第五章　解线性方程组的直接方法

如何利用电子计算机来快速、有效地求解线性方程组的问题是数值线性代数研究的核心问题, 而且也是目前仍在继续研究的重大课题之一. 这是因为各种各样的科学与工程问题往往最终都要归结为一个线性方程组的求解问题. 例如结构分析、网络分析、大地测量、数据分析、最优化及非线性方程组和微分方程组数值解等, 都常遇到线性方程组的求解问题.

线性方程组的求解问题是一个古老的数学问题. 早在中国古代的《九章算术》中, 就已详细地载述了解线性方程组的消元法. 到了19世纪初, 西方也有了Gauss 消去法. 然而求解未知数多的大型线性方程组则是在20世纪中叶电子计算机问世后才成为可能.

求解线性方程组的数值方法大体上可分为直接法和迭代法两大类. 直接法是指在没有舍入误差的情况下经过有限次运算可求得方程组的精确解的方法. 因此, 直接法又称为精确法. 迭代法则是采取逐次逼近的方法, 亦即从一个初始向量出发,按照一定的计算格式, 构造一个向量的无穷序列, 其极限才是方程组的精确解, 只经过有限次运算得不到精确解.

这一章, 我们将主要介绍解线性方程组的一类最基本的直接法—Gauss 消去法. Gauss 消去法是目前求解中小规模线性方程组(即阶数不要太高, 例如不超过1000)最常用的方法, 它一般用于系数矩阵稠密(即矩阵的绝大多数元素都是非零的)而又没有任何特殊结构的线性方程组. 如若系数矩阵具有某种特殊形式, 则为了尽可能地减少计算量与存储量, 需采用其他专门的方法来求解, 限于篇幅, 本书不涉及这些专门的方法.

1　三角形方程组和三角分解

1.1　三角形方程组的解法

由于三角形方程组简单易于求解, 而且它又是用分解方法解一般线性方程组的基础, 所以我们首先考虑这种特殊类型的线性方程组的解法.

先考虑下三角形方程组

$$Ly = b, \tag{1.1}$$

这里$b = (b_1, \ldots, b_n)^{\mathrm{T}} \in \boldsymbol{R}^n$是已知的, $y = (y_1, \ldots, y_n)^{\mathrm{T}} \in \boldsymbol{R}^n$ 是未知的, 而$L = [l_{ij}] \in \boldsymbol{R}^{n\times n}$是已知的非奇异下三角阵, 即

$$L = \begin{bmatrix} l_{11} & & & & \\ l_{21} & l_{22} & & & \\ l_{31} & l_{32} & l_{33} & & \\ \vdots & \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{nn} \end{bmatrix},$$

而且$l_{ii} \neq 0$, $i = 1, 2, \ldots, n$. 由方程组(1.1)的第一个方程

$$l_{11}y_1 = b_1$$

得

$$y_1 = b_1/l_{11};$$

再由第二个方程

$$l_{21}y_1 + l_{22}y_2 = b_2$$

可得

$$y_2 = \big(b_2 - l_{21}y_1\big)/l_{22}.$$

一般地, 如果我们已求出$y_1, \ldots, y_{i-1}$, 就可根据(1.1)的第i个方程

$$l_{i1}y_i + l_{i2}y_2 + \cdots + l_{i,i-1}y_{i-1} + l_{ii}y_i = b_i$$

求出

$$y_i = \Big(b_i - \sum_{j=1}^{i-1} l_{ij}y_j\Big)\Big/l_{ii}.$$

这种解方程组(1.1)的方法称之为**前代法**. 如果在实际计算时将得到的y_i就存放在b_i所用的存储单元内, 并适当地调整一下运算次序, 可得如下算法:

算法1 (解下三角形方程组: 前代法)

for $j = 1 : n-1$
$\quad b(j) = b(j)/L(j,j)$
$\quad b(j+1:n) = b(j+1:n) - b(j)L(j+1:n, j)$
end
$b(n) = b(n)/L(n,n)$

该算法所需要的加、减、乘、除运算的次数为:

$$\sum_{i=1}^{n}(2i-1) = 2 \times \frac{n(n+1)}{2} - n = n^2,$$

即该算法的运算量为n^2.

再考虑上三角形方程组

$$Ux = y, \tag{1.2}$$

其中$U = [u_{ij}] \in \boldsymbol{R}^{n\times n}$是非奇异上三角阵, 即$u_{ij} = 0,\ i > j$, 而且$u_{ii} \neq 0,\ i = 1, 2, \ldots, n$, $y = (y_1, \ldots, y_n)^{\mathrm{T}} \in \boldsymbol{R}^n$是已知的, $x = (x_1, \ldots, x_n)^{\mathrm{T}} \in \boldsymbol{R}^n$ 是未知的. 这一方程组可以用所谓的**回代法**解之, 即从方程组的最后一个方程出发依次求出$x_n, x_{n-1}, \ldots, x_1$, 其计算公式为

$$x_i = \Big(y_i - \sum_{j=i+1}^{n} u_{ij}x_j\Big)\Big/u_{ii}, \quad i = n, n-1, \ldots, 1;$$

其具体算法如下:

算法2 (解上三角形方程组: 回代法)

for $j = n : -1 : 2$
$\quad y(j) = y(j)/U(j,j)$
$\quad y(1:j-1) = y(1:j-1) - y(j)U(1:j-1, j)$
end
$y(1) = y(1)/U(1,1)$

显然, 该算法的运算量亦为n^2.

对于一般的线性方程组

$$Ax = b, \tag{1.3}$$

其中$A \in \boldsymbol{R}^{n\times n}$和$b \in \boldsymbol{R}^n$是已知的, $x \in \boldsymbol{R}^n$是未知的, 如果我们能够将A分解为：$A = LU$, 即一个下三角阵L与一个上三角阵U的乘积, 那么原方程组的解x便可由下面两步得到：

(1) 用前代法解$Ly = b$得y;

(2) 用回代法解$Ux = y$得x.

所以对于求解一般的线性方程组来说, 关键是如何将A分解为一个下三角阵L与一个上三角阵U的乘积, 这正是我们本节的中心任务.

1.2 Gauss变换

欲把一个给定的矩阵A分解为一个下三角阵L与一个上三角阵U的乘积, 最自然的做法便是通过一系列的初等变换, 逐步将A约化为一个上三角阵, 而又能保证这些变换的乘积是一个下三角矩阵. 这可归结为：对于一个任意给定的向量$x \in \boldsymbol{R}^n$, 找一个尽可能简单的下三角矩阵, 使x经这一矩阵作用之后的第$k+1$至第n个分量均为零. 能够完成这一任务的最简单的下三角矩阵便是如下形式的初等下三角阵

$$L_k = I - l_k e_k^{\mathrm{T}},$$

其中

$$l_k = (0, \cdots, 0, l_{k+1,k}, \cdots, l_{nk})^{\mathrm{T}},$$

即

$$L_k = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{k+1,k} & 1 & \\ & & \vdots & & \ddots \\ & & -l_{n,k} & & 1 \end{bmatrix}.$$

这种类型的初等下三角矩阵称作**Gauss变换**, 而称向量l_k为**Gauss向量**.

对于一个给定的向量$x = (x_1, \ldots, x_n)^{\mathrm{T}} \in \boldsymbol{R}^n$, 我们有

$$L_k x = (x_1, \ldots, x_k, x_{k+1} - x_k l_{k+1,k}, \ldots, x_n - x_k l_{nk})^{\mathrm{T}}$$

由此立即可知,只要取

$$l_{ik} = \frac{x_i}{x_k}, \qquad i = k+1, \ldots, n,$$

便有

$$L_k x = (x_1, \cdots, x_k, 0, \cdots, 0)^{\mathrm{T}}.$$

当然, 这里我们要求$x_k \neq 0$.

Gauss变换L_k具有许多良好的性质. 例如, 它的逆是很容易求的. 因为$e_k^{\mathrm{T}} l_k = 0$, 所以

$$(I - l_k e_k^{\mathrm{T}})(I + l_k e_k^{\mathrm{T}}) = I - l_k e_k^{\mathrm{T}} l_k e_k^{\mathrm{T}} = I,$$

即

$$L_k^{-1} = I + l_k e_k^{\mathrm{T}}.$$

再如, 设$A \in \boldsymbol{R}^{n\times n}$, 则有

$$L_k A = (I - l_k e_k^{\mathrm{T}})A = A - l_k(e_k^{\mathrm{T}} A),$$

即Gauss变换作用于一个矩阵就相当于对该矩阵进行秩1修正.

1.3 三角分解的计算

假定$A \in \boldsymbol{R}^{n\times n}$, 三角分解是指分解$A = LU$, 其中$L \in \boldsymbol{R}^{n\times n}$为下三角矩阵, $U \in \boldsymbol{R}^{n\times n}$为上三角矩阵. 基于分解式的这种表达方式, 有时亦称三角分解为LU分解.

下面我们来讨论怎样利用Gauss变换来实现A的三角分解. 先来考察一个简单的例子. 设

$$A = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 10 \end{bmatrix}.$$

我们首先计算一个Gauss变换L_1使得$L_1 A$的第1列的后两个元素为0. 容易算出这样的L_1为

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix},$$

且有

$$L_1 A = \begin{bmatrix} 1 & 4 & 7 \\ 0 & -3 & -6 \\ 0 & -6 & -11 \end{bmatrix}.$$

然后再计算Gauss变换L_2使得$L_2(L_1 A)$的第2列的最后一个元素为0, 即取

$$L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix},$$

便有

$$L_2(L_1 A) = \begin{bmatrix} 1 & 4 & 7 \\ 0 & -3 & -6 \\ 0 & 0 & 1 \end{bmatrix}.$$

对于一般的n阶矩阵A, 在一定条件下, 我们也可以计算$n-1$个Gauss变换$L_1, \ldots, L_{n-1}$, 使得$L_{n-1}\cdots L_1 A$为上三角矩阵. 事实上, 记$A^{(0)} = A$, 并假定已求出$k-1$个Gauss变换$L_1, \ldots, L_{k-1} \in \boldsymbol{R}^{n\times n}$ $(k < n)$使得

$$A^{(k-1)} = L_{k-1}\cdots L_1 A = \begin{bmatrix} A_{11}^{(k-1)} & A_{12}^{(k-1)} \\ 0 & A_{22}^{(k-1)} \end{bmatrix},$$

其中$A_{11}^{(k-1)}$是$k-1$阶上三角阵, $A_{22}^{(k-1)}$为

$$A_{22}^{(k-1)} = \begin{bmatrix} a_{kk}^{(k-1)} & \cdots & a_{kn}^{(k-1)} \\ \vdots & \ddots & \vdots \\ a_{nk}^{(k-1)} & \cdots & a_{nn}^{(k-1)} \end{bmatrix}.$$

如果$a_{kk}^{(k-1)} \neq 0$, 则我们就又可以确定一个Gauss变换L_k, 使得$L_k A^{(k-1)}$中第k列的最后$n-k$个元素为0. 由前面所介绍的Gauss变换可知, 这样的L_k应为

$$L_k = I - l_k e_k^{\mathrm{T}},$$

其中

$$l_k = (0, \ldots, 0, l_{k+1,k}, \ldots, l_{nk})^{\mathrm{T}}, \quad l_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}, \quad i = k+1, \ldots, n.$$

因为$a_{kk}^{(k-1)} \neq 0$, 故L_k是唯一确定的. 对于这样确定的L_k, 我们有

$$A^{(k)} = L_k A^{(k-1)} = \begin{bmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ 0 & A_{22}^{(k)} \end{bmatrix} \begin{matrix} k \\ n-k \end{matrix},$$

其中$A_{11}^{(k)}$是k阶上三角阵. 从$k=1$出发, 如此进行$n-1$步, 最终所得矩阵$A^{(n-1)}$即为我们所要求的上三角形式. 现令

$$L = \left(L_{n-1} L_{n-2} \cdots L_1\right)^{-1}, \qquad U = A^{(n-1)},$$

则有$A = LU$. 这样只要证明了L是下三角矩阵, 则我们就已经实现了A的三角分解. 事实上, 根据Gauss变换的特点, 我们很容易证明L是一个$n \times n$的单位下三角阵, 即L是一个对角元均为1的下三角矩阵. 注意到对$j < i$有$e_j^{\mathrm{T}} l_i = 0$, 便有

$$\begin{aligned} L &= L_1^{-1} \cdots L_{n-1}^{-1} \\ &= (I + l_1 e_1^{\mathrm{T}})(I + l_2 e_2^{\mathrm{T}}) \cdots (I + l_{n-1} e_{n-1}^{\mathrm{T}}) \\ &= I + l_1 e_1^{\mathrm{T}} + \cdots + l_{n-1} e_{n-1}^{\mathrm{T}}, \end{aligned}$$

即L具有形式

$$L = I + [l_1, l_2, \cdots, l_{n-1}, 0] = \begin{bmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & l_{n3} & \cdots & 1 \end{bmatrix}.$$

由此可见, L不仅是一个单位下三角矩阵, 而且是非常容易得到的.

