NAL TM-315

UDC 512.83:
681.3.06

航空宇宙技術研究所資料

TECHNICAL MEMORANDUM OF NATIONAL AEROSPACE LABORATORY

TM-315

1 次 方 程 式 系 の 解 法 III

——長方行列を対象とするもの——

福 田 正 大 ・ 末 松 俊 二

1976 年 10 月

航 空 宇 宙 技 術 研 究 所
NATIONAL AEROSPACE LABORATORY

目　　次

1 次 方 程 式 系 の 解 法 III

―― 長方行列を対象とするもの ――*

福 田 正 大**・ 末 松 俊 二**

概　要

本報告は，一連の一次方程式系を取り扱ったもののうち，長方行列を対象としたものである。そのうち主として最小自乗問題の解法を取り扱い，特殊な場合のものとして，実正方行列の逆行列を計算するものがある。方法としては，ハウスホルダー変換を利用したもの，消去法型の直交化法を利用したもの，及び singular value decomposition を利用したものの3つがある。

1章　緒　言

本報告は，「1次方程式系の解法　I」[1] にも述べてあるように，科学技術計算用サブルーチンライブラリーの標準化ならびに充実化という目標に沿ったものの一つである。これは，計算センターにとって各種解法の知識を集積していくことが必要と考えるからである。このような研究は地味ではあるが計算センターの質を高める上で甚だ重要なことと思われる。

今回報告するものは，長方行列 A を係数行列にもつ overdetermined system

$$Ax = b$$

を最小自乗法によって解くものを主としている。行列 A としては稠密なものを想定しており，帯状をした行列に対する特殊化は行っていない。なお，各サブルーチンの仕様で，「1次方程式系の解法　II」で採用した「1次方程式系の解法　I」からの変更をここでも採用した。それは次の二点である。

1) 1次元のワーキングエリアも引き数に加える。こうすることによって引き数の並びは若干煩雑になるが，その代わりに無駄なエリアを取らなくてすむ。

2) 整合配列の引き数を各2次元アレイごとにつけるのではなく，行列として同じ行数をもつものについては同じ整合配列引き数を使用する。

(例)　A が $m\times n$，B が $m\times r$，X が $n\times r$ という行列のとき，A と B には同じ整合配列引き数 KA が X には異なる引き数 KX が使用されている。従って，サブルーチンの中での dimension 宣言は

DIMENSION　$A(KA,N)$, $B(KA,R)$,
$X(KX,R)$

となっている。

「1次方程式系の解法　I，II，III」全体のサブルーチンの用いられ方については，本報告末尾の綴じ込みに図版Aとしてフローチャートを載せてある。

2章　数値テスト

ここでいう最小自乗問題とは，m 行 n 列 $(m>n)$ の行列 A と右辺のベクトル b に対して，

$$\|b-Ax\| = \min_{\xi}\|b-A\xi\|$$

となる x を求めることである。複数個のベクトル b_i $i=1, r$ について解くときには，右辺はそれらのベクトルより作られる $m\times r$ 行列 B として扱う。この解を求めるルーチンのテストとして次の4つのものを使用した。なお，r は右辺にあるベクトルの個数である。又，各長方行列の singular value の最大値 σ_1 と最小値 σ_n の値を記しておく。ここで，singular value というのは A^TA の固有値の正の平方根のところである。正値対称行列 A^TA のスペクトルノルムに関する条件数は $(\sigma_1/\sigma_n)^2$ になる。

(1)　$m=6$　$n=5$　$r=2$

$$A = \begin{pmatrix} 36 & -630 & 3360 & -7560 & 7560 \\ -630 & 14700 & -88200 & 211680 & -220500 \\ 3360 & -88200 & 564480 & -1411200 & 1512000 \\ -7560 & 211680 & -1411200 & 3628800 & -3969000 \\ 7560 & -220500 & 1512000 & -3969000 & 4410000 \\ -2772 & 83160 & -582120 & 1552320 & -1746360 \end{pmatrix}$$

$$\begin{pmatrix} b_1 & b_2 \\ 463 & -4157 \\ -13860 & -17820 \end{pmatrix}$$

* 昭和51年8月11日　受付
** 計算センター

$$\begin{pmatrix} 97020 & 93555 \\ -258720 & -261800 \\ 291060 & 288288 \\ -116424 & -118944 \end{pmatrix}$$

$\sigma_1=8.888\times10^6$　$\sigma_5=1.892$　$\sigma_1/\sigma_5=4.699\times10^6$

この行列Aは，6次ヒルベルト行列の逆行列の最初の5列より成っているものである。右辺のベクトルb_1は，$\|b_1-Ax\|=0$となるものであり，b_2はAの各列に直交するベクトルをb_1に加えたものである。従って最小自乗問題の解はどちらも同じである。

(2)　$m=6$　$n=5$　$r=3$

$$A=\begin{pmatrix} -74 & 80 & 18 & -11 & -4 \\ 14 & -69 & 21 & 28 & 0 \\ 66 & -72 & -5 & 7 & 1 \\ -12 & 66 & -30 & -23 & 3 \\ 3 & 8 & -7 & -4 & 1 \\ 4 & -12 & 4 & 4 & 0 \end{pmatrix}\quad \begin{matrix} b_1 \\ \\ \\ \\ \\ \\ \end{matrix}\quad \begin{matrix} b_2 \\ \\ \\ \\ \\ \\ \end{matrix}\quad \begin{matrix} b_3 \\ \\ \\ \\ \\ \\ \end{matrix}$$

$$\begin{pmatrix} 51 & -56 & -5 \\ -61 & 52 & -9 \\ -56 & 764 & 708 \\ 69 & 4096 & 4165 \\ 10 & -13276 & -13266 \\ -12 & 8421 & 8409 \end{pmatrix}$$

$\sigma_1=173.5$　$\sigma_5=0.1599$　$\sigma_1/\sigma_5=1.085\times10^3$

b_1は$\|b_1-Ax\|=0$となるものであり，b_2はAの各列に直交しており（$A^Tb_2=0$），b_3はb_1とb_2を加えたものである。従って，b_1とb_3に対する最小自乗問題の解は同じであり，b_2に対するそれは0ベクトルである。

(3)　$m=8$　$n=5$　$r=3$

$$A=\begin{pmatrix} 22 & 10 & 2 & 3 & 7 \\ 14 & 7 & 10 & 0 & 8 \\ -1 & 13 & -1 & -11 & 3 \\ -3 & -2 & 13 & -2 & 4 \\ 9 & 8 & 1 & -2 & 4 \\ 9 & 1 & -7 & 5 & -1 \\ 2 & -6 & 6 & 5 & 1 \\ 4 & 5 & 0 & -2 & 2 \end{pmatrix}\quad \begin{matrix} b_1 \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix}\quad \begin{matrix} b_2 \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix}\quad \begin{matrix} b_3 \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix}$$

$$\begin{pmatrix} -1 & 1 & 0 \\ 2 & -1 & 1 \\ 1 & 10 & 11 \\ 4 & 0 & 4 \\ 0 & -6 & -6 \\ -3 & 6 & 3 \\ 1 & 11 & 12 \\ 0 & -5 & -5 \end{pmatrix}$$

$\sigma_1=35.33$　　$\sigma_5=0.0$

この行列Aのランクは3なので，通常の最小自乗問題としては解が一意には定まらない。$\|x\|$を最小にするという条件を付ければ一意に定まる。b_1，b_2，b_3の構成は(2)と同じである。

(4)　$m=7$　$n=5$　$r=3$

$$A=\begin{pmatrix} 2 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 & 2 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 2 & 4 & 1 \end{pmatrix}\quad \begin{matrix} b_1 \\ \\ \\ \\ \\ \\ \\ \end{matrix}\quad \begin{matrix} b_2 \\ \\ \\ \\ \\ \\ \\ \end{matrix}\quad \begin{matrix} b_3 \\ \\ \\ \\ \\ \\ \\ \end{matrix}$$

$$\begin{pmatrix} 0 & 7 & 21 \\ 47 & 40 & 32 \\ 22 & 22 & 25 \\ 69 & 55 & 36 \\ -4 & 3 & 17 \\ 15 & 8 & -21 \\ 8 & 15 & 26 \end{pmatrix}$$

$\sigma_1=8.242$　　$\sigma_5=1.0$　　$\sigma_1/\sigma_5=8.242$

この最小自乗問題の解は3つとも同じである。

以上4つのテストをそれぞれテスト1，テスト2，テスト3，テスト4と呼ぶことにする。これら4つの問題の正確な解と，残差の自乗$\|b-Ax\|^2$は次の通りである。

(1)　$x_1=x_2$；(1, 1/2, 1/3, 1/4, 1/5)
$\|b_2-Ax_2\|^2=7.255\times10^7$

(2)　$x_1=x_3$；(1, 2, -1, 3, -4)
$\|b_2-Ax_2\|^2=\|b_3-Ax_3\|^2=2.645\times10^8$

(3)　$x_1=x_3$；(-1/12, 0, 1/4, -1/12, 1/12)
$\|b_2-Ax_2\|^2=\|b_3-Ax_3\|^2=320$

(4)　$x_1=x_2=x_3$；(5, 4, 3, 2, 1)
$\|b_1-Ax_1\|^2=4880$　$\|b_2-Ax_2\|^2=2577$
$\|b_3-Ax_3\|^2=1913$

結果の整理には，計算機より出力された値$\bar{x}$をそのまま記し，さらに正確な解をxとして

$$y=x-\bar{x}$$
$$z_i=(x_i-\bar{x}_i)/x_i$$

を計算して，$a_m=\max|y_i|$，$a_f=\|y\|$，$s_m=\max|z_i|$，$s_f=\|z\|$の値を求めた。さらに，

$$T_S=26\ \log 2-\log\|b-Ax\|$$
$$T_D=61\ \log 2-\log\|b-Ax\|$$

の2つの値も併せ記してある。なお，z_i（相対誤差）の計算で，$x_i=0$なるときには，分子の値をz_iとして採用した。

なおテスト計算はすべて，FACOM 230-75を使用して，FORTRAN-H, V02, L05, OPT 2のもとに行った。

3章　ハウスホルダー法による最小自乗問題の解法

3－1　方法の基礎

行列Aはm行n列（$m\geqq n$）で，そのランクはnであるとする。ベクトルbが与えられたとき，$b-Ax$のユークリッドノルム$\|b-Ax\|=\left\{\sum_{i=1}^{m}\left(b_i-\sum_{j=1}^{n}a_{ij}x_j\right)^2\right\}^{1/2}$を最小にするベクトル$x$を求める問題を考える。

ユークリッドノルムは直交変換のもとで不変であるから，

$$\|b-Ax\|=\|c-QAx\|\qquad c=Qb,\ Q^TQ=I$$

となる。そこでQを

$$QA=R=\begin{pmatrix}\tilde{R}\\0\end{pmatrix}\begin{matrix}n\times n\\(m-n)\times n\end{matrix}\qquad(3.1)$$

