $\boldsymbol{\Sigma}_0\neq I_p$.

此时由于Σ_0正定，于是存在可逆矩阵D，满足$D\Sigma_0D^T=I_p$，令

$$Y_{(a)}=DX_{(a)},a=1,2,\cdots,n \tag{8.119}$$

则由多元正态分布的性质可知

$$Y_{(a)} \sim N_p(D\mu, D\Sigma D^T), \text{ 记为} Y_{(a)} \sim N_p(\mu^*, \Sigma^*) \tag{8.120}$$

此时，原假设$H_0: \Sigma = \Sigma_0$等价于$H_0: \Sigma^* = I_p$，此时利用样本$\boldsymbol{Y}_{(1)}, \boldsymbol{Y}_{(2)}, \cdots, \boldsymbol{Y}_{(n)}$，可构造似然比统计量为

$$\lambda = |\boldsymbol{S}^*|^{\frac{n}{2}} \left(\frac{e}{n}\right)^{\frac{np}{2}} \exp\left\{-\frac{1}{2} tr \boldsymbol{S}^*\right\} \tag{8.121}$$

其中$\boldsymbol{S}^* = \sum_{a=1}^{n} (Y_{(a)} - \bar{Y})(Y_{(a)} - \bar{Y})^T = DSD^T$。但是研究该统计量的抽样分布非常困难，因此也可以根据情形1的方法构造检验方法。

第八章习题

1. 为了调查某广告对销售收入的影响，某商店记录了7个月的销售收入y(万元)与广告费用x(万元)，数据如下表，

x	1	2	3	4	5	6	7
y	10	10	16	20	23	40	43

求参数β_0, β_1的最小二乘估计。

2. 10组观测数据由下表给出:

x	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4
y	1.1	1.2	1.3	1.5	1.6	1.7	1.9	2	2.1	2.2

应用正态线性回归模型$y=\beta_0+\beta_1 x+\varepsilon$， 其中误差项$\varepsilon$服从正态分布$N(0,\sigma^2)$，

(1) 求回归方程;

(2) σ^2的无偏估计值$\hat{\sigma}^2$;

(3) 求β_1的置信水平为95%的置信区间;

(4) 在显著性水平$\alpha=0.05$下相关系数的显著性检验;

(5) 观察值y的置信水平为95%的预测区间。

3. 对于一元线性回归的样本模型

$$\begin{cases} y_i=\beta_0+\beta_1 x_i+\varepsilon_i \quad (i=1,2,\cdots,n) \\ \varepsilon_i\text{相互独立}, \ E(\varepsilon_i)=0, \mathrm{var}(\varepsilon_i)=\sigma^2 \end{cases}$$

请给出参数β_0,β_1的最小二乘估计值$\hat{\beta}_0,\hat{\beta}_1$的推导。

4. 某课题研究四种衣料内棉花吸附十硼氢量。 每种衣料各做五次测量，所得数据如下表，试检验各种衣料棉花吸附十硼氢量有没有差异。 ($\alpha=0.05$)

衣料1	衣料2	衣料3	衣料4
2.33	2.48	3.06	4.00
2.00	2.34	3.06	5.13
2.93	2.68	3.00	4.61
2.73	2.34	2.66	2.80
2.33	2.22	3.06	3.60

5. 设有三种机器A，B，C制造一种产品，对每种机器各观测5天，其日产量如下表所示，问机器与机器之间是否存在差异?($\alpha = 0.05$)

天数机器	1	2	3	4	5
A	41	48	41	49	57
B	65	57	54	72	64
C	45	51	56	48	48

6. 以淀粉为原料生产葡萄糖过程中，残留的许多糖蜜可用于酱色生产。生产酱色之前应尽可能彻底除杂，以保证酱色质量。今选用4种除杂方法，每种方法做4次试验，试验结果如下表，试分析不同除杂方法的除杂效果有无差异?($\alpha = 0.05$)