这种计算三角分解的方法称作**Gauss消去法**. 实际计算时, 我们还需弄清的是：当L_k作用于$A^{(k-1)}$后, $A^{(k-1)}$ 的哪些元素作了改变？以及作了怎样的改变？此外, L_k及$A^{(k)}$的元素又是怎样存储起来的？因为

$$A^{(k)} = L_k A^{(k-1)} = (I - l_k e_k^{\mathrm{T}}) A^{(k-1)} = A^{(k-1)} - l_k e_k^{\mathrm{T}} A^{(k-1)},$$

并注意到$e_k^{\mathrm{T}}A^{(k-1)}$是$A^{(k-1)}$的第$k$行以及$l_k$的前$k$个分量为0, 我们即知$A^{(k)}$和$A^{(k-1)}$的前$k$行元素相同, 而

$$\begin{aligned} a_{ik}^{(k)} &= 0, \quad i = k+1, \ldots, n, \\ a_{ij}^{(k)} &= a_{ij}^{(k-1)} - l_{ik}a_{kj}^{(k-1)}, \quad i,\ j = k+1, \ldots, n. \end{aligned}$$

$A^{(k)}$与L_k的存储是这样考虑的. $A^{(k-1)}$中第$k+1$行至第n行的元素在计算出$A^{(k)}$以后不再有用, 故可以用新计算出的$A^{(k)}$的元素冲掉$A^{(k-1)}$ 中相应位置上的元素. 此外, 由于$A^{(k)}$的第k列对角元以下的元素$a_{ik}^{(k)}$ $(i = k+1, \cdots, n)$为零, 无需存储, 故l_k中非零元即可存储在这些位置上. 例如一个4×4的矩阵A在经过二步消元后, 其形式为

$$\begin{bmatrix} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & a_{14}^{(0)} \\ l_{21} & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ l_{31} & l_{32} & a_{33}^{(2)} & a_{34}^{(2)} \\ l_{41} & l_{42} & a_{43}^{(2)} & a_{44}^{(2)} \end{bmatrix}.$$

综合上面的讨论, 可得如下算法:

算法3 (计算三角分解:Gauss 消去法)

for $k = 1 : n-1$
$A(k+1:n, k) = A(k+1:n, k)/A(k, k)$
$A(k+1:n, k+1:n) = A(k+1:n, k+1:n)$
$-A(k+1:n, k)A(k, k+1:n)$
end

该算法所需要的加、减、乘、除运算次数为

$$\begin{aligned} \sum_{k=1}^{n-1}\big((n-k) + 2(n-k)^2\big) &= \frac{n(n-1)}{2} + \frac{n(n-1)(2n-1)}{3} \\ &= \frac{2}{3}n^3 + O(n^2), \end{aligned}$$

即该算法的运算量为$\dfrac{2}{3}n^3$.

通常称Guass消去过程中的$a_{kk}^{(k-1)}$为主元. 显然, 当且仅当$a_{kk}^{(k-1)}(k = 1, \ldots, n-1)$均不为零时, 算法1.1.3才能进行到底. 那么自然要问：给定的矩阵A满足什么条件, 才能保证所有主元均不为零？这一问题可由下面定理回答.

定理1 主元$a_{ii}^{(i-1)}$ $(i = 1, \ldots, k)$均不为零的充分与必要条件是A 的i阶顺序主子阵A_i $(i = 1, \ldots, k)$都是非奇异的.

证明 对k用归纳法. 当$k = 1$时, $A_1 = a_{11}^{(0)}$, 定理显然成立. 假定定理直至$k-1$成立, 下面只需证明"若$A_1, \cdots, A_{k-1}$非奇异, 则A_k非奇异的充要条件是$a_{kk}^{(k-1)} \neq 0$"即可. 由归纳法假定知,

$a_{ii}^{(i-1)} \neq 0,\ i = 1, \ldots, k-1$. 因此, Guass消去过程至少可进行$k-1$步, 即可得到$k-1$个Gauss变换$L_1, \ldots, L_{k-1}$, 使得

$$A^{(k-1)} = L_{k-1} \cdots L_1 A = \begin{bmatrix} A_{11}^{(k-1)} & A_{12}^{(k-1)} \\ 0 & A_{22}^{(k-1)} \end{bmatrix}, \tag{1.4}$$

其中$A_{11}^{(k-1)}$是对角元为$a_{ii}^{(i-1)}$ $(i = 1, \ldots, k-1)$的上三角阵. 由此可知$A^{(k-1)}$的k阶顺序主子阵有如下形式

$$\begin{bmatrix} A_{11}^{(k-1)} & * \\ & a_{kk}^{(k-1)} \end{bmatrix}.$$

若将$L_1, \cdots, L_{k-1}$的k阶顺序主子阵分别记为$(L_1)_k, \ldots, (L_{k-1})_k$, 则由(1.4)及下三角阵的性质可知

$$(L_{k-1})_k (L_{k-2})_k \cdots (L_1)_k A_k = \begin{bmatrix} A_{11}^{(k-1)} & * \\ & a_{kk}^{(k-1)} \end{bmatrix}.$$

注意到L_i是单位下三角阵, 由此立即得到

$$\det A_k = a_{kk}^{(k-1)} \det A_{11}^{(k-1)},$$

从而有A_k非奇异当且仅当$a_{kk}^{(k-1)} \neq 0$. □

将定理1与前面的讨论相结合, 就得到了如下一个矩阵的三角分解存在的充分条件.

定理2 若$A \in \boldsymbol{R}^{n\times n}$的顺序主子阵$A_k \in \boldsymbol{R}^{k\times k}$ $(k = 1, \cdots, n-1)$均非奇异, 则存在唯一的单位下三角阵$L \in \boldsymbol{R}^{n\times n}$和上三角阵$U \in \boldsymbol{R}^{n\times n}$使得$A = LU$.

2 选主元三角分解

大家知道, 对于方程组$Ax = b$来说, 只要A非奇异, 方程组就存在唯一的解. 然而, A非奇异并不能保证其顺序主子阵A_i, $(i = 1, \ldots, n-1)$均非奇异. 因此, A非奇异并不能保证Gauss消去过程能够进行到底. 这样, 我们的问题自然便是, 怎样修改算法3才能使其适应于非奇异矩阵呢? 此外, 在算法3中计算l_{ik}时, 位于分母上的主元虽不为零但很小时, 是否会对算法产生不良影响呢? 若有影响, 该如何解决? 下面来看一个例子.

例1. 假定我们是在3位10进制的浮点数系下来解方程组

$$\begin{bmatrix} 0.001 & 1.00 \\ 1.00 & 2.00 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.00 \\ 3.00 \end{bmatrix}.$$

用算法1.1.3得

$$\widehat{L} = \begin{bmatrix} 1 & 0 \\ 1000 & 1 \end{bmatrix}, \qquad \widehat{U} = \begin{bmatrix} 0.001 & 1.00 \\ 0 & -1000 \end{bmatrix},$$

从而得该方程组的计算解为$\hat{x} = (0,\ 1)^{\mathrm{T}}$, 这与精确解

$$x = (1.002 \cdots,\ 0.998 \cdots)^{\mathrm{T}}$$

相差甚远.

上例中的问题是由小主元引起的. 当然, 如果用更高精度的计算机来计算, 可使计算解的精度提高. 然而仅以提高计算机的精度的方法去解决这个问题是不明智的, 因为计算机的精度毕竟是有限的. 事实上, 我们可以用下面的方法来避免小主元的出现.

如果交换例1的第一个与第二个方程的位置, 原方程组变为

$$\begin{bmatrix} 1.00 & 2.00 \\ 0.001 & 1.00 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3.00 \\ 1.00 \end{bmatrix}.$$

再用算法3在同样的数系下进行计算, 可得

$$\widehat{L} = \begin{bmatrix} 1 & 0 \\ 0.001 & 1 \end{bmatrix}, \qquad \widehat{U} = \begin{bmatrix} 1.00 & 2.00 \\ 0 & 1.00 \end{bmatrix},$$

进而得原方程组的计算解为$\hat{x} = (1.00, 1.00)^{\mathrm{T}}$, 这已与方程组的精确解相当接近了.

这种交换方程顺序的方法其实并不是解决小主元问题的唯一方法, 当然亦可通过交换未知向量x的分量在方程组中的顺序来解决. 这样, 如果出现小的主元, 我们就可以选择一个合适的元素, 交换A的行和列, 将此元素换到主元位置上. 例如在第k步中, 若$a_{kk}^{(k-1)}$ 太小, 并且选择$a_{pq}^{(k-1)} \neq 0$作为主元, 则我们需要先交换第k行和第p行, 再交换第k列和第q列, 从而将$a_{pq}^{(k-1)}$移至(k, k) 位置上, 消去过程用新的主元继续进行. 为了不打乱在消去过程中已经引入的零元素的分布,所选的$a_{pq}^{(k-1)}$的位置应该满足$p, q \geqslant k$.

为了下面叙述简单起见, 我们引入初等置换矩阵I_{pq}, 它是单位矩阵I的第p列与第q列交换所得到的矩阵, 即

$$I_{pq} = \big[e_1, \ldots, e_{p-1}, e_q, e_{p+1}, \ldots, e_{q-1}, e_p, e_{q+1}, \ldots, e_n\big].$$

用I_{pq}左乘矩阵A, 便交换了A的第p行与第q行；用I_{pq}右乘A便交换了A的第p列与第q列.

现在我们来看结合选主元的消去过程的具体做法. 假定消去过程已经进行了$k-1$步, 即已经确定了$k-1$个Gauss变换$L_1, \ldots, L_{k-1} \in \boldsymbol{R}^{n\times n}$和$2(k-1)$个初等置换矩阵

$$P_1, \ldots, P_{k-1} \in \boldsymbol{R}^{n\times n} \quad \text{和} \quad Q_1, \ldots, Q_{k-1} \in \boldsymbol{R}^{n\times n},$$

使得

$$\begin{aligned} A^{(k-1)} &= L_{k-1}P_{k-1}\cdots L_1P_1AQ_1\cdots Q_{k-1} \\ &= \begin{bmatrix} A_{11}^{k-1} & A_{12}^{(k-1)} \\ 0 & A_{22}^{(k-1)} \end{bmatrix}, \end{aligned}$$

其中$A_{11}^{(k-1)}$为$k-1$阶上三角阵, $A_{22}^{(k-1)}$为

$$A_{22}^{(k-1)} = \begin{bmatrix} a_{kk}^{(k-1)} & \cdots & a_{kn}^{(k-1)} \\ \vdots & & \vdots \\ a_{nk}^{(k-1)} & \cdots & a_{nn}^{(k-1)} \end{bmatrix}.$$

那么, 第k步是先在$A_{22}^{(k-1)}$中选择尽可能大的主元, 即选

$$\left|a_{pq}^{(k-1)}\right| = \max\left\{\left|a_{ij}^{(k-1)}\right| \,:\, k \leqslant i,\ j \leqslant n\right\}.$$

如果$a_{pq}^{(k-1)} = 0$, 则说明A的秩为$k-1$, 消去过程结束；否则, 交换$A^{(k-1)}$的第k行与第p行以及第k列与q列. 记交换后的$A_{22}^{(k-1)}$为

$$\widetilde{A}_{22}^{(k-1)} = \begin{bmatrix} \widetilde{a}_{kk}^{(k-1)} & \cdots & \widetilde{a}_{kn}^{(k-1)} \\ \vdots & & \vdots \\ \widetilde{a}_{nk}^{(k-1)} & \cdots & \widetilde{a}_{nn}^{(k-1)} \end{bmatrix}.$$

然后再计算Gauss变换$L_k = I - l_k e_k^{\mathrm{T}}$, 其中

$$l_k = (0, \ldots, 0, \widetilde{l}_{k+1,k}, \ldots, \widetilde{l}_{n,k})^{\mathrm{T}}, \quad \widetilde{l}_{ik} = \frac{\widetilde{a}_{ik}^{(k-1)}}{\widetilde{a}_{kk}^{(k-1)}}, \quad i = k+1, \cdots, n.$$

这样便有

$$A^{(k)} = L_k P_k A^{(k-1)} Q_k = \begin{bmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ 0 & A_{22}^{(k)} \end{bmatrix} \begin{matrix} k \\ n-k \end{matrix},$$

其中$A_{11}^{(k)}$是k阶上三角阵, $P_k = I_{kp}$, $Q_k = I_{kq} \in \boldsymbol{R}^{n\times n}$.