$\tilde{R}$：上3角行列

となるように選ぶ。そうすると，$\tilde{c}$をcの始めのnコの要素から成るベクトルとして，求める解xは

$$x=\tilde{R}^{-1}\tilde{c}$$

で与えられる。

3－2　計算方法とサブルーチンの説明

(3.1)のような分解を行うための効果的な方法の1つとして，ハウスホルダー変換がある。ハウスホルダー変換は基本対称直交変換とも呼ばれ，それは長さ1のベクトルuを用いて，

$$P=I-2uu^T$$

と表わされる。容易にわかるように，Pは対称な直交行列である。今，与えられたベクトルxとyの長さが等しければ($x^Tx=y^Ty$)，xをyに変換するような行列Pが常に存在する。そのためにはuとして，

$$u=(y-x)/\{(y-x)^T(y-x)\}^{1/2}$$

をとればよい。$x=y$のときには$u=0$とする。またuの第i要素が0ならば，Pvの第i要素は変化しない。即ち

$$u_i=0 \quad \text{ならば} \quad (Pv)_i=v_i$$

となる。この変換は数値計算の手法上甚だ応用の広いものである。

ここでは次のようにしてこの変換を利用する。$A^{(1)}=A$から始めて，

$$P^{(k)}=I-\beta_k u^{(k)}u^{(k)T}$$

という対称直交行列$P^{(k)}$を用いて，$A^{(2)}$，$A^{(3)}$，$A^{(n+1)}$を順次

$$A^{(k+1)}=P^{(k)}A^{(k)} \qquad k=1,2,\cdots,n$$

に従って計算していく。ここで$P^{(k)}$の要素は

$$a_{ik}^{(k+1)}=0 \qquad i=k+1,\cdots,n$$

となるように定める。そのためには$u^{(k)}$を次のようにして求めればよい。

$$\sigma_k=\Big(\sum_{i=k}^{m}(a_{ik}^{(k)})^2\Big)^{1/2}$$

$$\beta_k=(\sigma_k(\sigma_k+|a_{kk}^{(k)}|))^{-1}$$

$$u_i^{(k)}=0 \qquad i<k$$

$$u_k^{(k)}=\mathrm{sgn}(a_{kk}^{(k)})(\sigma_k+|a_{kk}^{(k)}|)$$

$$u_i^{(k)}=a_{ik}^{(k)} \qquad i>k$$

$A^{(k+1)}=P^{(k)}A^{(k)}$を計算するには，$P^{(k)}$を求めてから計算するということはせずに，

$$y_k^T=\beta_k u^{(k)T}A^{(k)}$$

$$A^{(k+1)}=A^{(k)}-u^{(k)}y_k^T$$

という順に計算する。y_kと$A^{(k+1)}$の計算には$u^{(k)}$の始めの$k-1$個の要素が0であるという事実を十分に利用する。

さらに，できるだけの精度を確保するために，次のようなピボッティングを行う。すなわち，第kステップで$|a_{k,k}^{(k+1)}|$が最大になるような$A^{(k)}$の列を選んで変換を行う。そのためには，

$$s_j^{(k)}=\sum_{i=k}^{n}(a_{ij}^{(k)})^2 \qquad j=k,\cdots,n$$

としたとき，$|a_{k,k}^{(k+1)}|=\sigma_k$だから，$s_j^{(k)}$が最大になる列を選べばよい。$s_j^{(k)}$の計算には，列の長さが直交変換のもとで不変であるという性質を用いて，$A^{(k)}$が計算された後に

$$s_j^{(k)}=s_j^{(k-1)}-(a_{k-1,j}^{(k)})^2$$

に従って求める。サブルーチンDECOMSはここまでのことをする。

サブルーチンSOLVESは，$c=Qb$の計算と$x=\widetilde{R}^{-1}\widetilde{c}$の計算を行う。

サブルーチンLSQHDSではイタレイションによって解の精度を上げる試みをしている。今xを初期解，修正解を$x'=x+e$とし，剰余rを

$$r=b-Ax$$

とすれば，

$$\|b-Ax'\|=\|r-Ae\|$$

となる。即ち，修正ベクトルeそれ自身も又同じ最小自乗問題の解となっている。Aを一度分解し，変換を保存しておけば，rを計算してeを求めることは容易である。なお，rの計算は倍精度で行う。イタレイションは収束するまで続けるが，収束解の全部の桁が正しいという保証はない。(3-5「数値テスト及び結果」を参照)なお，このルーチンでは右辺のベクトルが複数個の場合にも使用できるようになっているので，右辺はベクトルではなく行列である。

3－3　用　途

長方行列Aを取り扱っているとしばしば$(A^TA)^{-1}$や$\det(A^TA)$の計算が必要になる。このときAを(3.1)のように分解すれば，$\widetilde{R}$が上3角行列だからその逆行列は容易に求まり，前者は

$$(A^TA)=\widetilde{R}^T\widetilde{R} \qquad (\widetilde{R}^T\widetilde{R})^{-1}=\widetilde{R}^{-1}(\widetilde{R}^{-1})^T$$

から計算できる。又後者は

$$\det(A^TA)=\det(\widetilde{R})^2=r_{11}^2r_{22}^2\cdots r_{nn}^2$$

より求まる。3角行列$\widetilde{R}$の対角要素r_{ii}の値は，DECOMS或はLSQHDSのアウトプットDにストアされている。

Aが正方行列である場合には，最小自乗解の特殊な場合として，一次方程式系の解が得られる。この場合さらに右辺が単位行列であれば逆行列が求まることになる。

3－4　サブルーチンの使用方法

(1) DECOMS, DECOMD

このサブルーチンは，長方行列Aを$A=Q^TR$と分解する。変換に関する情報はAの全下3角部分に，上3角行列Rの非対角要素はAの真上3角部分にそれぞれストアされる。従って元のAの値は失われる。

呼び出し形式

DECOMS ($M, N, A, KA, D, IP, W1, W2, ILL$)

インプット

M，N：整数型変数名又は整定数

Mは方程式の数を，Nは未知数の数を表わす。

$M \geqslant N \geqslant 2$

A：大きさKAの2次元アレイ。

$M \times N$の長方行列Aの要素をストアする。

KA：整数型変数名又は整定数

アレイAの大きさを指定，$KA \geqslant M$

アウトプット

A：3角行列$\tilde{R}$の非対角要素が真上3角部分に，ハウスホルダー変換で使用するベクトルuの非零要素が全下3角部分にそれぞれストアされている。図3－1を参照。

D：大きさNの1次元アレイ

3角行列$\tilde{R}$の対角要素がストアされている。

IP：大きさNの整数型1次元アレイ

ピボットの値がストアされている。

$W1, W2$：大きさNの1次元アレイ

ワーキングエリアとして使用。

$W2$には$s_j^{(k)}$の値がストアされている。

ILL：整数型変数名

$=0$　正常に終了したとき

$=30000$　$M<N$又は$KA<M$又は$N<2$のとき

$=K$，$1 \leqslant K \leqslant N$　　$RANK(A)=K-1$

エラー処理

$N<2$又は$M<N$又は$KA<M$のときにはILL値をセットし，さらに改頁してその旨の印字をする。この場合それ以上の計算はせずにそのままリターンする。

K番目の変換を行なうとき，対象となる列の自乗和が正確に0となれば，$RANK(A)=K-1$と判定してILL値をセットする。さらに改頁して，“SUB DECOMS IS FAIL”と印字する。

詳細説明

インプットAとアウトプットA，Dは次のようになっている。

列に関するピボッティングを行なっているので，インプットとアウトプットとでは列の番号が違っている。その対応の仕方はアウトプットIPによってわかる。即ち，アウトプットの第i列はインプットの第$IP(i)$列に対応している。

DECOMDを用いるときには，A, D, $W1$, $W2$ を倍精度実数型として指定する。

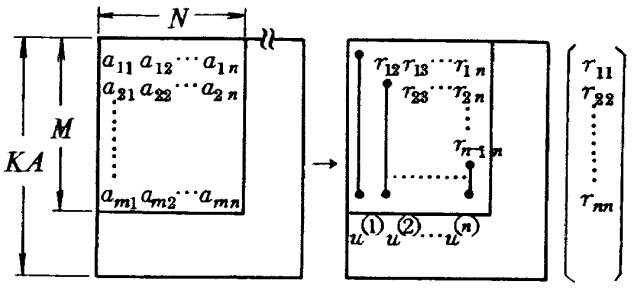


図3－1

(2) SOLVES, SOLVED

このサブルーチンは，与えられたベクトルbに対して，$c=Qb$の計算と$x=\tilde{R}^{-1}\tilde{c}$の計算を行なう。$b$の値がストアされていたエリアに新しく$c$をストアするので，元の$b$の値は失われる。本サブルーチンを使用する前に，DECOMSを用いて係数行列Aを分解しておく。

呼び出し形式

SOLVES ($M, N, A, KA, D, IP, B, X, W1$)

インプット

M，N：整数型変数名又は整定数

Mは方程式の数，Nは未知数の数を表わす。

$M \geqslant N \geqslant 2$

A：大きさKAの2次元アレイ

DECOMSからのアウトプットA

KA：整数型変数名又は整定数

アレイAの大きさを指定，$KA \geqslant M$

D：大きさNの1次元アレイ

DECOMSからのアウトプットD

IP：大きさNの整数型1次元アレイ

DECOMSからのアウトプットIP

B：大きさMの1次元アレイ

方程式の右辺の要素をストアする。

アウトプット

B：$c=Qb$の値がストアされている。

X：大きさNの1次元アレイ

最小自乗解が元の未知数の順にストアされている。

$W1$：大きさNの1次元アレイ

ワーキングエリアとして使用

エラー処理

$N<2$又は$M<N$又は$KA<M$であればなにもせずに

そのままリターンする。ILL値のセットやエラーメッセージの印字等も行なわない。

詳細説明

A, D, IP の値は変化しないが，Bの値は変化する。XとW1の実引数を同じものにしてはならない。BとX或はBとW1の実引数は同じものであってもよい。その場合前者ではB(X)に解がストアされ，後者ではBはワーキングエリアとして使用される。

SOLVEDを用いるときにはA, D, B, X, W1を倍精度実数型として指定する。

③ LSQHDS, LSQHDD

このサブルーチンは，長方行列Aを $A=Q^TR$ と分解し（DECOMSを使用），後述するイタレイションによって最小自乗解を得る（SOLVESを使用）。右辺が複数個あってもよい。逐次式からもわかるように，元のAとBの値は保存される。

呼び出し形式

LSQHDS (M, N, NR, A, KA, B, X, KX, QR, D, IP, W1, W2, W3, W4, ILL)

インプット

M, N, NR：整数型変数名又は整定数
Mは方程式の数，Nは未知数の数を表わす。
NRは方程式の右辺の列の数
$M\geqq N\geqq 2$, $NR\geqq 1$

A：大きさKAの2次元アレイ
$M\times N$の長方行列Aの要素をストアする。

KA：整数型変数名又は整定数
アレイA, B, QRの大きさを指定，$KA\geqq M$

B：大きさKAの2次元アレイ
方程式の右辺である$M\times NR$行列Bの要素をストアする。

KX：整数型変数名又は整定数
アレイXの大きさを指定，$KX\geqq N$

アウトプット

X：大きさKXの2次元アレイ
最小自乗問題の解が元の未知数の順にストアされている。

QR：大きさKAの2次元アレイ
3角行列$\tilde{R}$の非対角要素が真上3角部分に，ハウスホルダー変換で使用するベクトルuの非零要素が全下3角部分に，それぞれストアされている。DECOMSのアウトプットAを参照。