除杂方法	除杂质量			
方法一	25.6	24.7	21.6	25.3
方法二	26.4	28.9	19.7	24.6
方法三	22.1	23.1	24.2	30
方法四	25.1	24.6	22.6	25.7

7. 高新技术公司的工程师为了确认不同厂家的塑胶材料对成型品拉拔力的影响，分别对该公司正在使用的5家供应商提供的塑胶料成型品进行拉拔试验，取得数据如下：

供应商A	供应商B	供应商C	供应商D	供应商E
5.86	4.95	5.60	5.88	4.97
5.95	5.72	6.21	6.42	4.86
5.74	5.36	5.81	6.51	5.03
5.62	5.89	5.30	5.98	6.05
5.08	4.99	4.97	5.74	4.10

试分析不同供应商所提供的塑胶材料是否有显著不同？($\alpha = 0.05$)

8. 设$\boldsymbol{X} = (X_1, X_2, \cdots X_n)'$是来自二项分布总体$B(m,p)$的$i.i.d$样本，其中$m$已知，$0 < p < 1$未知。

   1) 验证$\pi(p) \propto p^{-1}(1-p)^{-1}$是广义先验分布;

   2) 在上述先验分布下，求p的后验分布。并找出p的充分统计量;

   3) 若p的先验分布取为$\beta(a,b)$, 验证$\beta(a,b)$是否为二项分布总体$B(m,p)$共轭先验分布。

9. 设$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$是来自Poisson分布总体$P(\lambda)$的$i.i.d$。样本，$\lambda$的先验分布取为$\Gamma$分布$\Gamma(\alpha,\mu)$，则$\Gamma$布族$\{\Gamma(\alpha,\mu):\alpha>0,\mu>0\}$是Poisson分布族$\{P(\lambda),\lambda>0\}$的共轭先验分布族。

10. 设$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$是来自0-1分布总体$B(1,p)$的$i.i.d$样本，设$p$的先验分布为

$$\pi(p)=\begin{cases}\frac{1}{p},\ p_0<p<1\\ 0,\ \text{其它}\end{cases}$$

其中p_0是$(0,1)$中的固定常数。求证p的条件期望估计为

$$\hat{p}=\frac{\sum\limits_{i=1}^{n}X_i}{n+2}\cdot\frac{1-I_{p_0}\left(\sum\limits_{i=1}^{n}X_i+2,n-\sum\limits_{i=1}^{n}X_i+1\right)}{1-I_{p_0}\left(\sum\limits_{i=1}^{n}X_i+1,n-\sum\limits_{i=1}^{n}X_i+1\right)}$$

其中

$$I_{p_0}(\alpha,\beta)=\frac{1}{B(\alpha,\beta)}\int_0^{p_0}p^{\alpha-1}(1-p)^{\beta-1}\mathrm{d}p$$

11. 设$\boldsymbol{X}=(X_1,X_2,\cdots X_n)'$是来自$\Gamma$分布总体$\Gamma\left(\frac{1}{2},\frac{1}{2\theta}\right)$的$i.i.d$样本，

(1) 求证逆Γ分布$I\Gamma(\alpha,\lambda)$分布是Γ分布$\Gamma\left(\frac{1}{2},\frac{1}{2\theta}\right)$的共轭先验分布；