设全主元Gauss消去法进行到r步终止. 则我们得到初等变换阵P_k, Q_k 和初等下三角阵L_k, $k = 1, \ldots, r$, 使得

$$L_r P_r \cdots L_1 P_1 A Q_1 \cdots Q_r = U$$

为上三角矩阵. 令

$$\begin{aligned} Q &= Q_1 \cdots Q_r, \\ P &= P_r \cdots P_1, \\ L &= P(L_r P_r \cdots L_1 P_1)^{-1}, \end{aligned}$$

则有

$$PAQ = LU. \tag{2.1}$$

可以证明这样得到的L是一个单位下三角阵, 而且它的第k列对角线以下的元素是由构成L_k的Gauss向量l_k的分量作相应的排列而得到的. 因此, L的所有元素之模均不会超过1.

事实上, 由于

$$L = P_r \cdots P_2 L_1^{-1} P_2 L_2^{-1} \cdots P_r L_r^{-1},$$

所以, 若定义

$$L^{(1)} = L_1^{-1}, \quad L^{(k)} = P_k L^{(k-1)} P_k L_k^{-1}, \quad k = 2, \ldots r,$$

则有$L = L^{(r)}$, 而且可应用归纳法证明$L^{(k)}$具有如下的形状

$$L^{(k)} = \begin{bmatrix} L_{11}^{(k)} & 0 \\ L_{21}^{(k)} & I_{n-k} \end{bmatrix}, \qquad k = 1, \ldots, r, \tag{2.2}$$

其中$L_{11}^{(k)}$是所有元素之模均小于1的k阶单位下三角矩阵, $L_{21}^{(k)}$是所有元素之模均小于1的$(n-k)\times k$阶矩阵, I_{n-k}表示$n-k$阶单位矩阵.

由$L^{(1)}=L_1^{-1}$和L_1的定义立即知道$k=1$时(2.2)自然成立. 现假定对$k-1$已证(2.2)成立, 注意到$P_k=I_{kp}$而且$p\geqslant k$, 便有

$$L^{(k)}=P_kL^{(k-1)}P_kL_k^{-1}\begin{bmatrix} L_{11}^{(k-1)} & 0 \\ \widetilde{L}_{21}^{(k-1)} & \widetilde{L}_k^{-1} \end{bmatrix}, \tag{2.3}$$

其中$\widetilde{L}_{21}^{(k-1)}$是由$L_{21}^{(k-1)}$交换了第1行和第$p-k+1$行而得到的, 而

$$\widetilde{L}_k^{-1}=\begin{bmatrix} 1 & & & & \\ \widetilde{l}_{k+1,k} & 1 & & & \\ \widetilde{l}_{k+2,k} & 0 & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \widetilde{l}_{nk} & 0 & \cdots & 0 & 1 \end{bmatrix},$$

且$|\widetilde{l}_{ik}|=\left|\widetilde{a}_{ik}^{(k-1)}/\widetilde{a}_{kk}^{(k-1)}\right|\leqslant 1$, 故(2.3)表明(2.2)对$k$亦成立. 于是, 由归纳法原理知, (2.2)对$k=1,\ldots r$都成立, 从而$L=L^{(r)}$是一个所有元素之模均小于1的$n\times n$单位下三角矩阵.

此外, (2.3)亦给出了从$L^{(1)}=L_1^{-1}$出发逐步构造$L=L^{(r)}$的方法.

通常我们称(2.1)为A的全主元三角分解. 从(2.1)可以看出, 对A作全主元的三角分解, 就相当于先对A作好所有的行列交换得PAQ, 然后再对PAQ应用不选主元的Gauss消去法进行三角分解. 虽说这种做法并不能在实际中进行(因为我们是在计算过程中逐步得到每个主元, 从而知道它们是否很小或为零), 然而在理论上, 这是有用的. 因此我们将其总结为如下定理.

定理3 设$A\in\boldsymbol{R}^{n\times n}$, 则存在排列矩阵$P$, $Q\in\boldsymbol{R}^{n\times n}$, 以及单位下三角阵$L\in\boldsymbol{R}^{n\times n}$和上三角阵$U\in\boldsymbol{R}^{n\times n}$, 使得

$$PAQ=LU,$$

而且L的所有元素均满足$|l_{ij}|\leqslant 1$, U的非零对角元的个数正好等于矩阵A的秩.

以上讨论可总结为如下算法:

算法4 (计算全主元三角分解: 全主元Gauss消去法)

for $k=1:n-1$
确定p, q $(k\leqslant p,\ q\leqslant n)$使得
$|A(p,q)|=\max\{|A(i,j)|\ :\ i=k:n,\ j=k:n\}$
$A(k,1:n)\leftrightarrow A(p,1:n)$ (交换k行和p行)
$A(1:n,k)\leftrightarrow A(1:n,q)$ (交换k列和q列)
$u(k)=p$ (记录置换矩阵P_k)
$v(k)=q$ (记录置换矩阵Q_k)
if $A(k,k)\neq 0$
$A(k+1:n,k)=A(k+1:n,k)/A(k,k)$
$A(k+1:n,k+1:n)=A(k+1:n,k+1:n)$

$$-A(k+1:n,k)A(k,k+1:n)$$

else

stop (矩阵奇异)

end

end

虽然全主元Gauss消去法弥补了不选主元的Gauss消去法的不足, 但是选主元付出的代价也是极其昂贵的. 因为在A非奇异的情况下, 选主元必须进行

$$\sum_{k=1}^{n-1}(n-k+1)^2=\frac{1}{3}n^3+O(n^2)$$

次两两元素之间的比较和相应的逻辑判断, 这在计算机上是相当费时的. 为了尽可能地减少所进行的比较, 人们提出了列主元Gauss消去法. 这种方法与全主元Guass 消去法的差别仅在于, 第k步只在$A_{22}^{(k-1)}$的第k列上寻找模最大元, 即选

$$\left|a_{pk}^{(k-1)}\right|=\max\left\{\left|a_{ik}^{(k-1)}\right| \ : \ k\leqslant i\leqslant n\right\}.$$

这样, 第k步就不需进行列交换只进行行交换即可, 即有$P_k=I_{kp}$而$Q_k=I$. 而且从前面的讨论容易看出, 只要A非奇异, 则列主元Gauss消去法就可进行到底, 最终得到分解

$$PA=LU,$$

其中

$$\begin{aligned}&U=A^{(n-1)},\\&P=P_{n-1}\cdots P_1,\\&L=P(L_{n-1}P_{n-1}\cdots L_1P_1)^{-1}.\end{aligned}$$

这一分解通常称为列主元三角分解, 其具体算法如下.

算法5 (计算列主元三角分解: 列主元Gauss消去法)

for $k=1:n-1$

确定p $(k\leqslant p\leqslant n)$使得

$|A(p,k)|=\max\{|A(i,k)| \ : \ i=k:n\}$

$A(k,1:n)\leftrightarrow A(p,1:n)$ (交换k行和p行)

$u(k)=p$ (记录置换矩阵P_k)

if $A(k,k)\neq 0$

$A(k+1:n,k)=A(k+1:n,k)/A(k,k)$

$A(k+1:n,k+1:n)=A(k+1:n,k+1:n)$

$-A(k+1:n,k)A(k,k+1:n)$

else

stop (矩阵奇异)

end

end

注意, 这一算法与算法3一样, 也是将L和U分别存储在A的下三角部分和上三角部分.

设$A \in \boldsymbol{R}^{n\times n}$非奇异, 那么利用列主元Gauss消去法求解线性方程组$Ax = b$的计算过程就可按如下步骤进行：

(1) 用算法5计算A的列主元LU分解:$PA = LU$;

(2) 用算法1解下三角形方程组$Ly = Pb$;

(3) 用算法2解上三角形方程组$Ux = y$.

实际计算的经验和理论分析的结果表明, 列主元Gauss消去法与全主元Gauss消去法在数值稳定性方面完全可以媲美, 但它的运算量却大为减少. 因此, 它受到人们的青睐, 成为目前求解中小型稠密线性方程组最受欢迎的方法之一.

3 平方根法

平方根法又叫Cholesky分解法, 是求解对称正定线性方程组最常用的方法之一.

大家已经知道, 对于一般方阵, 为了消除LU分解的局限性和误差的过分积累, 而采用了选主元的方法. 但对于对称正定矩阵而言, 选主元却是完全不必要的.

设$A \in \boldsymbol{R}^{n\times n}$是对称正定的, 即$A$满足$A^{\mathrm{T}} = A$而且$x^{\mathrm{T}}Ax > 0$对一切的非零向量$x \in \boldsymbol{R}^n$成立. 此时, 由定理2容易推出

定理4 (Cholesky分解定理) 若$A \in \boldsymbol{R}^{n\times n}$对称正定, 则存在一个对角元均为正数的下三角阵$L \in \boldsymbol{R}^{n\times n}$, 使得

$$A = LL^{\mathrm{T}}.$$

上式中的L称作A的Cholesky因子.

证明 由于A对称正定蕴涵着A的全部主子阵均正定, 因此, 由定理1.1.2知, 存在一个单位下三角阵$\widetilde{L}$和一个上三角矩阵U, 使$A = \widetilde{L}U$. 令

$$D = \mathrm{diag}\,(u_{11}, \ldots, u_{nn}), \quad \widetilde{U} = D^{-1}U,$$

则有

$$\widetilde{U}^{\mathrm{T}}D\widetilde{L}^{\mathrm{T}} = A^{\mathrm{T}} = A = \widetilde{L}D\widetilde{U},$$

从而

$$\widetilde{L}^{\mathrm{T}}\widetilde{U}^{-1} = D^{-1}\widetilde{U}^{-T}\widetilde{L}D.$$

上式左边是一个单位上三角矩阵, 而右边是一个下三角矩阵, 故两边均为单位矩阵. 于是, $\widetilde{U} = \widetilde{L}^{\mathrm{T}}$, 从而$A = \widetilde{L}D\widetilde{L}^{\mathrm{T}}$. 由此即知, D的对角元均为正数. 令

$$L = \widetilde{L}\mathrm{diag}\,(\sqrt{u_{11}}, \ldots, \sqrt{u_{nn}}),$$

则$A = LL^{\mathrm{T}}$, 且L的对角元$l_{ii} = \sqrt{u_{ii}} > 0$, $i = 1, \ldots, n$. □

因此, 若线性方程组(1.3)的系数矩阵是对称正定的, 则我们自然可按如下的步骤求其解：

(1) 求A的Cholesky分解：$A = LL^{\mathrm{T}}$;

(2) 求解$Ly = b$得y;

(3) 求解$L^{\mathrm{T}}x = y$得x.

当然, 由定理4的证明可知, Cholesky分解可用不选主元的Gauss消去法来实现. 然而, 更简单而实用的方法是通过直接比较$A = LL^{\mathrm{T}}$两边的对应元素来计算L的. 设

$$L = \begin{bmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix}.$$

比较$A = LL^{\mathrm{T}}$两边对应的元素, 得关系式

$$a_{ij} = \sum_{p=1}^{j} l_{ip}l_{jp}, \qquad 1 \leqslant j \leqslant i \leqslant n. \tag{3.1}$$

首先, 由$a_{11} = l_{11}^2$, 得

$$l_{11} = \sqrt{a_{11}}.$$

再由$a_{i1} = l_{11}l_{i1}$, 得

$$l_{i1} = a_{i1}/l_{11}, \quad i = 1, \ldots, n.$$

这样便得到了矩阵L第一列元素. 假定已经算出L的前$k-1$列元素, 由

$$a_{kk} = \sum_{p=1}^{k} l_{kp}^2$$

得

$$l_{kk} = \left(a_{kk} - \sum_{p=1}^{k-1} l_{kp}^2\right)^{\frac{1}{2}}. \tag{3.2}$$

再由

$$a_{ik} = \sum_{p=1}^{k-1} l_{ip}l_{kp} + l_{ik}l_{kk}, \quad i = k+1, \ldots, n,$$

得

$$l_{ik} = \left(a_{ik} - \sum_{p=1}^{k-1} l_{ip}l_{kp}\right) \Big/ l_{kk}, \quad i = k+1, \ldots, n. \tag{3.3}$$

这样便又求出了L的第k列元素. 这种方法称为平方根法. 当然, 亦可按行来逐次计算L. 由于A的元素a_{ij}被用来计算出l_{ij}以后不再使用, 所以可将L的元素存储在A的对应位置上. 这样我们就得到如下算法.