D：大きさNの1次元アレイ
3角行列$\tilde{R}$の対角成分がストアされている。

IP：大きさNの整数型1次元アレイ
ピボットの値がストアされている。

W1, W2, W3：大きさNの1次元アレイ
ワーキングエリアとして使用

W4：大きさMの1次元アレイ
ワーキングエリアとして使用

ILL：整数型変数名
$=0$　正常に終了したとき
$=30000$　$M<N$又は$KA<M$又は$KX<N$又は$N<2$のとき
$=K$　$1\leqq K\leqq N$　RANK(A)$=K-1$
$=-L$　$1\leqq L\leqq NR$　右辺のL番目のベクトルに対してイタレイションが収束しなかった。

エラー処理

$N<2$又は$M<N$又は$KA<M$又は$KX<N$のときには，ILL値をセットしさらに改頁してその旨の印字をする。この場合それ以上の計算はせずにそのままリターンする。

DECOMSをコールしてRANK(A)$<N$と判定されたときには，ILL値をセットしてリターンする。このとき，改頁して“SUB DECOMS IS FAIL”と印字される。

右辺が複数個あるとき（$NR\geqq 2$），各ベクトルごとにイタレイションを行なうので，L番目のベクトルに対して収束しなければILL値をセットする。さらに改頁して，“SUB LSQHDS IS FAIL”と印字し，リターンする。

収束判定

イタレイションのスキームは

$$x^{(0)}=0$$
$$r^{(k)}=b-Ax^{(k)},\quad e^{(k)}:\ \min\|r^{(k)}-Ae^{(k)}\|$$
$$x^{(k+1)}=x^{(k)}+e^{(k)}\qquad k=0,\ 1,\ \cdots\cdots$$

である。収束の判定は

$$\|e^{(1)}\|/\|x^{(1)}\|>1/4$$

の場合には，イタレイションループには入らずにエラーとして終了し，

$$\|e^{(k+1)}\|/\|e^{(k)}\|>1/4 \quad \text{又は}$$
$$\|e^{(k)}\|/\|x^{(1)}\|<\varepsilon \quad \varepsilon=2^{1-t}$$

t：仮数部桁数

となったときに収束したとして正常に終了する。

詳細説明

Bは$M\times NR$行列，Xは$N\times NR$行列なのでNRが1であっても，それぞれ2次元アレイとして指定する。

ILLの値が負のときには，$|ILL|-1$までの右辺のベクトルに対してはイタレイションが収束しているので，その収束解がXにストアされている。

LSQHDDを用いるときには，A, B, X, QR, D, W1, W2, W3, W4を倍精度実数型として指定する。

3－5 数値テスト及び結果

テストは，LSQHDS と LSQHDD を用いて，2章の(1)～(4)の最小自乗問題について行った。結果は表1－1～表1－3である。

表1－1

単精度 テスト1

	x_{1S}	x_{2S}
	1.0000000	5.3354787
	0.50000000	1.9476525
	0.33333334	0.95377918
	0.25000000	0.52142189
	0.20000000	0.29650499
a_m	0.0	4.335
a_f	0.0	4.622
s_m	0.0	4.335
s_f	0.0	5.662
r	6.199×10^{-7}	7.255×10^{7}
T_S		3.896
$-\log s_m$		-0.6370
$-\log s_f$		-0.7529

表1－1

倍精度

	x_{1D}	x_{2D}
	1.00000000000000000	0.99999999945136063
	0.50000000000000000	0.49999999982012801
	0.33333333333333333	0.33333333325702088
	0.25000000000000000	0.24999999996683421
	0.20000000000000000	0.19999999998826207
a_m	0.0	5.486×10^{-11}
a_f	0.0	5.835×10^{-11}
s_m	0.0	5.486×10^{-11}
s_f	0.0	7.098×10^{-11}
r	1.352×10^{-26}	7.255×10^{7}
T_D		14.43
$-\log s_m$		10.26
$-\log s_f$		10.15

表1－2

単精度 テスト2

	x_{1S}	x_{2S}	x_{3S}
	1.0000000	-8	0.99921529
	2.0000000	-9	1.9991392
	-1.0000000	2×10^{-4}	-0.99987446
	3.0000000	-20	2.9981695
	-4.0000000	34	-3.9970396
a_m	0.0	3.474×10^{-3}	2.960×10^{-3}
a_f	0.0	4.196×10^{-3}	3.673×10^{-3}
s_m	0.0		7.847×10^{-4}
s_f	0.0		1.318×10^{-3}
r	0.0	2.645×10^{-8}	2.645×10^{8}
T_S			3.616
$-\log s_m$			3.105
$-\log s_f$			2.880

表1－2

倍精度

	x_{1D}	x_{2D}	x_{3D}
	1.00000000000000000	-5	0.99999999999954085
	2.00000000000000000	-5	1.99999999999994916
	-1.00000000000000000	1×10^{-14}	-0.99999999999989330
	3.00000000000000000	-12	2.99999999999988919
	-4.00000000000000000	20	-3.99999999999981238
a_m	0.0	1.959×10^{-13}	1.876×10^{-13}
a_f	0.0	2.401×10^{-13}	2.287×10^{-13}
s_m	0.0		4.691×10^{-14}
s_f	0.0		8.020×10^{-14}
r	0.0	2.645×10^{8}	2.645×10^{8}
T_D			14.15
$-\log s_m$			13.33
$-\log s_f$			13.10

表1－3　　テスト4

単精度

	x_{1S}	x_{2S}	x_{3S}
	4.9999999	4.9999995	5.0000000
	4.0000019	4.0000004	4.0000002
	2.9999989	2.9999992	2.9999999
	1.9999991	2.0000001	1.9999999
	1.0000001	1.0000007	0.99999999
a_m	1.907×10^{-6}	7.749×10^{-7}	2.384×10^{-7}
a_f	2.361×10^{-6}	1.200×10^{-6}	2.815×10^{-7}
s_m	4.768×10^{-7}	6.855×10^{-7}	5.961×10^{-8}
s_f	7.515×10^{-7}	7.465×10^{-7}	8.574×10^{-8}
r	4880	2577	1913
T_S	5.983	6.121	6.186
$-\log s_m$	6.322	6.164	7.225
$-\log s_f$	6.124	6.127	7.067

表を見てわかるように，どの場合にも残差はほぼ正確な解のそれと一致している。又，テスト3を除いていつも"収束解"が得られてはいるが，計算機精度内の全桁があっているとは限らない。（3－2「計算方法とサブルーチンの説明」を参照） どの程度の有効数字が失われるのか量的関係は定かでないが，残差の大きさが影響するようである。テスト1，2の残差が0になるベクトルでは，全桁が正しい。又，テスト2，4では残差のオーダーほどの桁が失われている。（T_S，T_Dと$-\log s_m$を比べる） テスト1とテスト2では残差のオーダーはほぼ等しいが，失われる有効桁数はテスト1の方が多い。これは，最大singular valueと最小singular valueの比の違いによるのかもしれない。（この比の値は，前者で10^6のオーダー，後者では10^3のオーダーである）

テスト3については，rank$(A)<n$のため，本来ならこのルーチンでは計算不可能である。但し，DECOMSでのrank$(A)<n$の判定が「＝0」で行われているため，そこではrank$(A)<n$とは判定されずに通過してきた。そのために，LSQHDSで通常どおり"解"が計算され，$ILL=-2$がセットされて戻ってきた。これは，右辺の第一ベクトルに対してはイタレイションが"収束"したことを意味している。しかし，結果の値は有効数字を1桁ももっていない。ちなみに，このときの3角行列$\tilde{R}$の対角要素の値を記しておくと，（上が単精度，下が倍精度）

$(-29.53, -18.64, -16.17, 1.347\times10^{-7}, -1.242\times10^{-7})$

$(-29.53, -18.64, -16.17, 7.361\times10^{-18}, 1.387\times10^{-18})$

である。これを見ると，Aのランクは3であると判断してさしつかえない。

4章　直交化法による最小自乗問題の解法

4－1　方法の基礎

行列Aはm行n列（$m\geqq n$）であるとする。このときAの各列が1次独立であるとき（即ちAのランクはn）に限り，任意のベクトルbに対して$b-Ax$のユークリッドノルム$\|b-Ax\|$を最小にするようなベクトルxが一意に存在する。このとき次の3つの事柄が成り立っている。

(i) 残差ベクトル$b-Ax$はAの各列に直交している。即ち

$$A^T(b-Ax)=0$$

(ii) $A^\dagger=(A^TA)^{-1}A^T$をAの擬逆行列とするとxは

$$x=A^\dagger b$$

で与えられる。ここで，Aの擬逆行列とは

$$AXA=A,\quad XAX=X$$
$$(AX)^T=AX,\quad (XA)^T=XA$$

の4つの条件を満足するn行m列の行列Xのことである。

(iii) $b-Ax$は，Aを消去するようなb上への適当な射影によって得られる。即ち実エルミート行列Hで

$$H^2=H,\quad HA=0$$

となる或るHによって

$$b-Ax=Hb$$

と表わされる。このようなHとしては

$$H=I-A(A^TA)^{-1}A^T$$

をとればよい。

普通の計算方法は，(i)から直接に得られる正規方程式

$$A^TAx=A^Tb \qquad (4.1)$$

を解く。或は上式の代わりに，正則な $n\times n$ 行列 S を用いて，方程式

$$(AS)^TAx=(AS)^Tb$$

を解いてもよい。ここで S を適当に選ぶと，$AS=U$ は A の各列の正規直交基底となるようにできる。その結果得られる方程式

$$U^TAx=U^Tb$$

の係数行列 U^TA の条件数は(4.1)の係数行列 A^TA のそれよりも，一般に良い。

4－2 計算方法とサブルーチンの説明

ここで用いているアルゴリズムは行列 A を，0ではない n 個の互いに直交する列より成る行列 U と n 次の正則な上3角行列 R と $D=(U^TU)^{-1}$ とによって，$A=UDR$ と分解することに基づいている。

$$U^T(Ax-b)=0$$

という条件から

$$Rx=U^Tb$$

という系が得られる。これは正則な上3角行列を係数としているので逆進過程によって容易に解ける。ここで $R=U^TA$ と右辺の U^Tb は A と b に同じ変換 U^T を施すことによって得られる。

$Rx=U^Tb$ へ変形する計算過程は n 段階で終了する。第 i 段階では，中間行列 $A^{(i-1)}$ から新しい中間行列 $A^{(i)}$ が，第 i 列より後の各列を第 i 列に直交させることによって得られる。これは，本質的にはシュミットの直交化法である。ただ後者では，第 i 列のベクトルを第 $i-1$ 列までの全ベクトルに直交させていく過程をとっている。

$A^{(0)}=A$ から始めて，$A^{(i-1)}$ から $A^{(i)}$ を導くには，$A^{(i)}$ の第 k 列を $a_k^{(i)}$ で表わすことにして，

$$a_k^{(i)}=a_k^{(i-1)}-a_i^{(i-1)}(a_i^{(i-1)})^Ta_k^{(i-1)}/(a_i^{(i-1)})^Ta_i^{(i-1)} \qquad i<k$$