(2) 求θ的满足杰弗莱原则的先验分布；

(3) 在(1)，(2)两种先验分布下，求θ的条件期望估计与最大后验估计。

参考文献

1. R.A. Horn，C.R. Johnson 著，杨奇译，矩阵分析，机械工业出版社，2005.
2. G.H. 戈卢布，C.F.范洛恩[美]著，袁亚湘等译，矩阵计算，科学技术出版社，2005.
3. S. Mallat 著，戴道清，杨力华译，信号处理的小波导引：稀疏方法，机械工业出版社，2012.
4. 曹志浩，变分迭代法，科学技术出版社，2006.
5. 陈宝林，最优化理论与算法（第二版），清华大学出版社，2015.
6. 谷超豪 等，数学物理方程（第3版），高等教育出版社，2012.
7. 顾基发主编，刘宝碇，施泉生副主编，运筹学，科学出版社，2011.
8. 何晓群，刘文卿，应用回归分析（第二版），中国人民大学出版社，2007.
9. 焦李成，赵进，杨淑媛，刘芳，深度学习、优化与识别，清华大学出版社，2017.
10. 雷纪刚，唐平，矩阵论及其应用，机械工业出版社，2005.
11. 李刚，刘文军，数学物理方程：模型、方法与应用，科学出版社，2018.
12. 李庆扬等，数值计算原理，清华大学出版社，2001.
13. 李治平，偏微分方程数值解讲义，北京大学出版社，2010.
14. 梁昆淼，数学物理方法（第四版），高等教育出版社，2010.
15. 陆金甫，关治，偏微分方程数值解法（第3版），清华大学出版社，2016.
16. 茆诗松，贝叶斯统计(第二版),中国统计出版社，2012.
17. 石钟慈，王鸣著，有限元方法，科学出版社，2010.
18. 孙文瑜，徐成贤，朱德通，最优化方法（第二版），高等教育出版社，2010.
19. 汪定伟 等，智能优化方法，高等教育出版社，2007.

20. 王凌， 智能优化算法及其应用， 清华大学出版社， 2004.

21. 王元明，工程数学-数学物理方程与特殊函数，高等教育出版社，2012.

22. 魏毅强等，数值计算方法，科学出版社，2018.

23. 谢政，李建平，陈挚，非线性最优化理论与算法，高等教育出版社，2012.

24. 邢文训，谢金星，现代优化计算方法（第二版），清华大学出版社， 2005.

25. 邢志栋等，矩阵数值分析（第二版），陕西科学技术出版社，2005.

26. 徐仲，张凯院，陆全，冷国伟，矩阵论简明教程，科学出版社，2005.

27. 叶慈南，曹伟丽，应用数理统计，机械工业出版社，2004.

28. 张尧庭，方开泰，多元统计分析引论，科学出版社，2017.

29. 张文生，科学计算中的偏微分方程有限差分法，高等教育出版社，2006.

30. 张文生，微分方程数值解：有限差分理论方法与数值计算，科学出版社，2015.

31. 朱元国，饶玲，严涛，张军，李宝成 编，矩阵分析与计算，国防工业出版社，2010.

附表

表 8.4: 相关系数临界值表

$n-2$	5%	1%	$n-2$	5%	1%	$n-2$	5%	1%
1	0.997	1	16	0.468	0.59	35	0.325	0.418
2	0.95	0.99	17	0.456	0.575	40	0.304	0.393
3	0.878	0.959	18	0.444	0.561	45	0.288	0.372
4	0.811	0.947	19	0.433	0.549	50	0.273	0.354
5	0.754	0.874	20	0.423	0.537	60	0.250	0.325
6	0.707	0.834	21	0.413	0.526	70	0.232	0.302
7	0.666	0.798	22	0.404	0.515	80	0.217	0.283
8	0.632	0.765	23	0.396	0.505	90	0.205	0.267
9	0.602	0.735	24	0.388	0.496	100	0.195	0.254
10	0.576	0.708	25	0.381	0.487	125	0.174	0.228
11	0.553	0.684	26	0.374	0.478	150	0.159	0.208
12	0.532	0.661	27	0.367	0.470	200	0.138	0.181
13	0.514	0.641	28	0.361	0.463	300	0.113	0.148
14	0.497	0.623	29	0.355	0.456	400	0.098	0.128
15	0.482	0.606	30	0.349	0.449	1000	0.062	0.081
$n-2$	5%	1%	$n-2$	5%	1%	$n-2$	5%	1%