算法6 (计算Cholesky分解: 平方根法)

for $k = 1 : n$
$A(k,k) = \sqrt{A(k,k)}$
$A(k+1:n, k) = A(k+1:n, k)/A(k,k)$

for $j = k+1 : n$

$A(j:n, j) = A(j:n, j) - A(j:n, k)A(j, k)$

end

end

容易算出, 该算法的运算量为$\frac{1}{3}n^3$, 仅是Gauss消去法运算量的一半.

由公式(3.2)可以看出, 用平方根法解对称正定线性方程组时, 计算L 的对角元素l_{ii}需用到开方运算. 为了避免开方, 我们可求A 之如下形式的分解

$$A = LDL^{\mathrm{T}}, \tag{3.4}$$

其中L是单位下三角矩阵, D是对角元素均为正数的对角矩阵. 这一分解称作$\mathrm{LDL}^{\mathrm{T}}$分解, 是Cholesky分解的变形. 比较(3.4)两边对应元素, 得

$$a_{ij} = \sum_{k=1}^{j-1} l_{ik} d_k l_{jk} + l_{ij} d_j, \quad 1 \leqslant j \leqslant i \leqslant n.$$

由此可得确定l_{ij}和d_j的计算公式如下：

$$v_k = d_k l_{jk}, \quad k = 1, \ldots, j-1,$$

$$d_j = a_{jj} - \sum_{k=1}^{j-1} l_{jk} v_k,$$

$$l_{ij} = \Big(a_{ij} - \sum_{k=1}^{j-1} l_{ik} v_k\Big) \Big/ d_j, \quad i = j+1, \ldots, n,$$

这里$j = 1, 2, \ldots, n$. 上述这种确定A的分解的方法称作改进的平方根方法. 实际计算时, 是将L的严格下三角元素存储在A的对应位置上, 而将D的对角元存储在A的对应的对角位置上. 这样我们就得到如下的实用算法.

算法7 (计算$\mathrm{LDL}^{\mathrm{T}}$分解: 改进的平方根法)

for $j = 1 : n$

for $i = 1 : j-1$

$v(i) = A(j, i)A(i, i)$

end

$A(j, j) = A(j, j) - A(j, 1 : j-1)v(1 : j-1)$

$A(j+1 : n, j) =$

$\big(A(j+1 : n, j) - A(j+1 : n, 1 : j-1)v(1 : j-1)\big) \big/ A(j, j)$

end

这一算法的运算量与算法1.3.1一样也是$\frac{1}{3}n^3$, 而且还不需开方运算.

一但求得A的$\mathrm{LDL}^{\mathrm{T}}$分解, 只需再解如下两个三角方程组

$$Ly = b \quad \text{和} \quad DL^{\mathrm{T}} x = y$$

即可求得线性方程组(1.3)的解. 利用这种方法求解对称正定线性方程组所需的运算量仅是Gauss消去法的一半, 而且还不需要选主元. 此外, Cholesky 分解的计算过程是稳定的. 事实上, 由关系式

$$a_{ii} = \sum_{k=1}^{i} l_{ik}^2$$

得

$$|l_{ij}| \leqslant \sqrt{a_{ii}}. \tag{3.5}$$

上式说明Cholesky分解中的量l_{ij}能够得以控制, 因此其计算过程是稳定的.

例2 设

$$A = \begin{bmatrix} 4 & -2 & 4 & 2 \\ -2 & 10 & -2 & -7 \\ 4 & -2 & 8 & 4 \\ 2 & -7 & 4 & 7 \end{bmatrix}, \qquad b = \begin{bmatrix} 8 \\ 2 \\ 16 \\ 6 \end{bmatrix}.$$

利用改进的平方根法, 得

$$\begin{aligned}
&d_1 = a_{11} = 4, \\
&l_{21} = a_{21}/d_1 = \frac{-2}{4} = -\frac{1}{2}, \\
&l_{31} = 1, \\
&l_{41} = \frac{1}{2}, \\
&d_2 = a_{22} - l_{21}^2 d_1 = 10 - 1 = 9, \\
&l_{32} = \big(a_{32} - l_{31} d_1 l_{21}\big)/d_2 = \frac{-2 - (-0.5) \times 4}{9} = 0, \\
&l_{42} = -\frac{2}{3}, \\
&d_3 = a_{33} - l_{31}^2 d_1 - l_{32}^2 d_2 = 8 - 4 = 4, \\
&l_{43} = \big(a_{43} - l_{41} d_1 l_{31} - l_{42} d_2 l_{32}\big)/d_3 = \frac{4-2}{4} = \frac{1}{2}, \\
&d_4 = a_{44} - l_{41}^2 d_1 - l_{42}^2 d_2 - l_{43}^2 d_3 = 7 - 1 - 4 - 1 = 1.
\end{aligned}$$

这样, 我们便得到A的$\mathrm{LDL}^{\mathrm{T}}$分解的因子为

$$L = \begin{bmatrix} 1 & & & \\ -\frac{1}{2} & 1 & & \\ 1 & 0 & 1 & \\ \frac{1}{2} & -\frac{2}{3} & \frac{1}{2} & 1 \end{bmatrix}, \quad D = \mathrm{diag}\,(4,\, 9,\, 4,\, 1).$$

于是, 欲解方程组$Ax = b$, 可先解$Lz = b$, 得$z = (8,\, 6,\, 8,\, 2)^{\mathrm{T}}$; 再解$Dy = z$, 得$y = (2,\, \frac{2}{3},\, 2,\, 2)^{\mathrm{T}}$; 最后解$L^{\mathrm{T}}x = y$, 得$x = (1,\, 2,\, 1,\, 2)^{\mathrm{T}}$.

前面我们详细地介绍了如何用Gauss消去法求解线性方程组, 但从数值计算的角度来看, 这是不够的. 这是因为在实际计算时, 我们不仅要面对如何求解的问题, 而且也要面对数据不精确的问题和机器的有限精度问题. 只有在讨论了后面这两个问题, 我们才可能知道, 所求得的解是否可靠. 而要讨论这两个问题, 我们需要用范数来描述向量与矩阵的扰动的大小等概念. 所以在这一章中,我们首先介绍向量范数和矩阵范数的概念及其基本性质；然后对线性方程组进行敏度分析, 讨论舍入误差问题, 并对列主元Gauss消去法进行的详细舍入误差分析; 最后介绍一种估计计算解的精度的实用方法以及改进其计算精度的迭代方法.

4 向量范数和矩阵范数

4.1 向量范数

向量范数的概念是复数模的概念的自然推广, 其定义如下:

定义1 一个从$\boldsymbol{R}^n$到$\boldsymbol{R}$的非负函数$\|\cdot\|$ 叫做$\boldsymbol{R}^n$上的**向量范数**, 如果它满足:
(1)正定性：对所有的$x \in \boldsymbol{R}^n$有$\|x\| \geqslant 0$, 而且$\|x\| = 0$当且仅当$x = 0$;
(2)齐次性：对所有的$x \in \boldsymbol{R}^n$和$\alpha \in \mathbf{R}$有$\|\alpha x\| = |\alpha| \, \|x\|$;
(3)三角不等式：对所有的$x, \, y \in \boldsymbol{R}^n$有$\|x + y\| \leqslant \|x\| + \|y\|$.

由范数的性质(2)和(3)容易导出, 对任意的$x, \, y \in \boldsymbol{R}^n$有

$$\big| \, \|x\| - \|y\| \, \big| \leqslant \|x - y\| \leqslant \max_{1 \leqslant i \leqslant n} \|e_i\| \sum_{i=1}^{n} |x_i - y_i|.$$

由此即知, $\|\cdot\|$作为$\boldsymbol{R}^n$上的实函数是连续的.

最常用的向量范数是p**范数**(亦称**Hölder范数**):

$$\|x\|_p = \big(|x_1|^p + \cdots + |x_n|^p \big)^{\frac{1}{p}}, \qquad p \geqslant 1.$$

其中$p = 1, 2, \infty$是最重要的, 即

$$\begin{aligned} \|x\|_1 &= |x_1| + \cdots + |x_n|, \\ \|x\|_2 &= \big(|x_1|^2 + \cdots + |x_n|^2 \big)^{\frac{1}{2}} = \sqrt{x^{\mathrm{T}} x}, \\ \|x\|_\infty &= \max\{ |x_i| \ : \ i = 1, 2, \ldots, n \}. \end{aligned}$$

它们分别叫做**1范数、2范数**和∞**范数**. 这三个范数的正定性与齐次性是显然的, 而且也容易证明$\|\cdot\|_1$ 和$\|\cdot\|_\infty$满足三角不等式. 对2范数要证明三角不等式成立, 需要用到Cauchy-Schwartz不等式

$$|x^{\mathrm{T}} y| \leqslant \|x\|_2 \|y\|_2, \qquad x, \, y \in \boldsymbol{R}^n,$$

这个不等式是Hölder不等式

$$|x^{\mathrm{T}} y| \leqslant \|x\|_p \|y\|_q, \qquad \frac{1}{p} + \frac{1}{q} = 1$$

的特殊情形. 事实上, 利用Cauchy-Schwartz不等式, 我们有

$$\begin{aligned}\|x+y\|_2^2 &= (x+y)^{\mathrm{T}}(x+y) = \|x\|_2^2 + x^{\mathrm{T}}y + y^{\mathrm{T}}x + \|y\|_2^2 \\ &\leqslant \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^{\mathrm{T}} = (\|x\|_2^2 + \|y\|_2^2)^2,\end{aligned}$$

由此即知2范数满足三角不等式.

尽管在$\boldsymbol{R}^n$上可以引进各种各样的范数, 但在下面定理所述的意义下所有这些范数都是等价的.

定理5 设$\|\cdot\|_\alpha$和$\|\cdot\|_\beta$是$\boldsymbol{R}^n$上任意二个范数. 则存在正常数c_1和c_2, 使得对一切$x \in \boldsymbol{R}^n$有

$$c_1\|x\|_\alpha \leqslant \|x\|_\beta \leqslant c_2\|x\|_\alpha.$$

这一定理的证明较为复杂这里不再给出, 有兴趣的读者可参看有关的参考书.

例如, 对于$\|\cdot\|_1$, $\|\cdot\|_2$和$\|\cdot\|_\infty$ 这三种常用的向量范数, 有

$$\begin{aligned}\|x\|_2 &\leqslant \|x\|_1 \leqslant \sqrt{n}\|x\|_2, \\ \|x\|_\infty &\leqslant \|x\|_2 \leqslant \sqrt{n}\|x\|_\infty, \\ \|x\|_\infty &\leqslant \|x\|_1 \leqslant n\|x\|_\infty.\end{aligned}$$

利用定理5可证如下的重要结果.

定理6 设$x_k \in \boldsymbol{R}^n$. 则$\lim\limits_{k\to\infty}\|x_k - x\| = 0$ 的充要条件是$\lim\limits_{k\to\infty}|x_i^{(k)} - x_i| = 0$, $i = 1, \ldots, n$, 即向量序列的范数收敛等价于分量收敛.

证明留作练习.

4.2 矩阵范数

若我们用E_{ij}表示在(i,j)位置上的元素是1, 其余元素都是零的$n \times n$矩阵, 则E_{ij}是线性无关的, 而且任一个$n \times n$矩阵$A = [a_{ij}]$都可表为

$$A = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}E_{ij}.$$

这也就是说, 全体$n \times n$实矩阵构成的空间的维数是n^2. 因此, $\boldsymbol{R}^{n\times n}$亦可看作一个$n^2$维的向量空间. 这样, 我们自然想到将向量范数的概念直接推广到矩阵上. 然而这样推广的缺点是未考虑到矩阵的乘法运算. 因而实用的矩阵范数的定义是按如下的方式定义的.