に従って各列を計算する。上式からわかるように，計算としては本質的にはガウス消去法型である。こうして n 段階後に得られる行列 $A^{(n)}$ は，直交基底の行列 U となっており，対角成分が1の上3角行列 S によって $U=AS$ と表わされる。又 U^TA と U^Tb の計算も同時に進めることができ，消去過程が終わると基底行列 U，正則な上3角行列 $R=U^TA$，系の右辺 U^Tb がすべて求まっている。

なお正規方程式 $A^TAx=A^Tb$ をガウス消去法で解いた場合にも同じ系に導かれる。というのも

$$A^TA=R^TDR$$

となり，R^TD は対角成分が1の下3角行列である。又，U を正規化しておけば，R は A^TA をコレスキー分解して得られる上3角行列に一致する。

最終的に解を得るにあたり，次のようなイタレイションによって解の精度を上げる試みをしている。初期近似解を x_0 として順次 x_n を

$$Re_n=U^T(b-Ax_n)$$

$$x_{n+1}=x_n+e_n \qquad n=0,1,\cdots\cdots$$

に従って計算する。このとき $b-Ax_n$ の計算は倍精度演算によって行なう。このために行列 A，U の要素を保存しておく必要がある。イタレイションは収束するまで続けるが，収束解の全部の桁が正しいという保証はない。（4－5「数値テスト及び結果」を参照）

ここには4種のサブルーチンがあって，そのうちの2つは最小自乗解を求めるもの，残りの2つは正方行列の逆行列を求めるものである。前者は，右辺が単一ベクトルの場合（LSQE1S）と複数個ある場合（LSQE2S）とからなっており，それぞれ上記のイタレイションを行っている。後者は，イタレイションを行なうもの（OTIVIS）と行なわないもの（OTIVNS）とからなっている。

4－3 用　途

4つのサブルーチン各々について自明である。なお $A^TA=R^TDR$ で R^TD は対角成分が1の3角行列だから，

$$\det(A^TA)=\det(R)$$

となる。上3角行列 R の要素は，アウトプットの1次元アレイ R にストアされているので，次のようにすれば $\det(R)$ の計算ができる。

```
K=N*(N+1)/2
DET=R(K)
DO 10 I=2,N
K=K-I
DET=DET*R(K)
10 CONTINUE
```

又，$(A^TA)^{-1}=R^{-1}D^{-1}(R^{-1})^T$ から $(A^TA)^{-1}$ の計算も可能である。

4－4　サブルーチンの使用方法

(1) LSQE1S, LSQE1D

このサブルーチンは，右辺が1つしかない場合，即ち$Ax=b$という形（b, xはベクトル）の最小自乗問題を解くものである。イタレイションによって解の精度を上げる試みをしているので，インプットであるAとbの値は保存される。

呼び出し形式

LSQE1S（M, N, A, KA, B, X, U, R, $W1$, $W2$, ILL）

インプット

M, N：整数型変数名又は整定数

Mは方程式の数を，Nは未知数の数を表わす

$2 \leqslant N \leqslant M$

A：大きさKAの2次元アレイ

$M \times N$の長方行列Aの要素をストアする。

KA：整数型変数名又は整定数

アレイAとUの大きさを指定，$KA \geqslant M$

B：大きさMの1次元アレイ

方程式の右辺であるベクトルbの要素をストアする。

アウトプット

X：大きさNの1次元アレイ

最小自乗問題の解がストアされている。

U：大きさKAの2次元アレイ

$M \times N$行列Uの要素がストアされている。

R：大きさ$N(N+1)/2$の1次元アレイ

3角行列Rの要素がストアされている。

図4－1を参照。

$W1$：大きさMの1次元アレイ

ワーキングエリアとして使用。

$W2$：大きさNの1次元アレイ

ワーキングエリアとして使用。

ILL：整数型変数名

＝0　正常に終了したとき

＝30000　$M<N$又は$KA<M$又は$N<2$のとき

＝1　イタレイションが収束しなかったとき

エラー処理

$N<2$又は$M<N$又は$KA<M$のときにはILL値をセットし，さらに改頁してその旨の印字をする。この場合それ以上の計算はせずにそのままリターンする。

イタレイションが収束しなかったときには，ILL値をセットし，さらに改頁して"SUB LSQE1S IS FAIL"と印字する。このときXには最終近似解がストアされている。

収束判定

イタレイションは，近似解をx_n修正ベクトルをe_nとして，

$$R x_0 = U^T b \quad , \quad x_{n+1} = x_n + e_n$$
$$R e_n = U^T (b - A x_n)$$

に従って行う。収束の判定は

$$\| e_n \| > \| x_n \| / 2$$

となったとき，収束しないとして$ILL=1$をセットして終了する。

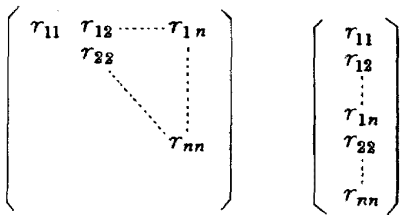
$$\| e_n \| / \| x_n \| < \varepsilon \text{ 又は } \| e_n \| > \| e_{n-1} \| / 10$$
$$\varepsilon = 2^{1-t}$$

t：仮数部桁数

となったときに収束したとして正常に終了する。

詳細説明

3角行列Rの要素は下図のようにストアされている。



行列 R　　アレイ R

図4－1

LSQE1Dを用いるときにはA, B, X, U, R, $W1$, $W2$を倍精度実数型として指定する。

(2) LSQE2S, LSQE2D

このサブルーチンは，右辺が複数個ある場合，即ち$AX=B$という形（B, Xは行列）の最小自乗問題を解くものである。イタレイションによって解の精度を上げる試みをしているので，インプットであるAとBの値は保存される。

呼び出し形式

LSQE2S（M, N, NR, A, KA, B, X, KX, U, R, $W1$, $W2$, $W3$, ILL）

インプット

M, N, NR：整数型変数名又は整定数

Mは方程式の数を，Nは未知数の数を表わす。

NRは方程式の右辺の列の数

$2 \leqslant N \leqslant M$, $2 \leqslant NR$

A：大きさKAの2次元アレイ

$M \times N$の長方行列Aの要素をストアする。

KA：整数型変数名又は整定数

アレイA，B，Uの大きさを指定，$KA \geqq M$

B：大きさKAの2次元アレイ

方程式の右辺である$M \times NR$　行列Bの要素をストアする。

KX：整数型変数名又は整数定数

アレイの大きさを指定，$KX \geqq N$

アウトプット

X：大きさKXの2次元アレイ

$N \times NR$行列で最小自乗問題の解がストアされている。

U：大きさKAの2次元アレイ

$M \times N$行列Uの要素がストアされている。

R：大きさ$N(N+1)/2$の1次元アレイ

3角行列Rの要素がストアされている。

図4－1を参照。

$W1$：大きさMの1次元アレイ

ワーキングエリアとして使用。

$W2, W3$：大きさNの1次元アレイ

ワーキングエリアとして使用。

ILL：整数型変数名

＝0　正常に終了したとき。

＝30000　$M<N$又は$KA<M$又は$KX<N$又は$N<2$又は$NR<2$のとき。

＝K　$1 \leqq K \leqq NR$　右辺の第K番目のベクトルに対してイタレイションが収束しなかったとき。

エラー処理

$N<2$又は$M<N$又は$KA<M$又は$KX<N$又は$NR<2$のときにはILL値をセットし，さらに改頁してその旨の印字をする。この場合それ以上の計算はせずにそのままリターンする。

右辺の第K番目のベクトルに対してイタレイションが収束しなかったときには，ILL値をセットし，さらに改頁して"SUB LSQE2S IS FAIL"と印字する。このときXの第$K-1$列までには正常に終了した収束解がストアされている。

収束判定

LSQE1Sと同じイタレイションを右辺の各ベクトルごとに行い，収束判定も各ベクトルごとに同様の基準が適用される。

詳細説明

3角行列Rの要素の配列の仕方については図4－1を参照。

LSQE2Dを用いるときにはA，B，X，U，R，$W1$，$W2$，$W3$を倍精度実数型として指定する。

(3) OTIVNS, OTIVND

このサブルーチンは，正方行列Aの逆行列AIを$AI=R^{-1}U^T$として求めるものである。インプット　あるAの値は失われる。

呼び出し形式

OTIVNS（N，A，KA，R，$W1$，ILL）

インプット

N：整数型変数名又は整数定数

正方行列Aの次数，$2 \leqq N$

A：大きさKAの2次元アレイ

正方行列Aの要素をストアする。

KA：整数型変数名又は整数定数

アレイAの大きさを指定，$KA \geqq N$

アウトプット

A：行列Aの逆行列の要素がストアされている。

R：大きさ$N(N+1)/2$の1次元アレイ

3角行列Rの要素がストアされている。

$W1$：大きさNの1次元アレイ

ワーキングエリアとして使用

ILL：整数型変数名

＝0　正常に終了したとき。

＝30000　$KA<N$又は$N<2$のとき。

エラー処理

$N<2$又は$KA<N$のときにはILL値をセットし，さらに改頁してその旨の印字をする。この場合それ以上の計算はせずにそのままリターンする。

詳細説明

このサブルーチンではイタレイションを行なわないので，行列Aの値を保存しておく必要がない。そこでAから計算されるUの要素をアレイAにストアし，ついでこれから計算されるAの逆行列の要素を又アレイAにストアしている。

3角行列Rの要素の配列の仕方については図4－1を参照。

OTIVNDを用いるときにはA，R，$W1$を倍精度実数型として指定する。

(4) OTIVIS, OTIVID

このサブルーチンは，正方行列Aの逆行列AIを$AI=R^{-1}U^T$で求め，その後イタレイションによって精度を上げている。インプットであるAの値は保存される。

呼び出し形式

OTIVIS（N，A，KA，AI，U，R，$W1$，$W2$，$W3$，ILL）

インプット

N：整数型変数名又は整定数

正方行列Aの次数，$2 \leqq N$

A：大きさKAの2次元アレイ

正方行列Aの要素をストアする。

KA：整数型変数名又は整定数

アレイA，AI，Uの大きさを指定，$KA \geqq N$

アウトプット

AI：大きさKAの2次元アレイ

行列Aの逆行列の要素がストアされている。

U：大きさKAの2次元アレイ

N次正方行列Uの要素がストアされている。

R：大きさ$N(N+1)/2$の1次元アレイ

3角行列Rの要素がストアされている。

$W1$，$W2$，$W3$：大きさNの1次元アレイ

ワーキングエリアとして使用。

ILL：整数型変数名

＝0　正常に終了したとき。

＝30000　$KA<N$又は$N<2$のとき。

＝1　イタレイションが収束しなかったとき。

エラー処理

$N<2$又は$KA<N$のときにはILL値をセットし，さらに改頁してその旨の印字をする。この場合それ以上の計算はせずにそのままリターンする。

イタレイションが収束しなかったときには，ILL値をセットし，さらに改頁して"SUB OTIVIS IS FAIL"と印字する。このときAIには最終近似逆行列の要素がストアされている。