表 8.5: 标准正态分布函数表

$$\Phi(x)=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt=P\{X\leq x\}$$

x	0	1	2	3	4	5	6	7	8	9	x
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359	0.0
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	0.1
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	0.2
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517	0.3
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879	0.4
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224	0.5
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	0.6
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	0.7
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	0.8
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389	0.9
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	1.0
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830	1.1
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015	1.2
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177	1.3
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319	1.4
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441	1.5
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545	1.6
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633	1.7
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706	1.8
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767	1.9
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817	2.0
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857	2.1
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890	2.2
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916	2.3
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936	2.4
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952	2.5
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964	2.6
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974	2.7
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981	2.8
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986	2.9
3.0	0.9987	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000	3.0
x	0	1	2	3	4	5	6	7	8	9	x

注：表中3.0所在行是函数值 $\Phi(3.0)$, $\Phi(3.1)$, $\Phi(3.2)$, $\Phi(3.3)$, $\Phi(3.4)$, $\Phi(3.5)$, $\Phi(3.6)$, $\Phi(3.7)$, $\Phi(3.8)$, $\Phi(3.9)$.

表 8.6: t 分布上分位点表

$$P\{t(n) > t_\alpha(n)\} = \alpha$$

n \ α	0.25	0.1	0.05	0.025	0.01	0.005	n
1	1.0000	3.0777	6.3138	12.7062	31.8205	63.6567	1
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248	2
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409	3
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041	4
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321	5
6	0.7176	1.4398	1.9432	2.24469	3.1427	3.7074	6
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995	7
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554	8
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498	9
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693	10
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058	11
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545	12
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123	13
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768	14
15	0.6912	1.3406	1.7531	2.1314	2.6025	2.9467	15
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208	16
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982	17
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784	18
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609	19
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453	20

续表

n \ α	0.995	0.99	0.975	0.95	0.9	0.75	n
21	0.6864	1.3232	1.7207	2.0796	2.5176	2.8314	21
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188	22
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073	23
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.7969	24
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874	25
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787	26
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707	27
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633	28
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564	29
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500	30
31	0.6825	1.3095	1.6955	2.0395	2.4528	2.7440	31
32	0.6822	1.3086	1.6939	2.0369	2.4487	2.7385	32
33	0.6820	1.3077	1.6924	2.0345	2.4448	2.7333	33
34	0.6818	1.3070	1.6909	2.0322	2.4411	2.7284	34
35	0.6816	1.3062	1.6896	2.0301	2.4377	2.7238	35
36	0.6814	1.3055	1.6883	2.0281	2.4345	2.7195	36
37	0.6812	1.3049	1.6871	2.0262	2.4314	2.7154	37
38	0.6810	1.3042	1.6860	2.0244	2.4286	2.7116	38
39	0.6808	1.3036	1.6849	2.0227	2.4258	2.7079	39
40	0.6807	1.3031	1.6839	2.0211	2.4233	2.7045	40
41	0.6805	1.3025	1.6829	2.0195	2.4208	2.7012	41
42	0.6804	1.3020	1.6820	2.0181	2.4185	2.6981	42
43	0.6802	1.3016	1.6811	2.0167	2.4163	2.6951	43
44	0.6801	1.3011	1.6802	2.0154	2.4141	2.6923	44
45	0.6800	1.3006	1.6794	2.0141	2.4121	2.6896	45