定义2 一个从$\boldsymbol{R}^{n\times n}$到$\boldsymbol{R}$的非负函数$\|\cdot\|$叫做$\boldsymbol{R}^{n\times n}$上的**矩阵范数**, 如果它满足:

(1)正定性：对所有的$A \in \boldsymbol{R}^{n\times n}$有$\|A\| \geqslant 0$, 而且$\|A\| = 0$当且仅当$A = 0$;

(2)齐次性：对所有的$A \in \boldsymbol{R}^{n\times n}$和$\alpha \in \boldsymbol{R}$有$\|\alpha A\| = |\alpha|\,\|A\|$;

(3)三角不等式：对所有的$A, B \in \boldsymbol{R}^{n\times n}$有$\|A + B\| \leqslant \|A\| + \|B\|$;

(4)相容性：对所有的$A, B \in \boldsymbol{R}^{n\times n}$有$\|AB\| \leqslant \|A\|\,\|B\|$.

因为$\boldsymbol{R}^{n\times n}$上的矩阵范数自然可以看作是$\boldsymbol{R}^{n^2}$上的向量范数, 所以矩阵范数具有向量范数的一切性质. 例如, 有

(i) $\boldsymbol{R}^{n\times n}$ 上的任意两个矩阵范数是等价的.

(ii) 矩阵序列的范数收敛等价于元素收敛, 即

$$\lim_{k\to\infty}\|A_k - A\| = 0 \Longleftrightarrow \lim_{k\to\infty} a_{ij}^{(k)} = a_{ij},\ i,j = 1,\ldots,n,$$

其中$A_k = \left[a_{ij}^{(k)}\right] \in \boldsymbol{R}^{n\times n}$.

矩阵与向量的乘积在矩阵计算中经常出现, 因此我们自然希望矩阵范数与向量范数之间最好具有某种协调性. 若将向量看作是矩阵的特殊情形, 那么由矩阵范数的相容性, 我们便得到了这种协调性, 即矩阵范数与向量范数的相容性.

定义3 若矩阵范数$\|\cdot\|_M$和向量范数$\|\cdot\|_v$满足

$$\|Ax\|_v \leqslant \|A\|_M\|x\|_v, \qquad A\in\boldsymbol{R}^{n\times n},\ x\in\boldsymbol{R}^n,$$

则称矩阵范数$\|\cdot\|_M$与向量范数$\|\cdot\|_v$是**相容**的.

在本书中, 如果没有特别说明, 凡同时涉及到向量范数和矩阵范数的均假定它们是相容的. 事实上, 对任意给定的向量范数, 我们都可以构造一个与该向量范数相容的矩阵范数, 其方法如下面的定理所述.

定理7 设$\|\cdot\|$是$\boldsymbol{R}^n$上的一个向量范数. 若定义

$$|\!|\!|A|\!|\!| = \max_{\|x\|=1}\|Ax\|, \quad A\in\boldsymbol{R}^{n\times n}, \tag{4.1}$$

则$|\!|\!|\cdot|\!|\!|$ 是$\boldsymbol{R}^{n\times n}$ 上的一个矩阵范数.

证明 由于$\mathcal{D} = \{x\in\boldsymbol{R}^n\ :\ \|x\| = 1\}$是$\boldsymbol{R}^n$中有界闭集, 而$\|\cdot\|$又是$\boldsymbol{R}^n$ 上的连续函数, 所以存在向量$x_0\in\mathcal{D}$使得$\|Ax_0\| = \max\limits_{x\in\mathcal{D}}\|Ax\|$. 因此, 由(4.1) 所定义的$|\!|\!|\cdot|\!|\!|$是有意义的.

其次, 对任意$x\in\boldsymbol{R}^n$, $x\neq 0$, 由(4.1)知

$$\frac{\|Ax\|}{\|x\|} = \left\|A\frac{x}{\|x\|}\right\| \leqslant |\!|\!|A|\!|\!|,$$

从而有

$$\|Ax\| \leqslant |\!|\!|A|\!|\!|\,\|x\|, \qquad x\in\boldsymbol{R}^n. \tag{4.2}$$

下面对任取的A, $B\in\boldsymbol{R}^{n\times n}$, 证明$|\!|\!|\cdot|\!|\!|$ 满足矩阵范数的四条性质.

(1) 正定性. 设$A\neq 0$, 不妨设A的第i列非零, 即$Ae_i\neq 0$. 由(4.2) 和向量范数的正定性, 有

$$0 < \|Ae_i\| \leqslant |\!|\!|A|\!|\!|\,\|e_i\|,$$

从而$|\!|\!|A|\!|\!| > 0$.

(2) 齐次性. 任取$\alpha\in\boldsymbol{R}$, 有

$$|\!|\!|\alpha A|\!|\!| = \max_{\|x\|=1}\|\alpha Ax\| = |\alpha|\max_{\|x\|=1}\|Ax\| = |\alpha|\,|\!|\!|A|\!|\!|.$$

(3) 三角不等式. 设x满足$\|x\| = 1$使得$\|(A+B)x\| = \|\!|A+B|\!\|$, 则由向量范数的三角不等式和(4.2), 有

$$\begin{aligned}\|\!|A+B|\!\| = \|(A+B)x\| &\leqslant \|Ax\| + \|Bx\| \\ &\leqslant \|\!|A|\!\|\,\|x\| + \|\!|B|\!\|\,\|x\| = \|\!|A|\!\| + \|\!|B|\!\|.\end{aligned}$$

(4)相容性. 设$x \in \boldsymbol{R}^n$满足$\|x\| = 1$使得$\|ABx\| = \|\!|AB|\!\|$, 则由(4.2), 有

$$\|\!|AB|\!\| = \|ABx\| \leqslant \|\!|A|\!\|\,\|Bx\| \leqslant \|\!|A|\!\|\,\|\!|B|\!\|\,\|x\| = \|\!|A|\!\|\,\|\!|B|\!\|.$$

这样, 我们就完成了定理2.1.3的证明. □

由(4.1)定义的矩阵范数$\|\!|\cdot|\!\|$称为**从属于向量范数**$\|\cdot\|$**的矩阵范数**, 也称其为由向量范数$\|\cdot\|$诱导出的**算子范数**.

在上面的讨论中, 为了不致引起混淆, 我们将算子范数记作$\|\!|\cdot|\!\|$. 今后为了简单起见, 我们仍将其记作$\|\cdot\|$. 此外, 还需指出的是, 为了讨论方便, 我们在本节中仅考虑了方阵的范数, 但其大部分结论都适宜于长方阵的情形.

根据定理7, 我们可从$\boldsymbol{R}^n$上最常用的p范数得到$\boldsymbol{R}^{n\times n}$上的算子范数$\|\cdot\|_p$:

$$\|A\|_p = \max_{\|x\|_p=1} \|Ax\|_p, \quad A \in \boldsymbol{R}^{n\times n}.$$

对于$p = 1, 2, \infty$所对应的算子范数, 我们有

定理8 设$A = [a_{ij}] \in \boldsymbol{R}^{n\times n}$, 则有

$$\begin{aligned}\|A\|_1 &= \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|, \\ \|A\|_\infty &= \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|, \\ \|A\|_2 &= \sqrt{\lambda_{\max}(A^{\mathrm{T}}A)},\end{aligned}$$

其中$\lambda_{\max}(A^{\mathrm{T}}A)$表示$A^{\mathrm{T}}A$的最大特征值.

证明 $A = 0$时定理显然成立. 因此在下面的证明中总假定$A \neq 0$. 对于1范数, 将给定的$A \in \boldsymbol{R}^{n\times n}$按列分块为$A = [a_1, \cdots, a_n]$, 并记$\delta = \|a_{j_0}\|_1 = \max\limits_{1\leqslant j\leqslant n} \|a_j\|_1$. 则对任意的$x \in \boldsymbol{R}^n$满足$\|x\|_1 = \sum\limits_{i=1}^{n} |x_i| = 1$有

$$\begin{aligned}\|Ax\|_1 = \left\|\sum_{j=1}^{n} x_j a_j\right\|_1 &\leqslant \sum_{j=1}^{n} |x_j|\,\|a_j\|_1 \\ &\leqslant \left(\sum_{j=1}^{n} |x_j|\right) \max_{1\leqslant j\leqslant n} \|a_j\|_1 = \|a_{j_0}\|_1 = \delta.\end{aligned}$$

此外, 若取e_{j_0}为n阶单位矩阵的第j_0列, 则有$\|e_{j_0}\|_1 = 1$, 而且

$$\|Ae_{j_0}\|_1 = \|a_{j_0}\|_1 = \delta.$$

因此, 我们有

$$\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1 = \delta = \max_{1\leqslant j\leqslant n} \|a_j\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|.$$

对于∞范数, 记$\eta = \max\limits_{1\leqslant j\leqslant n} \sum\limits_{j=1}^{n} |a_{ij}|$. 则对任一$x \in \boldsymbol{R}^n$满足$\|x\|_\infty = 1$有

$$\begin{aligned} \|Ax\|_\infty &= \max_{1\leqslant i\leqslant n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leqslant \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|\,|x_j| \\ &\leqslant \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}| = \eta. \end{aligned}$$

设A的第k行的1范数最大, 即$\eta = \sum\limits_{j=1}^{n} |a_{kj}|$. 令

$$\widetilde{x} = \big(\operatorname{sgn}(a_{k1}), \cdots, \operatorname{sgn}(a_{kn})\big)^{\mathrm{T}},$$

则$A \neq 0$蕴含着$\|\widetilde{x}\|_\infty = 1$, 而且容易证明$\|A\widetilde{x}\|_\infty = \eta$. 这样, 我们就已经证明了

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

对于2范数, 应有

$$\begin{aligned} \|A\|_2 &= \max_{\|x\|_2=1} \|Ax\|_2 = \max_{\|x\|_2=1} [(Ax)^{\mathrm{T}} Ax)]^{\frac{1}{2}} \\ &= \max_{\|x\|_2=1} [x^{\mathrm{T}} (A^{\mathrm{T}} A) x]^{\frac{1}{2}}. \end{aligned}$$

注意, $A^{\mathrm{T}}A$ 是半正定的对称阵. 设其特征值为

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n \geqslant 0,$$

以及其对应的正交规范特征向量为$v_1, \cdots, v_n \in \boldsymbol{R}^n$. 则对任一满足$\|x\|_2 = 1$的向量$x \in \boldsymbol{R}^n$有

$$x = \sum_{i=1}^{n} \alpha_i v_i \quad \text{和} \quad \sum_{i=1}^{n} \alpha_i^2 = 1.$$

于是, 有

$$x^{\mathrm{T}} A^{\mathrm{T}} A x = \sum_{i=1}^{n} \lambda_i \alpha_i^2 \leqslant \lambda_1.$$

另一方面, 若取$x = v_1$, 则有

$$x^{\mathrm{T}} A^{\mathrm{T}} A x = v_1^{\mathrm{T}} A^{\mathrm{T}} A v_1 = v_1^{\mathrm{T}} \lambda_1 v_1 = \lambda_1.$$

所以

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \sqrt{\lambda_1} = \sqrt{\lambda_{\max}(A^{\mathrm{T}} A)}.$$

于是, 定理得证. □

基于定理8, 我们通常分别称矩阵的1范数、∞范数和2 范数为**列和范数、行和范数**和**谱范数**. 此外, 从这一定理容易看出, 矩阵列和范数与行和范数是很容易计算的, 而矩阵的谱范数就不适宜于实际计算, 它需要计算$A^{\mathrm{T}}A$的最大特征值. 但是, 谱范数所具有的许多好的性质, 使它在理论研究中很有用处. 下面的定理列举了谱范数的几条最常用的性质.

定理9 设$A \in \boldsymbol{R}^{n\times n}$, 则

(1) $\|A\|_2 = \max\{|y^{\mathrm{T}}Ax| \ : \ x, y \in \boldsymbol{C}^n, \ \|x\|_2 = \|y\|_2 = 1\}$;

(2) $\|A^{\mathrm{T}}\|_2 = \|A\|_2 = \sqrt{\|A^{\mathrm{T}}A\|_2}$;

(3) 对任意的正交矩阵U和V有, $\|UA\|_2 = \|AV\|_2 = \|A\|_2$.