収束判定

イタレイションは

$$A_{n+1}=A_n+E_n$$

E_nの各列ベクトルはLSQE2Sと同じ方法で計算

に従って行なう。収束の判定は

$$\|A\|_E=\left(\sum_{i,j=1}^{N} a_{ij}^2\right)^{1/2}$$

としたとき

$$\|E_n\|_E>\|A_n\|_E/2$$

となると，収束しないとして$ILL=1$をセットして終了する。

$$\|E_n\|_E/\|A_n\|_E<\varepsilon \quad \varepsilon=2^{1-t}$$

t：仮数部桁数

又は　$\|E_n\|_E>\|E_{n-1}\|_E/10$

となったときに収束したとして正常に終了する。

詳細説明

このサブルーチンではイタレイションによって解の精度を上げているので，行列A及びUの要素を保存しておく必要がある。そのために，OTIVNSに比べてメモリーを多く必要とする。

3角行列Rの要素の配列の仕方については図4－1を参照。

OTIVIDを用いるときにはA，AI，U，R，$W1$，$W2$，$W3$を倍精度実数型として指定する。

4－5 数値テスト及び結果

最小自乗問題を解くルーチンのテストは，LSQE2SとLSQE2Dを用いて，2章の(1)～(4)の最小自乗問題について行った。結果は表2－1～表2－3である。テスト1の単精度，テスト3については収束せずに，$ILL=1$が出力されてきた。

ここでも表を見てわかるように，残差は正解のそれにほぼ一致しており，残差が0となるベクトルに対しては，"収束解"の全桁が正しい。しかし，残差が0でないものについては，"収束解"でも全桁が正しいとは限らない。テスト1の倍精度，テスト4についてはLSQHDD，LSQHDSよりも1桁ずつ精度が良いが，テスト2の倍精度では逆にLSQE2Dの方が1桁精度が悪い。

テスト3については，ここでも本来このルーチンでは計算不可能である。このときの3角行列Rの対角成分の値は，（上が単精度，下が倍精度）

（872，264.5，343.8，4.741×10^{-14}，1.160×10^{-14}）

（872，264.5，343.8，2.581×10^{-35}，1.866×10^{-35}）

である。これを見ると，Aのランクは3であると判断してさしつかえない。

正方行列の逆行列を求めるルーチンについては，表2－4と表2－5の2つの行列について行った。表には相対誤差の行列を掲げてある。X_SはOTIVNS，X_DはOTIVND，Y_SはOTIVIS，Y_DはOTIVIDに対応するものである。aとfは，X_Sについて書けば

表2－1 テスト1

倍精度

	x_{1D}	x_{2D}
	1.00000000000000000 0.50000000000000000 0.33333333333333333 0.25000000000000000 0.20000000000000000	1.00000000000106584 0.50000000000361341 0.33333333333489477 0.25000000000068665 0.20000000000024503
a_m a_f	0.0 0.0	1.066×10^{-12} 1.139×10^{-12}
s_m s_f	0.0 0.0	1.066×10^{-12} 1.403×10^{-12}
r	1.352×10^{-26}	7.255×10^{7}
T_D $-\log s_m$ $-\log s_f$		14.43 11.97 11.85

表2－2 テスト2

単精度

	x_{1S}	x_{2S}	x_{3S}
	1.0000000 2.0000000 −1.0000000 3.0000000 −4.0000000	86 −93 3×10^{-5} −190 268	0.99911812 1.9990406 −0.99995470 2.9980310 −3.9971867
a_m a_f	0.0 0.0	2.678×10^{-3} 3.519×10^{-3}	2.813×10^{-3} 3.673×10^{-3}
s_m s_f	0.0 0.0		8.819×10^{-4} 1.391×10^{-3}
r	0.0	2.645×10^{8}	2.645×10^{8}
T_S $-\log s_m$ $-\log s_f$			3.616 3.055 2.857

$$a=\max_{i,j}|(X_S)_{ij}|$$

$$f=\sqrt{\sum_{i,j=1}^{n}(X_S)_{ij}^2}=\|X_S\|_E \quad n：行列の次数$$

の値である。第2のテスト行列に対する Y_S，Y_D は O 行列である。

第1の行列のスペクトルノルムに関する条件数 χ_1 は 3.664×10^6 であり，行列の条件としては甚だ悪い。第2の行列のそれは，$\chi_2=2.984\times10^3$ である。χ と計算機精度，解の精度の間の関係として，「1次方程式系の解法 I」[1] で考えたものと同じのを採ると，X_S，X_D の a の解の精度の代表として，

$26\log2-\log\chi_1=1.263$　　$\log a=0.3947$
$61\log2-\log\chi_1=11.80$　　$\log a=11.21$
$26\log2-\log\chi_2=4.352$　　$\log a=4.361$
$61\log2-\log\chi_2=14.89$　　$\log a=15.07$

となる。ここでも，この2つの行列に関する限り，解の精度は，計算機の精度よりも条件数のオーダー程度悪くなっているといえる。

表2－2

倍精度

	x_{1D}	x_{2D}	x_{3D}
	1.00000000000000000 2.00000000000000000 -1.00000000000000000 3.00000000000000000 -4.00000000000000000	-19 -21 4 $\times10^{-14}$ -46 77	0.99999999999807834 1.99999999999978771 -0.99999999999959278 2.99999999999954037 -3.99999999999923328
a_m a_f	0.0 0.0	7.730×10^{-13} 9.469×10^{-13}	7.667×10^{-13} 9.396×10^{-13}
s_m s_f	0.0 0.0		1.922×10^{-13} 3.318×10^{-13}
r	0.0	2.645×10^{8}	2.645×10^{8}
T_D $-\log s_m$ $-\log s_f$			14.15 12.72 12.48

表2－3　　テスト4

単精度

	x_{1S}	x_{2S}	x_{3S}
	5.0000000 4.0000001 2.9999999 1.9999999 0.99999994	5.0000000 4.0000001 2.9999999 1.9999999 0.99999994	5.0000000 4.0000000 2.9999999 2.0000000 1.0000000
a_m a_f	1.192×10^{-7} 1.712×10^{-7}	1.192×10^{-7} 1.577×10^{-7}	5.960×10^{-8} 6.664×10^{-8}
s_m s_f	5.960×10^{-8} 8.266×10^{-8}	5.960×10^{-8} 7.565×10^{-8}	2.981×10^{-8} 3.583×10^{-8}
r	4880	2577	1913
T_S $-\log s_m$ $-\log s_f$	5.983 7.225 7.083	6.121 7.225 7.121	6.186 7.526 7.446

表 2－4

$$A=\begin{bmatrix} -74 & 80 & 18 & -11 & -4 & -8 \\ 14 & -69 & 21 & 28 & 0 & 7 \\ 66 & -72 & -5 & 7 & 1 & 4 \\ -12 & 66 & -30 & -23 & 3 & -3 \\ 3 & 8 & -7 & -4 & 1 & 0 \\ 4 & -12 & 4 & 4 & 0 & 1 \end{bmatrix} \qquad A^{-1}=\begin{bmatrix} 1 & 0 & -7 & -40 & 131 & -84 \\ 0 & 1 & 7 & 35 & -112 & 70 \\ -2 & 2 & 29 & 155 & -502 & 319 \\ 15 & -12 & -192 & -1034 & 3354 & -2130 \\ 43 & -42 & -600 & -3211 & 10406 & -6595 \\ -56 & 52 & 764 & 4096 & -13276 & 8421 \end{bmatrix}$$

$$X_S=10^{-3}\times\begin{bmatrix} -210 & -145 & -46 & -1 & 0.7 & 0.3 \\ 179 & 123 & -40 & -1 & 0.7 & 0.3 \\ -403 & 278 & -43 & -1 & 0.7 & 0.3 \\ -359 & 309 & -43 & -1 & 0.7 & 0.3 \\ -388 & 274 & -43 & -1 & 0.7 & 0.3 \\ -380 & 282 & -43 & -1 & 0.7 & 0.3 \end{bmatrix} \qquad Y_S=10^{-7}\times\begin{bmatrix} -3 & 10 & -5 & 4 & -3 & -5 \\ 2 & -9 & -4 & 3 & -3 & -5 \\ -5 & -20 & -5 & 3 & -3 & -5 \\ -4 & -22 & -5 & 4 & -3 & -5 \\ -5 & -20 & -5 & 3 & -3 & -5 \\ -5 & -20 & -5 & 3 & -3 & -5 \end{bmatrix}$$

$a=4.030\times10^{-1}$
$f=1.019$

$a=2.223\times10^{-6}$
$f=4.870\times10^{-6}$

$$X_D=10^{-14}\times\begin{bmatrix} 8 & -288 & 27 & -8 & 2 & 2 \\ -7 & 245 & 23 & -7 & 2 & 2 \\ 16 & 550 & 25 & -8 & 2 & 2 \\ 14 & 613 & 25 & -8 & 2 & 2 \\ 15 & 543 & 25 & -8 & 2 & 2 \\ 15 & 559 & 25 & -8 & 2 & 2 \end{bmatrix} \qquad Y_D=\begin{bmatrix} 0 & 6\times10^{-36} & 0 & 0 & 0 & 0 \\ 8\times10^{-37} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$a=6.125\times10^{-12}$
$f=1.197\times10^{-11}$

表 2－5

$$A=\begin{bmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{bmatrix} \qquad A^{-1}=\begin{bmatrix} 68 & -41 & -17 & 10 \\ -41 & 25 & 10 & -6 \\ -17 & 10 & 5 & -3 \\ 10 & -6 & -3 & 2 \end{bmatrix}$$

$$X_S=10^{-6}\times\begin{bmatrix} 5 & -15 & -27 & 43 \\ 5 & -14 & -28 & 44 \\ 5 & -15 & -23 & 36 \\ 6 & -15 & -23 & 32 \end{bmatrix} \qquad X_D=10^{-17}\times\begin{bmatrix} -8 & 22 & 41 & -85 \\ -9 & 22 & 42 & -85 \\ -9 & 23 & 35 & -71 \\ -9 & 22 & 34 & -63 \end{bmatrix}$$

$a=4.351\times10^{-5}$
$f=9.869\times10^{-5}$

$a=8.546\times10^{-16}$
$f=1.781\times10^{-15}$

5章　Singular Value Decomposition

5－1　方法の基礎

Aを$m\times n$（$m\geqq n$）の行列とする。このとき

$$U^TU=I_n \qquad I_n：n次単位行列$$

なる$m\times n$行列Uとn次直交行列V，対角成分がσ_iなる対角行列Σの3つの行列により，

$$A=U\Sigma V^T \tag{5.1}$$

とAを分解できる。UはAA^Tの大きい方からn個の固有値に対応する正規直交化された固有ベクトルより成り，VはA^TAの正規直交化された固有ベクトルより成る。σ_iはA^TAの固有値の非負平方根である。この値のことをAのsingular valueという。今

$$\sigma_1\geqq\sigma_2\geqq\cdots\cdots\geqq\sigma_n\geqq 0$$

と仮定する。Aのランクrが$r<n$のときには，

$$\sigma_{r+1}=\sigma_{r+2}=\cdots\cdots=\sigma_n=0$$

である。(5.1)のような分解のことをsingular value decomposition（SVD）という。

異なる分解の仕方として，AA^Tの自明な固有値0に対応する$m-n$個の固有ベクトル$\hat{U}$をつけ加えて，

$$\tilde{U}=(U,\hat{U}),\quad \tilde{\Sigma}=\begin{pmatrix}\Sigma\\0\end{pmatrix}\Big\}m-n$$