表 8.7: χ^2 分布上分位点表

$$P\{\chi^2(n) > \chi^2_\alpha(n)\} = \alpha$$

n \ α	0.995	0.99	0.975	0.95	0.9	0.75	n
1	0.000	0.000	0.001	0.004	0.016	0.102	1
2	0.010	0.020	0.051	0.103	0.211	0.575	2
3	0.072	0.115	0.216	0.352	0.584	1.213	3
4	0.207	0.297	0.484	0.711	1.064	1.923	4
5	0.412	0.554	0.831	1.145	1.610	2.675	5
6	0.676	0.872	1.237	1.635	2.204	3.455	6
7	0.989	1.239	1.690	2.167	2.833	4.255	7
8	1.344	1.646	2.180	2.733	3.490	5.071	8
9	1.735	2.088	2.700	3.325	4.168	5.899	9
10	2.156	2.558	3.247	3.940	4.865	6.737	10
11	2.603	3.053	3.816	4.575	5.578	7.584	11
12	3.074	3.571	4.404	5.226	6.304	8.438	12
13	3.565	4.107	5.009	5.892	7.042	9.299	13
14	4.075	4.660	5.629	6.571	7.790	10.165	14
15	4.601	5.229	6.262	7.261	8.547	11.037	15
16	5.142	5.812	6.908	7.962	9.312	11.912	16
17	5.697	6.408	7.564	8.672	10.085	12.792	17
18	6.265	7.015	8.231	9.390	10.865	13.675	18
19	6.844	7.633	8.907	10.117	11.651	14.562	19
20	7.434	8.260	9.591	10.851	12.443	15.452	20

续表

n \ α	0.995	0.99	0.975	0.95	0.9	0.75	n
21	8.034	8.897	10.283	11.591	13.240	16.344	21
22	8.643	9.542	10.982	12.338	14.041	17.240	22
23	9.260	10.196	11.689	13.091	14.848	18.137	23
24	9.886	10.856	12.401	13.848	15.659	19.037	24
25	10.520	11.524	13.120	14.611	16.473	19.939	25
26	11.160	12.198	13.844	15.379	17.292	20.843	26
27	11.808	12.879	14.573	16.151	18.114	21.749	27
28	12.461	13.565	15.308	16.928	18.939	22.657	28
29	13.121	14.256	16.047	17.708	19.768	23.567	29
30	13.787	14.953	16.791	18.493	20.599	24.478	30
31	14.458	15.655	17.539	19.281	21.434	25.390	31
32	15.134	16.362	18.291	20.072	22.271	26.304	32
33	15.815	17.074	19.047	20.867	23.110	27.219	33
34	16.501	17.789	19.806	21.664	23.952	28.136	34
35	17.192	18.509	20.569	22.465	24.797	29.054	35
36	17.887	19.233	21.336	23.269	25.643	29.973	36
37	18.586	19.960	22.106	24.075	26.492	30.893	37
38	19.289	20.691	22.878	24.884	27.343	31.815	38
39	19.996	21.426	23.654	25.695	28.196	32.737	39
40	20.707	22.164	24.433	26.509	29.051	33.660	40
41	21.421	22.906	25.215	27.326	29.907	34.585	41
42	22.138	23.650	25.999	28.144	30.765	35.510	42
43	22.859	24.398	26.785	28.965	31.625	36.436	43
44	23.584	25.148	27.575	29.787	32.487	37.363	44
45	24.311	25.901	28.366	30.612	33.350	38.291	45

续表

n \ α	0.25	0.1	0.05	0.025	0.01	0.005	n
1	1.323	2.706	3.841	5.024	6.635	7.879	1
2	2.773	4.605	5.991	7.378	9.210	10.597	2
3	4.108	6.251	7.815	9.348	11.345	12.838	3
4	5.385	7.779	9.488	11.143	13.277	14.860	4
5	6.626	9.236	11.070	12.833	15.086	16.750	5
6	7.841	10.645	12.592	14.449	16.812	18.548	6
7	9.037	12.017	14.067	16.013	18.475	20.278	7
8	10.219	13.362	15.507	17.535	20.090	21.955	8
9	11.389	14.684	16.919	19.023	21.666	23.589	9
10	12.549	15.987	18.307	20.483	23.209	25.188	10
11	13.701	17.275	19.675	21.920	24.725	26.757	11
12	14.845	18.549	21.026	23.337	26.217	28.300	12
13	15.984	19.812	22.362	24.736	27.688	29.819	13
14	17.117	21.064	23.685	26.119	29.141	31.319	14
15	18.245	22.307	24.996	27.488	30.578	32.801	15
16	19.369	23.542	26.296	28.845	32.000	34.267	16
17	20.489	24.769	27.587	30.191	33.409	35.718	17
18	21.605	25.989	28.869	31.526	34.805	37.156	18
19	22.718	27.204	30.144	32.852	36.191	38.582	19
20	23.828	28.412	31.410	34.170	37.566	39.997	20