证明留作练习.

此外, 在$\boldsymbol{R}^{n\times n}$上的另一个常用且易于计算的矩阵范数为

$$\|A\|_F = \Big(\sum_{i,j=1}^{n} |a_{ij}|^2\Big)^{\frac{1}{2}},$$

通常称作**Frobenius范数**, 它是向量2范数的自然推广.

作为本节的结束, 我们来证明几个本书经常使用的与范数有关的重要结果. 由于这些结果与谱半径有关, 因此, 为了讨论方便, 我们下面的讨论将在复数范围内展开. 大家容易看出, 本节前面所讲的所有概念与结果都可毫无困难地推广到复空间上.

定义4 设$A \in \boldsymbol{C}^{n\times n}$, 则称

$$\rho(A) = \max\{|\lambda| \ : \ \lambda \in \lambda(A)\}$$

为A的**谱半径**, 这里$\lambda(A)$表示A的特征值的全体.

谱半径与矩阵范数之间有如下关系:

定理10 设$A \in \boldsymbol{C}^{n\times n}$, 则有

(1) 对$\boldsymbol{C}^{n\times n}$上的任意矩阵范数$\|\cdot\|$, 有

$$\rho(A) \leqslant \|A\|;$$

(2) 对任给的$\varepsilon > 0$, 存在$\boldsymbol{C}^{n\times n}$上的算子范数$\|\cdot\|$, 使得

$$\|A\| \leqslant \rho(A) + \varepsilon.$$

证明 (1) 设$x \in \boldsymbol{C}^n$满足

$$x \neq 0, \quad Ax = \lambda x, \quad |\lambda| = \rho(A),$$

则有

$$\rho(A)\|xe_1^{\mathrm{T}}\| = \|\lambda xe_1^{\mathrm{T}}\| = \|Axe_1^{\mathrm{T}}\| \leqslant \|A\| \, \|xe_1^{\mathrm{T}}\|,$$

从而有

$$\rho(A) \leqslant \|A\|.$$

(2) 由Jordan 分解定理知, 存在非奇异矩阵$X \in \boldsymbol{C}^{n\times n}$, 使得

$$X^{-1}AX = \begin{bmatrix} \lambda_1 & \delta_1 & & & \\ & \lambda_2 & \delta_2 & & \\ & & \ddots & \ddots & \\ & & & \lambda_{n-1} & \delta_{n-1} \\ & & & & \lambda_n \end{bmatrix},$$

其中$\delta_i = 1$或0. 对于任意给定的$\varepsilon > 0$, 令

$$D_\varepsilon = \operatorname{diag}(1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{n-1}),$$

则有

$$D_\varepsilon^{-1}X^{-1}AXD_\varepsilon = \begin{bmatrix} \lambda_1 & \varepsilon\delta_1 & & & \\ & \lambda_2 & \varepsilon\delta_2 & & \\ & & \ddots & \ddots & \\ & & & \lambda_{n-1} & \varepsilon\delta_{n-1} \\ & & & & \lambda_n \end{bmatrix}.$$

现在定义

$$\|G\|_\varepsilon = \|D_\varepsilon^{-1}X^{-1}GXD_\varepsilon\|_\infty, \qquad G \in \boldsymbol{C}^{n\times n},$$

则容易验证这样定义的函数$\|\cdot\|_\varepsilon$是由如下定义的向量范数

$$\|x\|_{XD_\varepsilon} = \|(XD_\varepsilon)^{-1}x\|_\infty, \qquad x \in \boldsymbol{C}^n$$

诱导出的算子范数, 而且有

$$\|A\|_\varepsilon = \|D_\varepsilon^{-1}X^{-1}AXD_\varepsilon\|_\infty = \max_{1\leqslant i\leqslant n}(|\lambda_i| + |\varepsilon\delta_i|) \leqslant \rho(A) + \varepsilon,$$

其中假定$\delta_n = 0$. □

定理11 设$A \in \boldsymbol{C}^{n\times n}$, 则

$$\lim_{k\to\infty} A^k = 0 \Longleftrightarrow \rho(A) < 1.$$

证明 必要性 设$\lim\limits_{k\to\infty} A^k = 0$, 并假定$\lambda \in \lambda(A)$满足$\rho(A) = |\lambda|$. 由于对任意的$k$有$\lambda^k \in \lambda(A^k)$, 故由定理10知,

$$\rho(A)^k = |\lambda|^k \leqslant \rho(A^k) \leqslant \|A^k\|_2$$

对一切k成立, 从而必有$\rho(A) < 1$.

充分性 设$\rho(A) < 1$. 则由定理10知, 必有算子范数$\|\cdot\|$, 使得$\|A\| < 1$, 从而

$$0 \leqslant \|A^k\| \leqslant \|A\|^k \longrightarrow 0, \quad k \longrightarrow \infty,$$

于是$\lim\limits_{k\to\infty} A^k = 0$. □

利用定理11容易证明如下的重要结果:

定理12 设$A \in \boldsymbol{C}^{n\times n}$, 则有

(1) $\sum\limits_{k=0}^{\infty} A^k$收敛的充分与必要条件是$\rho(A) < 1$;

(2) 当$\sum\limits_{k=0}^{\infty} A^k$收敛时, 有

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

而且存在$\boldsymbol{C}^{n\times n}$上的算子范数$\|\cdot\|$, 使得

$$\left\|(I - A)^{-1} - \sum_{k=0}^{m} A^k\right\| \leqslant \frac{\|A\|^{m+1}}{1 - \|A\|}$$

对一切的自然数m 成立.

由这一定理立即得到如下常用的结果.

推论1 设$\|\cdot\|$是$\boldsymbol{C}^{n\times n}$上的一个满足条件$\|I\| = 1$的矩阵范数, 并假定$A \in \boldsymbol{C}^{n\times n}$满足$\|A\| < 1$, 则$I - A$可逆且有

$$\left\|(I - A)^{-1}\right\| \leqslant \frac{1}{1 - \|A\|}.$$

5 线性方程组的敏度分析

一个线性方程组$Ax = b$是由它的系数矩阵A和它的右端项b所确定的. 而在实际问题中, 通过观察或通过计算得到的A与b中的数据是带有误差的, 亦即A, b 受到了扰动, 通常这种扰动相对于精确数据是微小的. 那么, 自然要问: A 和b的微小扰动将对线性方程组的解有何影响? 即所谓的线性方程组的敏感性问题. 也许读者认为, 既然A, b受到的扰动是微小的, 那么对应的方程组的解x的变化也应该是微小的. 然而对于某些实际问题, 情况并非如此, 请看下例.

例3 线性方程组

$$\begin{bmatrix} 2.0002 & 1.9998 \\ 1.9998 & 2.0002 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$$

的解为$x = (1,\, 1)^{\mathrm{T}}$. 若方程组右端有扰动$\delta b = (2 \times 10^{-4}, -2 \times 10^{-4})^{\mathrm{T}}$, 则原方程组变为

$$\begin{bmatrix} 2.0002 & 1.9998 \\ 1.9998 & 2.0002 \end{bmatrix} \begin{bmatrix} \widetilde{x}_1 \\ \widetilde{x}_2 \end{bmatrix} = \begin{bmatrix} 4.0002 \\ 3.9998 \end{bmatrix},$$

其解为$\widetilde{x} = (1.5,\, 0.5)^{\mathrm{T}}$. 这样, 我们有

$$\frac{\|\widetilde{x} - x\|_\infty}{\|x\|_\infty} = \frac{1}{2}, \qquad \frac{\|\delta b\|_\infty}{\|b\|_\infty} = \frac{1}{20000},$$

即解的相对误差是右端项相对误差的10000倍.

这个例子表明, 确实有一些线性方程组其系数的微小变化会引起解的巨大变化. 下面我们就一般的非奇异线性方程组$Ax = b$来讨论其敏感性问题. 假定该方程组经微小扰动之后变为

$$(A + \delta A)(x + \delta x) = b + \delta b.$$

将$b = Ax$代入上式并整理可得

$$(A + \delta A)\delta x = \delta b - \delta A x. \tag{5.1}$$

由于A非奇异, 故在δA充分小时, $A + \delta A$仍是非奇异的. 事实上, 由推论1知, 只要$\|A^{-1}\| \, \|\delta A\| < 1$, 就有$A + \delta A$可逆, 而且

$$\|(I + A^{-1}\delta A)^{-1}\| \leqslant \frac{1}{1 - \|A^{-1}\| \, \|\delta A\|}. \tag{5.2}$$

因此, 在此条件下, 有$A + \delta A = A(I + A^{-1}\delta A)$是非奇异的, 而且由(5.1)可得

$$\begin{aligned} \delta x &= (A + \delta A)^{-1}(\delta b - \delta A x) \\ &= (I + A^{-1}\delta A)^{-1} A^{-1}(\delta b - \delta A x). \end{aligned}$$

两边取范数得

$$\begin{aligned} \|\delta x\| &\leqslant \|(I + A^{-1}\delta A)^{-1}\| \, \|A^{-1}\|(\|\delta b\| + \|\delta A\| \, \|x\|) \\ &\leqslant \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \, \|\delta A\|}\big(\|\delta b\| + \|\delta A\| \, \|x\|\big), \end{aligned}$$

上式最后一步利用了不等式(5.2). 上式两边都除以$\|x\|$ (当然, 这里假定$x \neq 0$), 并注意到$\|b\| \leqslant \|A\| \, \|x\|$, 便有

$$\frac{\|\delta x\|}{\|x\|} \leqslant \frac{\|A^{-1}\| \, \|A\|}{1 - \|A^{-1}\| \, \|\delta A\|} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right).$$

这样, 我们就证明了如下定理

定理13 设$\|\cdot\|$是$\boldsymbol{R}^{n\times n}$上的一个满足条件$\|I\| = 1$的矩阵范数, 并假定$A \in \boldsymbol{R}^{n\times n}$非奇异, $b \in \boldsymbol{R}^n$非零. 再假定$\delta A \in \boldsymbol{R}^{n\times n}$满足$\|A^{-1}\| \, \|\delta A\| < 1$. 若$x$和$x + \delta x$分别是线性方程组

$$Ax = b \quad \text{和} \quad (A + \delta A)(x + \delta x) = b + \delta b$$

的解, 则

$$\frac{\|\delta x\|}{\|x\|} \leqslant \frac{\kappa(A)}{1 - \kappa(A)\frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right), \tag{5.3}$$

其中$\kappa(A) = \|A^{-1}\| \, \|A\|$.

当$\dfrac{\|\delta A\|}{\|A\|}$较小时, 有

$$\frac{\kappa(A)}{1 - \kappa(A)\frac{\|\delta A\|}{\|A\|}} \approx \kappa(A),$$

从而, 有

$$\frac{\|\delta x\|}{\|x\|} \lessapprox \kappa(A) \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right).$$

由此可知, 线性方程组的解x的相对误差的上界是右端项b和系数矩阵A的相对误差之和乘以一个放大倍数$\kappa(A)$而得到的. 因此, 扰动对线性方程组之解的影响大小便与这个放大倍数有很大关系. 若这个放大倍数不大, 则扰动对解的影响也不会太大；而若这个放大倍数很大, 则扰动对解的影响可能就很大. 于是, 我们有如下定义.

定义5 数$\kappa(A)=\|A\|\|A^{-1}\|$称为线性方程组$Ax=b$的**条件数**.

条件数在一定程度上刻划了扰动对方程组解的影响程度. 通常, 若线性方程组的系数矩阵A的条件数$\kappa(A)$很大, 则我们就说该线性方程组的求解问题是病态, 有时亦说A是病态的；反之, 若$\kappa(A)$很小, 则我们就说该线性方程组的求解问题是良态的, 或说A是良态的.