とし，

$$A=\tilde{U}\tilde{\Sigma}V^T$$

とするのもある。又，非零singular valueに応じるものばかりに圧縮して，

$$A=U_r\Sigma_rV_r^T \qquad U_r^TU_r=V_r^TV_r=I_r$$

とも表わせる。Σ_rはAの非零singular valuesより成る対角行列である。

5－2　計算方法とサブルーチンの説明

このような分解をするために，直接AA^T, A^TAを計算するということはしない。その代わりに，まずハウスホルダー変換によって行列Aを2重対角行列に変換し，ついで固有値，固有ベクトルの計算を行なう。

5－2－1　2重対角行列への変換

一連のハウスホルダー変換

$$P^{(k)}=I-2x^{(k)}x^{(k)T} \qquad k=1,2,\cdots,n$$
$$Q^{(k)}=I-2y^{(k)}y^{(k)T} \qquad k=1,2,\cdots,n-2$$

$$x^{(k)T}x^{(k)}=y^{(k)T}y^{(k)}=1$$

を用いて，Aを次式のような上2重対角行列に変形する。

$$P^{(n)}\cdots P^{(1)}AQ^{(1)}\cdots Q^{(n-2)}=\begin{bmatrix}q_1 & e_2 & & & \\ & q_2 & e_3 & & \\ & & \ddots & \ddots & \\ & & & e_n & \\ & & & q_n & \\ 0 & & & & \end{bmatrix}\equiv J^{(0)}\Big\}(m-n)\times n$$

$A^{(1)}=A$として

$$A^{(k+1/2)}=P^{(k)}A^{(k)} \qquad k=1,2,\cdots,n$$
$$A^{(k+1)}=A^{(k+1/2)}Q^{(k)} \qquad k=1,2,\cdots,n-2$$

で定義すると，$P^{(k)}$は

$$a_{ik}^{(k+1/2)}=0 \qquad i=k+1,\cdots,m$$

となるように，$Q^{(k)}$は

$$a_{kj}^{(k+1)}=0 \qquad j=k+2,\cdots,n$$

となるように定める。

$J^{(0)}$のsingular valuesはAのそれと同じであり，$J^{(0)}$のsingular value分解を

$$J^{(0)}=G\Sigma H^T$$

とすると，Aのそれは

$$A=(PG)\Sigma(QH)^T$$

となる。ここで，$P=P^{(0)}\cdots P^{(n)}$，$Q=Q^{(1)}\cdots Q^{(n-2)}$である。これより，$U=PG$，$V=QH$となる。

5－2－2　2重対角行列$J^{(0)}$のsingular value分解

QR－アルゴリズムを用いて，$J^{(0)}$を逐次的に対角化していく。但し，直接的にQR－法を用いるのではなく，以下のような形で使用する。一連の直交行列$T^{(i)}$と$S^{(i)}$とで，

$$J^{(i+1)}=S^{(i)T}J^{(i)}T^{(i)}$$

という変換を行ない，$T^{(i)}$は$M^{(i)}=J^{(i)T}J^{(i)}$が対角行列に収束するように，又$S^{(i)}$はすべての$J^{(i)}$が2重対角行列であるように選ぶ。このために，$J=J^{(i)}$から$\bar{J}=J^{(i+1)}$への変換を，Jに右からと左からと交互に平面回転を施すことによって行なう。即ち，T_k，S_kを平面回転として，

$$\bar{J}=S_n^TS_{n-1}^T\cdots S_2^TJ\,T_2T_3\cdots T_n$$

と変換し，T_k，S_kを次のように選ぶ。

T_2	要素$(J)_{21}$を生成
S_2	要素$(J)_{21}$を消去，要素$(J)_{13}$を生成
T_3	要素$(J)_{13}$を消去，要素$(J)_{32}$を生成
⋮	⋮　，　⋮
S_n	要素$(J)_{n,n-1}$を消去

この一連の経過については図5－1を参照。ここで，まだ定まっていないT_2を，J^TJから$\bar{J}^T\bar{J}$への変換が，シフトつきのQR－変換と同等になるように定める。

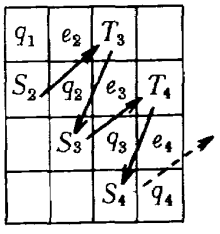


図5－1
q_i：Jの対角成分
e_i：Jの非対角成分

固有値が分離されたかどうかの判定は次のようにして行なう。

$$\delta=\varepsilon\cdot\max_i(|q_i|+|e_i|) \qquad \varepsilon：計算機精度$$

として，

$$|e_n|\leqq\delta$$

となれば，$|q_n|$ をsingular valueとし，行列の次数を1つ小さくする。

サブルーチンSVDSは，この方法に基ずいてAを$U\Sigma V^T$と分解する。

サブルーチンMINFISも同様に，Aを$\widetilde{U}\widetilde{\Sigma}V^T$解し，同時に行列積$\widetilde{U}B$も計算する。このサブルーチンでは$m<n$であってもよい。

なお，どちらのサブルーチンでも，求まったsingular valuesは大きさの順には並んでいない。

5－3 用　途

Aを$U\Sigma V^T$と分解すれば，Aの擬逆行列$A^\dagger$が次のようにして計算できる。（擬逆行列については4－1を参照）

$$A^\dagger=V\Sigma^\dagger U^T$$

ここで$\Sigma^\dagger=(\sigma_i^\dagger)$は

$$\sigma_i^\dagger=\begin{cases}\sigma_i^{-1} & \sigma_i>0\\ 0 & \sigma_i=0\end{cases}$$

として定まる対角行列である。

行列Aのランクrが$r<n$のとき，同次方程式

$$Ax=0$$

の解が，次のようにして求まる。このときには，

$$\sigma_i=0 \qquad i=r+1,\cdots,n$$

であり，$A=U\Sigma V^T$であることより，行列U，Vの列ベクトルをそれぞれu_i，v_iとすれば

$$Av_i=\sigma_i u_i \qquad i=1,\cdots,n$$

と書ける。従って，$i=r+1,\cdots,n$に対しては

$$Av_i=0$$

となり，これが同次方程式の解となる。

bを与えられたベクトルとして，$\|b-Ax\|$を最小にするという最小自乗問題の解xが次の式に従って求まる。

$$x=A^\dagger b=V\Sigma^\dagger U^T b=V\Sigma^\dagger c$$

このとき，Aのランクがnより小さいときには，このままでは解が一意には定まらないが，さらに$\|x\|$を最小にするという条件をつければ解は一意的に定まる。その解も同じ式によって計算される。

5－4　サブルーチンの使用方法

(1) SVDS, SVDD

このサブルーチンは長方行列Aを$A=U\Sigma V^T$とsingular value分解するものである。行列UとVの要素は必要に応じて計算されるようになっている。

呼び出し形式

SVDS（M, N, A, KA, U, V, KV, S, LU, LV, $W1$, ILL）

インプット

M, N：整数型変数名又は整定数
Mは行列Aの行数を，Nは列数を表わす。$M\geqq N\geqq 2$

A：大きさKAの2次元アレイ
$M\times N$行列Aの要素をストアする。

KA：整数型変数名又は整定数
アレイA, Uの大きさを指定，$KA\geqq M$

KV：整数型変数名又は整定数
アレイVの大きさを指定，$KV\geqq N$

LU, LV：論理型変数名又は論理型定数
アウトプットの行列Uが必要なときには，$LU=.TRUE.$，必要でないときには，$LU=.FALSE.$
アウトプットの行列Vが必要なときには，$LV=.TRUE.$，必要でないときには，$LV=.FALSE.$

アウトプット

U：大きさKAの2次元アレイ
$LU=.TRUE.$のとき，正規直交化された列をもつ$M\times N$行列Uの要素がストアされている。$LU=.FALSE.$のときには，ワーキングエリアとして使用される。

V：大きさKVの2次元アレイ
$LV=.TRUE.$のとき，N次直交行列Vの要素がストアされている。$LV=.FALSE.$のときには，Vは使用されない。

S：大きさNの1次元アレイ
行列Aの singular value がストアされている。

$W1$：大きさNの1次元アレイ
ワーキングエリアとして使用

ILL：整数型変数名
＝0　正常に終了したとき
＝30000　$M<N$又は$N<2$又は$KA<M$又は$KV<N$

エラー処理

$N<2$又は$M<N$又は$KA<M$又は$KV<N$のときには，ILL値をセットし，さらに改頁してその旨の印字をする。この場合それ以上の計算はせずにそのままリターンする。

詳細説明

$LV=.FALSE.$のときアレイVは使用されないので，このときには仮引数V, KVの実引数は型が合っていて，$KV\geqq N$を満足していれば何であってもよい。特に，Vがアレイである必要もない。一方Uについては，$LU=.FALSE.$であってもワーキングエリアとして使用されるので，実アレイが対応していなければならない。仮引数AとUの実引数が異っていれば，元のAの値は保存されるが，同じであれば$LU=.TRUE.$のときアレイAに行列Uの要素がストアされ，$LU=.FALSE.$のときアレイAはワーキングエリアとして使用される。従って，行列Aの singular values が欲しいだけのときには，

CALL　SVDS（M, N, A, KA, A, A, KA, S, $.FALSE.$, $.FALSE.$, $W1$, ILL）

でよい。

SVDDを用いるときには，A, U, V, S, $W1$を倍精度実数型として指定する。

(2) MINFIS, MINFID

このサブルーチンは，$m\times n$行列Aを$A=\tilde{U}\tilde{\Sigma}V^T$と singular value 分解する。$A$の要素がストアされていた場所に$V$の要素をストアするので，元の$A$の値は失われる。これと同時に，$m\times r$行列$B$に対して，$C=\tilde{U}^TB$の値が計算され，$B$がストアされていた場所に$C$をストアする。

呼び出し形式

MINFIS（M, N, NR, AB, KA, Q, $W1$, ILL）

インプット

M, N, NR：整数型変数名又は整定数
Mは行列Aの行数を，Nは列数を表わす。
NRは，最小自乗問題を解く場合の右辺の列数。
M, $N\geqq 2$, $NR\geqq 0$

AB：大きさKAの2次元アレイ
$\max(M,N)\times(N+NR)$行列であって，行列Aと行列Bの要素を次のようにストアする。

$AB(I,J)=a_{ij}$　　$1\leqq I\leqq M$, $1\leqq J\leqq N$
$AB(I,N+J)=b_{ij}$　　$1\leqq I\leqq M$, $1\leqq J\leqq NR$

図5－2を参照。

KA：整数型変数名又は整定数

アレイAの大きさを指定，$KA \geqq \max(M, N)$

アウトプット

AB：Vの要素とCの要素が次のようにストアされている。

$$AB(I, J) = v_{ij} \qquad 1 \leqq I,\ J \leqq N$$
$$AB(I, N+J) = c_{ij} \qquad 1 \leqq I \leqq \max(M, N),\quad 1 \leqq J \leqq NP$$