续表

n \ α	0.25	0.1	0.05	0.025	0.01	0.005	n
21	24.935	29.615	32.671	35.479	38.932	41.401	21
22	26.039	30.813	33.924	36.781	40.289	42.796	22
23	27.141	32.007	35.172	38.076	41.638	44.181	23
24	28.241	33.196	36.415	39.364	42.980	45.559	24
25	29.339	34.382	37.652	40.646	44.314	46.928	25
26	30.435	35.563	38.885	41.923	45.642	48.290	26
27	31.528	36.741	40.113	43.195	46.963	49.645	27
28	32.620	37.916	41.337	44.461	48.278	50.993	28
29	33.711	39.087	42.557	45.722	49.588	52.336	29
30	34.800	40.256	43.773	46.979	50.892	53.672	30
31	35.887	41.422	44.985	48.232	52.191	55.003	31
32	36.973	42.585	46.194	49.480	53.486	56.328	32
33	38.058	43.745	47.400	50.725	54.776	57.648	33
34	39.141	44.903	48.602	51.966	56.061	58.964	34
35	40.223	46.059	49.802	53.203	57.342	60.275	35
36	41.304	47.212	50.998	54.437	58.619	61.581	36
37	42.383	48.363	52.192	55.668	59.893	62.883	37
38	43.462	49.513	53.384	56.896	61.162	64.181	38
39	44.539	50.660	54.572	58.120	62.428	65.476	39
40	45.616	51.805	55.758	59.342	63.691	66.766	40
41	46.692	52.949	56.942	60.561	64.950	68.053	41
42	47.766	54.090	58.124	61.777	66.206	69.336	42
43	48.840	55.230	59.304	62.990	67.459	70.616	43
44	49.913	56.369	60.481	64.201	68.710	71.893	44
45	50.985	57.505	61.656	65.410	69.957	73.166	45

表 8.8: F 分布上分位点表

$$P\{F(n_1,n_2) > F_\alpha(n_1,n_2)\} = \alpha$$

$\alpha = 0.1$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	n_2
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	1
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	2
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	3
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	4
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	5
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	6
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	7
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	8
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	9
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	10
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	11
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	12
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	13
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	14
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	15
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	16
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	17
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	18
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	19
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	20
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	21
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	22
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	23
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	24
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	25
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	26
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	27
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	28
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	29
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	30
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	40
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	60
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	120

续表

n_2 \ n_1	10	15	20	24	30	40	60	120	n_2
1	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	1
2	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	2
3	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	3
4	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	4
5	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	5
6	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	6
7	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	7
8	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	8
9	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	9
10	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	10
11	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	11
12	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	12
13	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	13
14	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	14
15	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	15
16	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	16
17	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	17
18	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	18
19	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	19
20	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	20
21	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	21
22	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	22
23	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	23
24	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	24
25	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	25
26	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	26
27	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	27
28	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	28
29	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	29
30	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	30
40	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	40
60	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	60
120	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	120

$\alpha = 0.05$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	n_2
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	1
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	2
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	3
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	4
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	5
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	6
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	7
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	8
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	9
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	10
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	11
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	12
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	13
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	14
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	15
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	16
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	17
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	18
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	19
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	20
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	21
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	22
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	23
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	24
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	25
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	26
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	27
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	28
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	29
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	30
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	40
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	60
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	120