显然, 条件数与范数有关, 当要强调使用什么样的范数时,可在条件数上加上下标, 如

$$\kappa_2(A)=\|A\|_2\,\|A^{-1}\|_2.$$

既然条件数与范数有关, 那么自然要问: 一个方程组在一种范数下是病态的, 在另一种范数下又如何呢? 事实上, 由矩阵范数的等价性容易推出, $\boldsymbol{R}^{n\times n}$上任意两个范数下的条件数$\kappa_\alpha(A)$和$\kappa_\beta(A)$都是等价的, 即存在常数$c_1$和$c_2$, 使得

$$c_1\,\kappa_\alpha(A)\leqslant\kappa_\beta(A)\leqslant c_2\,\kappa_\alpha(A).$$

例如, 有

$$\begin{aligned}
&\frac{1}{n}\,\kappa_2(A)\leqslant\kappa_1(A)\leqslant n\,\kappa_2(A),\\
&\frac{1}{n}\,\kappa_\infty(A)\leqslant\kappa_2(A)\leqslant n\,\kappa_\infty(A),\\
&\frac{1}{n^2}\,\kappa_1(A)\leqslant\kappa_\infty(A)\leqslant n^2\,\kappa_1(A).
\end{aligned}$$

这样, 一个矩阵在α范数下是病态的, 则它在β范数下也是病态的.

此外, 由不等式(5.2)容易导出如下结果:

推论2 设$\|\cdot\|$是$\boldsymbol{R}^{n\times n}$上的一个满足条件$\|I\|=1$的矩阵范数, 并假定$A\in\boldsymbol{R}^{n\times n}$是非奇异的, 而且$\delta A\in\boldsymbol{R}^{n\times n}$满足$\|A^{-1}\|\,\|\delta A\|<1$, 则$A+\delta A$也是非奇异的, 而且有

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

这表明$\kappa(A)=\|A^{-1}\|\,\|A\|$亦可作为矩阵求逆问题的条件数.

最后我们再来看一下条件数的几何意义.

定理14 设$A\in\boldsymbol{R}^{n\times n}$非奇异, 则

$$\min\left\{\frac{\|\delta A\|_2}{\|A\|_2}\ :\ A+\delta A\text{ 奇异}\right\}=\frac{1}{\|A\|_2\,\|A^{-1}\|_2}=\frac{1}{\kappa_2(A)},\tag{5.4}$$

即在谱范数下, 一个矩阵的条件数的倒数正好等于该矩阵与全体奇异矩阵所成集合的相对距离.

证明 只需证明

$$\min\left\{\|\delta A\|_2\ :\ A+\delta A\text{ 奇异}\right\}=\frac{1}{\|A^{-1}\|_2}$$

即可. 由推论2.1.1可知, 当$\|A^{-1}\|_2\,\|\delta A\|_2 < 1$时, $A+\delta A$必是非奇异的, 从而有

$$\min\left\{\|\delta A\|_2 \,:\, A+\delta A \text{ 奇异}\right\} \geqslant \frac{1}{\|A^{-1}\|_2}. \tag{5.5}$$

此外, 由于谱范数是由向量2范数诱导出的算子范数, 因而必存在$x \in \boldsymbol{R}^n$ 满足$\|x\|_2 = 1$使得$\|A^{-1}x\|_2 = \|A^{-1}\|_2$. 现令

$$y = \frac{A^{-1}x}{\|A^{-1}x\|_2}, \qquad \delta A = -\frac{xy^{\mathrm{T}}}{\|A^{-1}\|_2},$$

则有$\|y\|_2 = 1$, 而且

$$(A+\delta A)y = Ay + \delta Ay = \frac{x}{\|A^{-1}x\|_2} - \frac{x}{\|A^{-1}\|_2} = 0,$$

$$\|\delta A\|_2 = \max_{\|z\|_2=1}\left\|\frac{xy^{\mathrm{T}}}{\|A^{-1}\|_2}z\right\|_2 = \frac{\|x\|_2}{\|A^{-1}\|_2}\max_{\|z\|_2=1}|y^{\mathrm{T}}z| = \frac{1}{\|A^{-1}\|_2}.$$

这也就是说, 我们已经找到了一个$\delta A \in \boldsymbol{R}^{n\times n}$使得$A+\delta A$奇异, 而且又有$\|\delta A\|_2 = \|A^{-1}\|_2^{-1}$. 这样, 结合(5.5)便有(5.4)成立. □

定理14表明, 当$A \in \boldsymbol{R}^{n\times n}$十分病态时, 就说明$A$已与一个奇异矩阵十分靠近.

6 迭代改进

若计算解$\hat{x}$的精度太低, 可将$\hat{x}$作为初值, 应用Newton迭代法于函数$f(x) = Ax - b$上, 来改进其精度. 具体计算过程可按如下步骤进行:

(1) 计算$r = b - A\hat{x}$ (用双精度和原始矩阵A);
(2) 求解$Az = r$ (利用A的三角分解);
(3) 计算$x = \hat{x} + z$;
(4) 若$\dfrac{\|x-\hat{x}\|_\infty}{\|x\|_\infty} \leqslant \varepsilon$, 则结束; 否则, 令$\hat{x} = x$, 转步(1).

实际计算的经验表明, 当A病态的并不是十分严重时, 利用这一方法最终可使其解的计算精度达到机器精度. 可是, 当A十分病态时, 这样做对解的精度并不会有太大的改进.

7 正交变换

前两小节我们来介绍两个最基本的初等正交变换, 它们是数值线性代数中许多重要算法的基础. 第三小节介绍重要的正交变换–QR分解.

7.1 Householder变换

使用Gauss变换将一个矩阵约化为上三角形式是基于一个简单的事实：对于任一个给定的向量x, 可构造一个初等下三角阵L, 使$Lx = \alpha e_1$, 这里e_1是I的第一列, $\alpha \in \boldsymbol{R}$. 这一节我们就来讨论如何求一个初等正交矩阵, 使其具有矩阵L的功能. 这样, 对一个矩阵的上三角化任务, 便可以由一系列的初等正交变换来完成.

定义6 设$w \in \boldsymbol{R}^n$满足$\|w\|_2 = 1$, 定义$H \in \boldsymbol{R}^{n\times n}$为

$$H = I - 2ww^{\mathrm{T}}, \tag{7.1}$$

则称H为**Householder变换**.

Householder变换也叫做**初等反射矩阵**或**镜像变换**, 它是著名的数值分析专家Householder在1958年为讨论矩阵特征值问题而提出来的. 下面的定理给出了Householder变换的一些简单而又十分重要的性质.

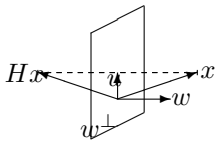
定理15 设H是由(7.1)定义的Householder变换, 那么H满足

(1) 对称性: $H^{\mathrm{T}} = H$;

(2) 正交性: $H^{\mathrm{T}}H = I$;

(3) 对合性: $H^2 = I$;

(4) 反射性: 对任意的$x \in \boldsymbol{R}^n$, 如下图所示, Hx 是x关于w的垂直超平面的镜像反射.



证明 (1) 显然. (2)和(3)可由(1)导出. 事实上, 我们有

$$\begin{aligned} H^{\mathrm{T}}H = H^2 &= (I - 2ww^{\mathrm{T}})(I - 2ww^{\mathrm{T}}) \\ &= I - 4ww^{\mathrm{T}} + 4ww^{\mathrm{T}}ww^{\mathrm{T}} = I. \end{aligned}$$

(4) 设$x \in \boldsymbol{R}^n$, 则x可表示为

$$x = u + \alpha w,$$

其中$u \in \operatorname{span}\{w\}^{\perp}$, $\alpha \in \boldsymbol{R}$. 利用$u^{\mathrm{T}}w = 0$和$w^{\mathrm{T}}w = 1$可得

$$\begin{aligned} Hx &= (I - 2ww^{\mathrm{T}})(u + \alpha w) \\ &= u + \alpha w - 2ww^{\mathrm{T}}u - 2\alpha ww^{\mathrm{T}}w \\ &= u - \alpha w, \end{aligned}$$

这就说明了Hx为x关于$\operatorname{span}\{w\}^{\perp}$的镜像反射. □

Householder变换除了具有定理15所述的良好性质外, 它的主要用途在于, 它能如Gauss变换一样, 可以通过适当选取单位向量w, 把一个给定向量的若干个指定的分量变为零.

定理16 设$0 \neq x \in \boldsymbol{R}^n$, 则可构造单位向量$w \in \boldsymbol{R}^n$, 使由(7.1) 定义的Householder变换H满足

$$Hx = \alpha e_1,$$

其中$\alpha = \pm\|x\|_2$.

证明 由于

$$Hx = (I - 2ww^{\mathrm{T}})x = x - 2(w^{\mathrm{T}}x)w,$$

故欲使$Hx = \alpha e_1$, 则w应为

$$w = \frac{x - \alpha e_1}{\|x - \alpha e_1\|_2}.$$

对$\alpha = \pm\|x\|_2$, 直接验证可知这样定义的w满足定理的要求. □

定理16 告诉我们, 对任意的$x \in \boldsymbol{R}^n$ $(x \neq 0)$ 都可构造出Householder矩阵H, 使Hx的后$n-1$个分量为零; 而且其证明亦告诉我们, 可按如下的步骤来构造确定H的单位向量w:

(1) 计算$v = x \pm \|x\|_2 e_1$;

(2) 计算$w = v/\|v\|_2$.

首先, 一个自然的问题是, 实际计算时, $\|x\|_2$前的符号如何选取最好. 为了使变换后得到的α为正数, 则应取

$$v = x - \|x\|_2 e_1.$$

但是这样选取就会出现一个问题, 如果x是一个很接近于e_1的向量, 计算

$$v_1 = x_1 - \|x\|_2$$

时, 就会出现两个相近的数相减, 而导致严重地损失有效数字, 这里v_1和x_1分别表示向量v和x的第一个分量. 不过, 幸运的是, 只要对上式做一简单的等价变形, 就可避免这一问题的出现. 事实上, 注意到

$$v_1 = x_1 - \|x\|_2 = \frac{x_1^2 - \|x\|_2^2}{x_1 + \|x\|_2} = \frac{-(x_2^2 + \cdots + x_n^2)}{x_1 + \|x\|_2},$$

只要在$x_1 > 0$时使用上面式子来计算v_1, 就会避免出现两个相近的数相减的情形.

其次, 注意到

$$H = I - 2ww^{\mathrm{T}} = I - \frac{2}{v^{\mathrm{T}}v}vv^{\mathrm{T}} = I - \beta vv^{\mathrm{T}}$$

其中$\beta = 2/v^{\mathrm{T}}v$, 我们就没有必要非求出$w$不可, 而只需求出$\beta$和$v$即可. 然而在实际计算时, 将$v$规格化为第一个分量为1的向量是方便的, 这是因为这样正好可以把v的后$n-1$个分量保存在x的后$n-1$个化为0的分量位置上, 而v的第一个分量1就无需保存了.

此外, 上溢和下溢也是计算中需要考虑的问题. 当下溢发生时, 一些计算机系统自动置其为零, 这就可能出现$v^{\mathrm{T}}v$为零的情形. 另外如果x的分量太大, 当该分量平方时, 便会出现上溢. 考虑到对任意的非零实数α有αv与v的单位化向量相同, 为了避免溢出现象的出现, 我们可用$x/\|x\|_\infty$代替x来构造v(这样做相当于在原来的v之前乘了常数$\alpha = 1/\|x\|_\infty$).

根据上面的讨论, 可得如下的基本算法.

算法8 (计算Householder变换)

function: $[v, \beta] = \mathbf{house}(x)$
$n = \mathbf{length}(x)$ (向量x的长度)
$\eta = \|x\|_\infty$ $x = x/\eta$
$\sigma = x(2:n)^{\mathrm{T}}x(2:n)$
$v(1) = 1;\ v(2:n) = x(2:n)$
if $\sigma = 0$
$\beta = 0$
else

$\alpha = \sqrt{x(1)^2 + \sigma}$
if $x(1) \leqslant 0$
$v(1) = x(1) - \alpha$
else
$v(1) = -\sigma/(x(1) + \alpha)$
end
$\beta = 2v(1)^2/(\sigma + v(1)^2)$; $v = v/v(1)$
end

利用Householder变换在一个向量中引入零元素, 并不局限于$Hx = \alpha e_1$ 的形式, 其实它可以将向量中任何若干相邻的元素化为零. 例如, 欲在$x \in \boldsymbol{R}^n$中从$k+1$至j位置引入零元素, 只需定义v为

$$v = (0, ..., 0, x_k - \alpha, x_{k+1}, ..., x_j, 0, ..., 0)$$

即可, 其中$\alpha^2 = \sum_{i=k}^{j} x_i^2$.