図５－２を参照。

Q：大きさNの１次元アレイ

行列Aの singular values がストアされている。

$W1$：大きさNの１次元アレイ

ワーキングエリアとして使用

ILL：整数型変数名

＝０　正常に終了したとき。

＝ 30000　$N<2$又は$M<2$又は$KA<\max(M,N)$

エラー処理

$N<2$又は$M<2$又は$KA<\max(M, N)$のときには，ILL値をセットし，さらに改頁してその旨の印字をする。この場合それ以上の計算はせずにそのままリターンする。

詳細説明

このサブルーチンでは$M<N$であってもよい。このときには，$N-M$個の自明な零 singular values がある。

インプットとアウトプットのアレイABは下図のようになっている。

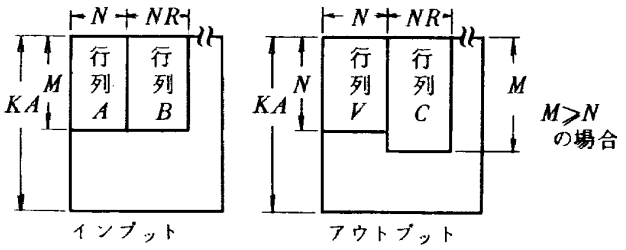


図５－２

行列$\tilde{U}$の要素が必要なときには，右辺BにM次の単位行列をセットしておけば（$M \geqq N$のとき），Cの所に$\tilde{U}^T$が計算されてくる。$M<N$のときには，

$$\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

という形の$N \times M$行列をセットしておき，$N-M$個の自明な零 singular values に対応する行を取り去ってできる$M \times M$行列が$\tilde{U}^T$を与える。

MINFIDを用いるときには，AB，Q，$W1$を倍精度実数型として指定する。

5－5　数値テスト及び結果

テスト1～テスト4の行列Aを，サブルーチンSVDS，SVDDを用いてsingular value分解した結果は，表3－1のように特徴づけられる。表中の記号は下記の通りである。

$A=(a_{ij})$，$Y=(y_{ij})$

$m(A)=\max|a_{ij}|$，$E(A)=\sqrt{\Sigma a_{ij}^2}$

$\bar{A}=U\Sigma V^T$，$X=A-\bar{A}$，$y_{ij}=(a_{ij}-\bar{a}_{ij})/a_{ij}$

この表から，行列U，Vの各列は殆んど正規直交化されているといってよい。又，$A=U\Sigma V^T$という分解も，テスト1の行列だけが若干悪いが，まずまずの精度であると判断できる。次に，これらの行列のsingular valueを単精度のものについてだけ，表3－2に掲げておく。そこでは，正確な値として倍精度で計算されたものの10桁を採用した。なお，相対誤差の欄で＊印のある所は絶対誤差の値である。

表を見ると，絶対値の大きいsingular valueほど精度よく求まっているのがわかる。これは，固有値を求めるルーチンにとって普通のことである。テスト3の行列には0というsingular valueが2つあるわけだが，それぞれSVDSでは2×10^{-7}と9×10^{-7}，SVDDでは6×10^{-18}と2×10^{-18}として求まっている。どちらの場合にも十分0と見なせる値である。これより，同次方程式$Ax=0$は，テスト3の行列を除いては自明な解$x=0$しかもたないことがわかる。（$\mathrm{rank}(A)=n$）テスト3の行列では，SVDSのアウトプットVの第2列と第5列が自明でない1次独立な解を与えてくれる。この解を単精度のものについて表にしたのが表3－3である。正規化されているため値は小数になっているが，0でない絶対値最小の要素で割り算して，適当整数倍することによって整数化すると，正確な解として，

（－137，144，－17，221，135）

（0，12，10，7，－23）

が得られる。

最小自乗問題を解くのには，サブルーチンMINFIS或はMINFIDを用いて，

$$x=A^{\dagger}b=V\Sigma^{\dagger}U^Tb=V\Sigma^{\dagger}c$$

という行列の掛け算によって求める。ここで，VとcはMINFISからのアウトプットABにセットされている。もちろん，中央の式に従ってSVDSによっても計算できる。ここで$\Sigma^{\dagger}$の計算にあたって，単精度では10^{-5}より小さいsingular valueを，倍精度では10^{-15}より小さいそれを0で置き代えた。その結果が表3－4～表3－7である。ここでも，解の精度が悪いときでも残差の値はほぼ正しいものになっている。又，テスト1，テスト2については残差が0になるベクトルに対しての方が精度はよくなっている。このルーチンでは残差の影響の他に，$A=U\Sigma V^T$と分解する精度の影響も考慮されねばならないだろう。

なお，このルーチンが前2者と最も異なる点は，テスト3のように$\mathrm{rank}(A)<n$となっている問題に対しても，$\|x\|$を最小にするという付帯条件の解を求められることにある。但し，このときには$n-\mathrm{rank}(A)$個の自明な零－singular valueがあるが，計算機では丁度0という値になって出力されることは殆んどないので，何らかの判定基準に従ってそれらの値を0に置き代えてやる必要がある。

表3－1

単精度

テストNo.	$m(X)$	$E(X)$	$m(Y)$	$E(Y)$	$m(I-U^TU)$	$m(I-V^TV)$
1	1.875×10^{-1}	3.206×10^{-1}	8.186×10^{-6}	1.429×10^{-5}	1.192×10^{-7}	8.941×10^{-8}
2	5.722×10^{-6}	1.317×10^{-5}	1.192×10^{-6}	1.782×10^{-6}	7.451×10^{-8}	8.941×10^{-8}
3	1.907×10^{-6}	3.088×10^{-6}	7.122×10^{-7}	1.039×10^{-6}	8.941×10^{-8}	5.960×10^{-8}
4	1.788×10^{-7}	4.059×10^{-7}	1.192×10^{-7}	2.882×10^{-7}	5.960×10^{-8}	5.960×10^{-8}

倍精度

テストNo.	$m(X)$	$E(X)$	$m(Y)$	$E(Y)$	$m(I-U^TU)$	$m(I-V^TV)$
1	3.688×10^{-12}	7.262×10^{-12}	4.305×10^{-16}	6.931×10^{-16}	1.735×10^{-18}	3.469×10^{-18}
2	3.331×10^{-16}	5.105×10^{-16}	7.932×10^{-18}	1.656×10^{-17}	2.168×10^{-18}	3.036×10^{-18}
3	4.857×10^{-17}	1.073×10^{-18}	1.214×10^{-17}	2.630×10^{-17}	3.469×10^{-18}	3.469×10^{-18}
4	5.204×10^{-18}	9.422×10^{-18}	1.741×10^{-18}	6.651×10^{-18}	1.301×10^{-18}	1.735×10^{-18}

表 3 － 2

テスト 1

$\bar{\sigma}_k$	$(\sigma_k-\bar{\sigma}_k)/\sigma_k$
8.8881582×10^{6}	2.197
6.9916152×10^{4}	−5.753
1.2492552×10^{2}	45.24×10^{-8}
1.8916602	38660
3.8969962×10^{1}	−703.0

テスト 2

$\bar{\sigma}_k$	$(\sigma_k-\bar{\sigma}_k)/\sigma_k$
1.7347876×10^{2}	0.8531
6.4858294×10^{1}	3.352
1.0034119×10^{1}	-0.1694×10^{-8}
9.0089682×10^{-1}	75.01
1.5990873×10^{-1}	137.6

テスト 3

$\bar{\sigma}_k$	$(\sigma_k-\bar{\sigma}_k)/\sigma_k$
3.5327044×10^{1}	−1.514
1.7087926×10^{-7}	-17.09×10^{-8} *
2.0000001×10^{1}	-5.000×10^{-8}
1.9595918×10^{1}	−0.2960
9.3931860×10^{-7}	-93.93×10^{-8} *

テスト 4

$\bar{\sigma}_k$	$(\sigma_k-\bar{\sigma}_k)/\sigma_k$
8.2423568	-1.329×10^{-8}
1.0000000	0.0
2.2502348	-4.684×10^{-8}
1.0000000	0.0
1.0000000	0.0

表 3 － 3

$\bar{v}_2$	$\bar{v}_5$	$v_2-\bar{v}_2$	$v_5-\bar{v}_5$	
−0.41909547	0.0	−1.5	0.0	(Def)
0.44050911	0.41854808	1.3	−1.6	
−0.052004562	0.34879006	1.3	−0.7	$\times10^{-8}$
0.67605914	0.24415303	0.0	0.7	
0.41297729	−0.80221714	1.3	1.8	

表 3 － 4

単精度　　テスト 1　　倍精度

	x_{1S}	x_{2S}
	0.97162955	0.067334449
	0.49080089	0.19346116
	0.32945643	0.20308967
	0.24832276	0.19333959
	0.19940837	0.17993266
a_m	2.837×10^{-2}	0.9327
a_f	3.013×10^{-2}	0.9922
s_m	2.837×10^{-2}	0.9327
s_f	3.650×10^{-2}	1.208
r	1.207×10^{-2}	7.255×10^{7}
T_S		3.896
$-\log s_m$	1.547	3.027×10^{-2}
$-\log s_f$	1.438	-8.215×10^{-2}

	x_{1D}	x_{2D}
	1.00000000000001649	1.00000000003793734
	0.50000000000004691	0.50000000012505591
	0.33333333333335179	0.33333333338655164
	0.25000000000000764	0.25000000002317488
	0.20000000000000262	0.20000000000821350
a_m	1.649×10^{-14}	3.794×10^{-11}
a_f	1.726×10^{-14}	4.037×10^{-11}
s_m	1.649×10^{-14}	3.794×10^{-11}
s_f	2.004×10^{-14}	4.922×10^{-11}
r	3.116×10^{-23}	7.255×10^{7}
T_S		14.43
$-\log s_m$	13.78	10.42
$-\log s_f$	13.70	10.31

表3－5

単精度 テスト2

	x_{1S}	x_{2S}	x_{3S}
	1.0000035	−6	0.99386187
	2.0000039	−7	1.9932397
	−0.99999948	1×10^{-8}	−0.99882910
	3.0000077	−15	2.9854727
	−4.0000067	24	−3.9760442
a_m	7.689×10^{-6}	2.410×10^{-2}	2.396×10^{-2}
a_f	1.147×10^{-5}	2.972×10^{-2}	2.949×10^{-2}
s_m	3.487×10^{-6}		6.138×10^{-3}
s_f	5.065×10^{-6}		1.048×10^{-2}
r	1.824×10^{-10}	2.645×10^{8}	2.645×10^{8}
T_S			3.616
$-\log s_m$	5.458		2.212
$-\log s_f$	5.295		1.980

倍精度

	x_{1D}	x_{2D}	x_{3D}
	0.99999999999999999	5	1.00000000000004893
	2.00000000000000000	5	2.00000000000005445
	−0.99999999999999999	-1×10^{-14}	−1.00000000000001384
	2.99999999999999998	12	3.00000000000012056
	−4.00000000000000000	−22	−4.00000000000021198
a_m	1.908×10^{-17}	2.155×10^{-13}	2.120×10^{-13}
a_f	1.922×10^{-17}	2.582×10^{-13}	2.550×10^{-13}
s_m	6.366×10^{-18}		5.300×10^{-14}
s_f	6.613×10^{-18}		8.804×10^{-14}
r	2.829×10^{-31}	2.645×10^{8}	2.645×10^{8}
T_D			14.15
$-\log s_m$			13.28
$-\log s_f$			13.06