续表

n_2 \ n_1	10	15	20	24	30	40	60	120	n_2
1	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	1
2	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	2
3	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	3
4	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	4
5	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	5
6	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	6
7	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	7
8	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	8
9	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	9
10	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	10
11	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	11
12	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	12
13	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	13
14	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	14
15	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	15
16	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	16
17	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	17
18	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	18
19	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	19
20	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	20
21	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	21
22	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	22
23	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	23
24	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	24
25	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	25
26	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	26
27	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	27
28	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	28
29	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	29
30	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	30
40	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	40
60	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	60
120	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	120

$\alpha = 0.025$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	n_2
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	1
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	2
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	3
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	4
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	5
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	6
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	7
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	8
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	9
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	10
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	11
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	12
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	13
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	14
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	15
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	16
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	17
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	18
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	19
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	20
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	21
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	22
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	23
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	24
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	25
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	26
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	27
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	28
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	29
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	30
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	40
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	60
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	120

续表

n_2 \ n_1	12	15	20	24	30	40	60	120	n_2
1	976.7	984.9	993.1	997.2	1001.4	1005.6	1009.8	1014.0	1
2	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	2
3	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	3
4	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	4
5	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	5
6	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	6
7	4.67	4.57	4.47	4.41	4.36	4.31	4.25	4.20	7
8	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	8
9	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	9
10	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	10
11	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	11
12	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	12
13	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	13
14	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	14
15	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	15
16	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	16
17	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	17
18	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	18
19	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	19
20	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	20
21	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	21
22	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	22
23	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	23
24	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	24
25	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	25
26	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	26
27	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	27
28	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	28
29	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	29
30	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	30
40	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	40
60	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	60
120	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	120

$\alpha = 0.01$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	n_2
1	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	1
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	2
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	3
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	4
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	5
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	6
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	7
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	8
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	9
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	10
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	11
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	12
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	13
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	14
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	15
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	16
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	17
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	18
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	19
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	20
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	21
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	22
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	23
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	24
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	25
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	26
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	27
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	28
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	29
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	30
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	40
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	120

续表

n_2 \ n_1	12	15	20	24	30	40	60	120	n_2
1	6106	6157	6209	6235	6261	6287	6313	6339	1
2	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	2
3	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	3
4	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	4
5	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	5
6	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6
7	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	7
8	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	8
9	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	9
10	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	10
11	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	11
12	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	12
13	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	13
14	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	14
15	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	15
16	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	16
17	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	17
18	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	18
19	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	19
20	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	20
21	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	21
22	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	22
23	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	23
24	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	24
25	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	25
26	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	26
27	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	27
28	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	28
29	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	29
30	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	30
40	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	40
60	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	60
120	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	120

$\alpha = 0.005$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	n_2
1	16211	19999	21615	22500	23056	23437	23715	23925	24091	24224	1
2	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.4	2
3	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	3
4	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	4
5	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	5
6	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	6
7	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	7
8	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	8
9	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	9
10	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	10
11	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	11
12	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	12
13	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	13
14	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	14
15	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	15
16	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	16
17	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	17
18	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	18
19	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	19
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	20
21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	21
22	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	22
23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	23
24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	24
25	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	25
26	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	26
27	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45	27
28	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	28
29	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38	29
30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	30
40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	40
60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	60
120	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	120

续表

n_2 \ n_1	10	15	20	24	30	40	60	120	n_2
1	24426	24630	24836	24940	25044	25148	25253	25359	1
2	199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5	2
3	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	3
4	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	4
5	13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	5
6	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	6
7	8.18	7.97	7.75	7.64	7.53	7.42	7.31	7.19	7
8	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	8
9	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	9
10	5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	10
11	5.24	5.05	4.86	4.76	4.65	4.55	4.45	4.34	11
12	4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	12
13	4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	13
14	4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	14
15	4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	15
16	4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22	16
17	3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	17
18	3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99	18
19	3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	19
20	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	20
21	3.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73	21
22	3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66	22
23	3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	23
24	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	24
25	3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	25
26	3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	26
27	3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	27
28	3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	28
29	3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	29
30	3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	30
40	2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	40
60	2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	60
120	2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.61	120