在应用Householder变换约化一个给定矩阵为某一需要的形式时, 其主要的工作量是计算一个Householder矩阵$H = I - \beta vv^{\mathrm{T}} \in \boldsymbol{R}^{m\times m}$ 与一个已知矩阵$A \in \boldsymbol{R}^{m\times n}$的乘积. 在实际计算时, H并不需要以显式给出, 而是根据如下的公式来计算

$$HA = (I - \beta vv^{\mathrm{T}})A = A - \beta v(A^{\mathrm{T}}v)^{\mathrm{T}} = A - vw^{\mathrm{T}},$$

其中$w = \beta A^{\mathrm{T}}v$, 即

(1) 计算$w = \beta A^{\mathrm{T}}v$;

(2) 计算$B = A - vw^{\mathrm{T}}$ (B即为所求的乘积HA).

完成这一计算任务所需的运算量为$4mn$.

7.2 Givens变换

欲把一个向量中许多分量化为零, 可以用Householder变换, 例如前面所讲到的把一个向量中若干相邻分量化为零. 如果只将其中一个分量化为零, 则应采用**Givens变换**, 它有如下形式:

$$G(i,k,\theta) = I + s(e_ie_k^{\mathrm{T}} - e_ke_i^{\mathrm{T}}) + (c-1)(e_ie_i^{\mathrm{T}} + e_ke_k^{\mathrm{T}})$$

$$= \begin{bmatrix} 1 & & & \vdots & & \vdots & & & \\ & \ddots & & \vdots & & \vdots & & & \\ \cdots & \cdots & & c & \cdots & s & \cdots & \cdots & i \\ & & & \vdots & & \vdots & & & \\ \cdots & \cdots & & -s & \cdots & c & \cdots & \cdots & k \\ & & & \vdots & & \vdots & \ddots & & \\ & & & \vdots & & \vdots & & 1 & \end{bmatrix},$$

其中$c=\cos\theta,\ s=\sin\theta$. 易证$G(i,k,\theta)$是一个正交阵.

设$x\in\boldsymbol{R}^n$. 令$y=G(i,k,\theta)x$, 则有

$$\begin{aligned} y_i &= cx_i+sx_k,\\ y_k &= -sx_i+cx_k,\\ y_j &= x_j,\quad j\neq i,k. \end{aligned}$$

因此, 若要$y_k=0$, 只要取

$$c=\frac{x_i}{\sqrt{x_i^2+x_k^2}},\qquad s=\frac{x_k}{\sqrt{x_i^2+x_k^2}}, \tag{7.2}$$

便有

$$y_i=\sqrt{x_i^2+x_k^2},\qquad y_k=0.$$

几何上来看, $G(i,k,\theta)x$是在(i,k)坐标平面内将x按顺时针方向作了θ度的旋转. 所以Givens变换亦称**平面旋转变换**.

若利用(7.2)计算c和s, 可能会发生溢出. 为避免这种情形的发生, 对给定的实数a和b, 实际上是按如下算法所述的方法计算$c=\cos(\theta)$和$s=\sin(\theta)$, 使得

$$\begin{bmatrix} c & s\\ -s & c \end{bmatrix}\begin{bmatrix} a\\ b \end{bmatrix}=\begin{bmatrix} r\\ 0 \end{bmatrix}$$

的.

算法9 (计算Givens变换)

function: $[c,\ s]=$**givens** $(a,\ b)$
  if $b=0$
    $c=1;\ \ s=0$
  else
    if $|b|>|a|$
      $\tau=a/b;\ \ s=1/\sqrt{1+\tau^2};\ \ c=s\tau$
    else
      $\tau=b/a;\ \ c=1/\sqrt{1+\tau^2};\ \ s=c\tau$
    end
  end

如果用一个Givens变换左(或右)乘一个矩阵$A\in\boldsymbol{R}^{n\times q}$, 则它只改变$A$的第$i,\ k$行(或列)的元素, 其余元素保持不变. 请读者作为练习写出其详细的算法.

7.3 正交分解

定理17 (QR分解定理) 设$A\in\boldsymbol{R}^{m\times n}\ (m\geqslant n)$, 则$A$有**QR分解**:

$$A=Q\begin{bmatrix} R\\ 0 \end{bmatrix}, \tag{7.3}$$

其中$Q \in \boldsymbol{R}^{m\times m}$是正交矩阵, $R \in \boldsymbol{R}^{n\times n}$是具有非负对角元的上三角阵；而且当$m = n$且$A$非奇异时, 上述的分解还是唯一的.

证明 先证明QR分解的存在性. 对n用数学归纳法. 当$n = 1$时, 自然成立. 现假定已经证明定理对所有的$p \times (n-1)$矩阵成立, 这里假设$p \geqslant (n-1)$. 设$A \in \boldsymbol{R}^{m\times n}$的第一列为$a_1$. 则由定理3.2.2知, 存在正交矩阵$Q_1 \in \boldsymbol{R}^{m\times m}$ 使得$Q_1^{\mathrm{T}} a_1 = \|a_1\|_2 e_1$, 于是, 有

$$Q_1^{\mathrm{T}} A = \begin{bmatrix} \|a_1\|_2 & v^{\mathrm{T}} \\ 0 & A_1 \end{bmatrix} \begin{matrix} 1 \\ m-1 \end{matrix}.$$

对$(m-1)\times(n-1)$矩阵A_1应用归纳法假定, 得

$$A_1 = Q_2 \begin{bmatrix} R_2 \\ 0 \end{bmatrix},$$

其中Q_2是$(m-1)\times(m-1)$正交矩阵, R_2 是具有非负对角元的$(n-1)\times(n-1)$上三角阵. 这样, 令

$$Q = Q_1 \begin{bmatrix} 1 & 0 \\ 0 & Q_2 \end{bmatrix}, \qquad R = \begin{bmatrix} \|a_1\|_2 & v^{\mathrm{T}} \\ 0 & R_2 \\ 0 & 0 \end{bmatrix},$$

则Q和R满足定理的要求. 于是, 由归纳法原理知存在性得证.

再证唯一性. 设$m = n$且A非奇异, 并假定$A = QR = \widetilde{Q}\widetilde{R}$, 其中$Q$, $\widetilde{Q} \in \boldsymbol{R}^{m\times m}$是正交矩阵, R, $\widetilde{R} \in \boldsymbol{R}^{n\times n}$是具有非负对角元的上三角阵. A非奇异蕴含着R, $\widetilde{R}$ 的对角元均为正数. 因此, 我们有

$$\widetilde{Q}^{\mathrm{T}} Q = \widetilde{R} R^{-1}$$

既是正交矩阵又是对角元均为正数的上三角矩阵, 这只能是单位矩阵. 从而, 必有$\widetilde{Q} = Q$, $\widetilde{R} = R$, 即分解是唯一的. □

下面我们来介绍实现QR分解最常用的Householder方法.

用Householder方法计算QR分解与不选主元的Gauss消去法很类似, 就是利用Householder变换逐步将A约化为上三角矩阵. 设$m = 7$, $n = 5$, 并假定已经计算出Householder变换H_1和H_2使得

$$H_2 H_1 A = \begin{bmatrix} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & 0 & + & \times & \times \\ 0 & 0 & + & \times & \times \\ 0 & 0 & + & \times & \times \\ 0 & 0 & + & \times & \times \\ 0 & 0 & + & \times & \times \end{bmatrix}.$$

那么我们现在的任务就是集中精力于第三列标为“+”的5个元素, 确定一个Householder变换$\widetilde{H}_3 \in$

$\boldsymbol{R}^{5\times 5}$使得

$$\widetilde{H}_3\begin{bmatrix}+\\+\\+\\+\\+\end{bmatrix}=\begin{bmatrix}\times\\0\\0\\0\\0\end{bmatrix}.$$

令$H_3=\mathrm{diag}\,(I_2,\,\widetilde{H}_3)$, 则有

$$H_3H_2H_1A=\begin{bmatrix}\times&\times&\times&\times&\times\\0&\times&\times&\times&\times\\0&0&\times&\times&\times\\0&0&0&\times&\times\\0&0&0&\times&\times\\0&0&0&\times&\times\\0&0&0&\times&\times\end{bmatrix}.$$

对于一般的矩阵$A\in\boldsymbol{R}^{m\times n}$, 假定我们已进行了$k-1$步, 得到了Householder变换$H_1,\ldots,H_{k-1}$, 使得

$$A_k=H_{k-1}\cdots H_1A=\begin{bmatrix}A_{11}^{(k)}&A_{12}^{(k)}\\0&A_{22}^{(k)}\end{bmatrix}\begin{matrix}k-1\\m-k+1\end{matrix},$$

其中$A_{11}^{(k)}$是上三角阵. 假定

$$A_{22}^{(k)}=[u_k,\ldots,u_n].$$

第k步是：先用算法3.2.1确定Householder变换

$$\widetilde{H}_k=I_{m-k+1}-\beta_kv_kv_k^{\mathrm{T}}\in\boldsymbol{R}^{(m-k+1)\times(m-k+1)},$$

使得

$$\widetilde{H}_ku_k=r_{kk}e_1,$$

其中$r_{kk}\geqslant 0$, $e_1=(1,0,...,0)^{\mathrm{T}}\in\boldsymbol{R}^{m-k+1}$; 然后, 再计算$\widetilde{H}_kA_{22}^{(k)}$. 令

$$H_k=\mathrm{diag}\,(I_{k-1},\widetilde{H}_k),$$

则

$$A_{k+1}=H_kA_k=\begin{bmatrix}A_{11}^{(k)}&A_{12}^{(k)}\\0&\widetilde{H}_kA_{22}^{(k)}\end{bmatrix}$$

$$=\begin{bmatrix}A_{11}^{(k+1)}&A_{12}^{(k+1)}\\0&A_{22}^{(k+1)}\end{bmatrix}\begin{matrix}k\\m-k\end{matrix},$$

其中$A_{11}^{(k+1)}$是上三角阵. 这样, 从$k=1$出发, 依次进行n次, 我们就可将A约化为上三角阵. 现在记

$$R = A_{11}^{(n)}, \qquad Q = H_1 \cdots H_n,$$

则

$$A = Q \begin{bmatrix} R \\ 0 \end{bmatrix}.$$

注意, 这样得到的上三角矩阵R的对角元均是非负的.

下面考虑计算A的QR分解的存储问题. 当分解完成后, 一般来说, A就不再需要, 便可用来存放Q与R. 通常并不是将Q算出, 而是只存放构成它的n个Householder矩阵H_k, 而对每个H_k, 我们只需保存v_k和β_k即可. 注意到v_k有如下形式

$$v_k = \left(1, v_{k+1}^{(k)}, \ldots, v_n^{(k)}\right)^{\mathrm{T}},$$

我们正好可以将$v_k(2:m-k+1)$存储在A的对角元以下位置上. 例如, 对于$m=4$, $n=3$的问题, 其存储方式如下:

$$A := \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ v_2^{(1)} & r_{22} & r_{23} \\ v_3^{(1)} & v_3^{(2)} & r_{33} \\ v_4^{(1)} & v_4^{(2)} & v_4^{(3)} \end{bmatrix}, \qquad d := \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}.$$

综合上面的讨论, 可得如下算法.

算法10 (计算QR分解:Householder 变换法)

for $j = 1:n$
  $[v,\, \beta] = \mathbf{house}\,(A(j:m, j))$
  $A(j:m,\, j:n) = (I_{m-j+1} - \beta v v^{\mathrm{T}}) A(j:m,\, j:n)$
  $d(j) = \beta$
  if $j < m$
    $A(j+1:m,\, j) = v(2:m-j+1)$
  end
end

容易算出, 这一算法的运算量为$2n^2(m - n/3)$.

Householder方法并不是实现QR分解的唯一方法, 例如, 我们亦可利用Givens变换或Gram–Schmidt正交化来实现. 通常用Givens变换来实现QR分解所需的运算量大约是Householder方法的二倍, 但如果A有较多的零元素, 则灵活地使用Givens变换往往会使运算量大为减少.

此外, QR分解可用来求解最小二乘问题, 而且它也是数值代数许多重要算法的基础. 例如, 著名的求解特征值问题的QR 方法就是利用这一分解而得到的; 再如, 我们亦可利用QR分解求解线性方程组(1.3), 而且对于某些病态方程组QR分解法的计算结果往往要比三角分解法好的多, 当然, 前者比后者的运算量也要大的多.