表3－6 テスト3

単精度

	x_{1S}	x_{2S}	x_{3S}
	−0.083333334	−3	−0.083333334
	-2.346×10^{-9}	−11	-3.668×10^{-9}
	0.25000001	22×10^{-10}	0.25000001
	−0.083333349	4	−0.083333347
	0.083333336	5	0.083333336
a_m	1.490×10^{-8}	2.193×10^{-9}	1.490×10^{-8}
a_f	2.129×10^{-8}	2.550×10^{-9}	2.022×10^{-8}
s_m	1.788×10^{-7}		1.565×10^{-7}
s_f	1.898×10^{-7}		1.690×10^{-7}
r	1.372×10^{-13}	320	320
T_S			6.574
$-\log s_m$	6.748		6.806
$-\log s_f$	6.722		6.772

倍精度

	x_{1D}	x_{2D}	x_{3D}
	-0.083333333333333334 2.169×10^{-20} 0.250000000000000001 -0.083333333333333333 0.083333333333333336	1 -6 8 $\times10^{-19}$ 4 2	-0.083333333333333334 -4.522×10^{-19} 0.250000000000000002 -0.083333333333333330 0.083333333333333338
a_m a_f	6.505×10^{-19} 7.154×10^{-19}	8.315×10^{-19} 1.153×10^{-18}	1.518×10^{-18} 1.690×10^{-18}
s_m s_f	3.249×10^{-18} 4.407×10^{-18}		6.068×10^{-18} 9.324×10^{-18}
r	1.945×10^{-34}	320	320
T_D $-\log s_m$ $-\log s_f$	 17.49 17.36		17.11 17.22 17.03

表3－7 テスト4

単精度

	x_{1S}	x_{2S}	x_{3S}
	5.0000004 3.9999996 2.9999997 2.0000008 1.0000005	5.0000004 3.9999999 2.9999999 2.0000006 1.0000002	5.0000006 3.9999998 2.9999999 2.0000004 1.0000002
a_m a_f	8.345×10^{-7} 1.126×10^{-6}	5.960×10^{-7} 7.468×10^{-7}	5.960×10^{-7} 8.205×10^{-7}
s_m s_f	4.768×10^{-7} 6.515×10^{-7}	2.980×10^{-7} 3.900×10^{-7}	2.086×10^{-7} 3.262×10^{-7}
r	4880	2577	1913
T_S $-\log s_m$ $-\log s_f$	5.983 6.322 6.186	6.121 6.526 6.409	6.186 6.681 6.487

6章 結 語

4章でも述べたように，overoteterminedsystem

$$Ax = b$$

を最小自乗法で解く最も普通の計算方法は，

$$A^TAx = A^Tb \qquad (6.1)$$

を解くものである。こうすると，A^TAは正値対称行列になるので，これを解くには「1次方程式系の解法 I」[1]で述べた方法が採用できる。このときに失われる精度は，A^TAの条件数－行列が正定値の場合の条件数は（最大固有値）÷（最小固有値）－のオーダー程度である。（cf.「1次方程式系の解法 I」[1]） ここで用いたテスト行列(1), (2), (4)に対する条件数はそれぞれ，2.2×10^{13}，1.2×10^{6}，6.8×10^{1}であるから，(6.1)という方程式を「1次方程式系の解法 I」のサブルーチンで解いたときに失われる精度は，それぞれ大略13桁，6桁，2桁と予想される。（実際の結果は表5－2を参照） 上記の数字を本報告のサブルーチンによるものと比較すれば下表のようになる。（数字は失われた精度の桁数）

表4－1

	テスト1	テスト2		テスト4
	倍精度	単精度	倍精度	単精度
LSQHDS	8.1～8.2	4.7～4.9	5.0～5.3	0.6～0.8
LSQE2S	6.4～6.5	4.8～5.0	5.6～5.9	0.3～0.4
MINFIS	7.9～8.1	5.6～5.8	5.1～5.3	1.1～1.3

これを見ると，精度についてならテスト1の行列では，(6.1)を解くのは圧倒的に不利であるのがわかる。もう

一つ(6.1)を解くことの欠点は，行列の掛け算を計算する必要があるので計算量が多くなることである。このことはとりわけ行列の次数が大きくなったときに影響が大きくなる。（$m\times n$行列AについてA^TAを計算するには約$2n^2m$回の積と和の計算が必要である）

しかし，計算量を不問にし，A^TA, A^Tbが精度よく計算できれば，有効数字2桁程度の近似解が得られると，(6.1)の解はイタレイションによってかなりの精度まで求められる。ところが，本報告のルーチンではイタレイションした結果が表4－1である。ここのサブルーチンでは，残差が0に近い場合には殆んど正しい値が得られているが，残差が大きくなると有効数字の桁数が減っている。この点について改良しなければ，次数の小さい問題については(6.1)をイタレイションで解く方が無難だったということになってしまう。そのためには，それがプログラム技術的な改良によって可能なのか，或は方法を改良する必要があるのか，それとも又別な方法を開発しなければならないのか，といった検討等が今後の課題である。

なお，「1次方程式系の解法　Ⅰ」のサブルーチン『CLDT2S』と『CLSL2S』を用いて(6.1)を解いた結果は下表の通りである。又，イタレイションを行っているサブルーチン『ACLSLS』を用いた場合には，下表にあるテスト全部について正しい解が得られた。

表4－2

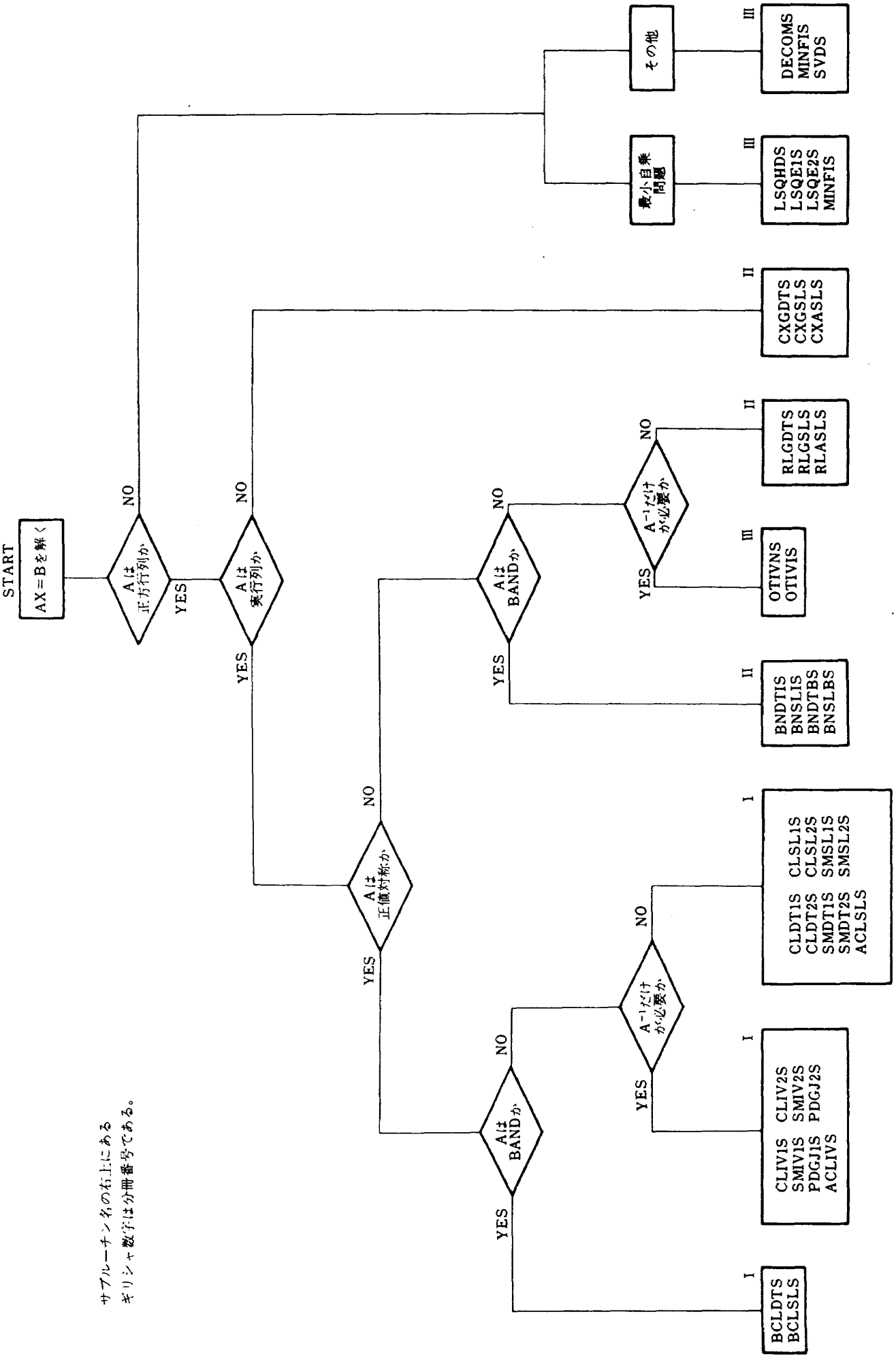
	テスト1	テスト2		テスト1
	倍精度	単精度	倍精度	単精度
$-\log s_m$	8.917	2.965	13.99	6.097
$-\log s_f$	8.809	2.746	13.76	6.035
失われた精度	9.4～9.6	4.9～5.1	4.4～4.6	1.7～1.8

テスト1の単精度は『CLDT2S』の中で計算続行が不可能となった。（$ILL=1$ cf.「1次方程式系の解法Ⅰ」）又，テスト1の右辺b_1, b_2に対する答として同じ値が出力されてきた。テスト2の右辺b_1, b_3，テスト4の右辺b_1, b_2, b_3に対しても同様である。さらにテスト2の右辺b_2に対しては正確に零ベクトルが出力されてきた。これらのことと，『ACLSLS』を使用した場合に全部正しい値が得られたこととは，A^TA, A^Tbの計算が桁落ちなく正確に行われた結果によるものと思われる。

本報告によって1次方程式系を取り扱うサブルーチンについては一応終了する。といっても完了する性格のものではなく，新しい手法の開発や異なる問題へのアプローチ等といった研究もあり，持続的研究が要求されるものである。なおこの後，行列の固有値，固有ベクトルを求める問題，さらには特殊函数の値を求めるルーチンの作成へと進む予定である。

参考文献

1) 福田正大他；1次方程式系の解法　Ⅰ，航技研資料TM－277（1975.5）



サブルーチン名の右上にある
ギリシャ数字は分冊番号である。

航空宇宙技術研究所資料315号

昭和51年10月発行

発行所　航空宇宙技術研究所
東京都調布市深大寺町1880
電話武蔵野三鷹(0422)47-5911(大代表)〒182

印刷所　株式会社　共進
東京都杉並区久我山4－1－7(羽田ビル)

Printed in Japar