数理解析研究所講究録
548巻 1985年 50-70

大規模消去演算における数値誤差（実験例）

岡山大学　谷口健男(Takeo Taniguchi)
岡山大学　曽我　明(Akira Soga)

1．序論

近年、電子計算機の著しい発達とともに有限要素法や差分法は工学的、数学的分野に広く一般に利用されるようになってきた。一般に工学的問題においては、これらの適用は線形な大型疎行列の問題に帰着することが多く、これらの大型疎行列に対して効果的な方法、例えばBand Matrix法やProfile法、Wave front法などの種々の解法を用いて解を得る。そして、これらの解法はいずれも消去法を基礎としている。ところが、この消去法を用いて得られた解の信頼性は、元数が増えれば次第に低下していくことが知られており、実際にその現象は、工学分野では数千元の系で顕著に現われてくると言われている。そこで、我々がこれらの解法を用いて解を求めるとなると一番心配なのが、どの程度の規模計算で、得られた解はどの程度まで信頼できるかということであり、その意

味で、解が包含する数値誤差の傾向や規模を調べてみることは興味深く、また重要なことである。

消去法を用いた時の数値誤差の研究を最初に行なったのがWilkinsonである[1]。そしてその後、Royの数値実験[2]などから、「大型疎行列の線形方程式において、十分信頼できる解を得ようとするならば、すべて倍精度演算を行なうべきである。」という提案がなされた。一方、解における有効桁数の推定として、行列の最大固有値と最小固有値の比で表わした条件数が良い評価を与えている、というのもよく知られた事実である[2]。しかしながら、これらの固有値の計算には多くの数値演算が必要であることより、一般に正しい条件数の算定は経済的な面からみても非常に困難であるといえる。

そこで本研究では、実際の解析法としてプログラミングやデータ構造が簡単で、記憶容量も小さくてすむことより、中規模以下の問題によく用いられているBand Matrix法で、一般的によく解かれる4000元程度までの係数行列を対象として単精度消去演算を行ない、消去演算過程で発生する数値誤差の伝播やその発生要因を明らかにする。さらに、最終的には、数値実験結果に基づいて、どの程度の規模計算でどの程度まで解が信頼できるかという、目安の検討を行なってみることにする。

2. 数値誤差

いま解くべき線形方程式を

$$Ax = b \qquad (1)$$

で表わす。ここに、Aは正定値対称で正則、また非常に多くの零要素を含む疎行列であり、xは未知ベクトル、bは既知ベクトルである。

ここでは、理論的な見地から解ベクトルxに現われる数値誤差について考えてみることにする。一般に、式(1)の線形方程式を電子計算機を用いて解く場合、発生する数値誤差の要因としては、

i) 入力データの不確定性

ii) 丸め誤差

iii) 打ち切り誤差

の3つが挙げられる。

2.1 入力データの不確定性

入力データが不確定性を有している場合、解ベクトルxには数値誤差の発生を引き起こす。いま、行列Aが不確定性dAを有しているとき、解ベクトルxに及ぼす影響をdxであるとすると、次式が成り立つ。

$$(A + dA)(x + dx) = b \qquad (2)$$

ここで、式(2)の両辺のノルムをとり、

$$\|dA\| / \|A\| \ll 1 \tag{3}$$

を仮定すれば、次式を得る。

$$\frac{\|dx\|}{\|x\|} \leq cond(A) \cdot \frac{\|dA\|}{\|A\|} \tag{4}$$

ここで cond(A) は行列Aの条件数であり、これは行列Aの最大固有値と最小固有値を用いて次式で定義される。

$$cond(A) = \frac{\lambda_{max}}{\lambda_{min}} \tag{5}$$

式(5)は、cond(A) は最大固有値 λ_{max} と最小固有値 λ_{min} の比が大きくなるに従って、大となることを示している。

以上のことより、式(4)は入力データの不確定性の解に与える影響度は cond(A) に依存することを示している。また条件数は行列Aの元数が大きくなれば大となる。それ故、式(1)に示すような大次元行列においては、解の精度を保つために有効桁数を増やさなければならない。本研究では、入力データは不確定性を有していないものと仮定する。

2.2 丸め誤差

我々が電子計算機を使用するかぎり、数値は有限な桁数ですべて評価される。そのため、得られた結果のほとんどが、計算機においては真値の近似値となる。データの丸めは各々の計算段階で行なわれ、我々は多くの数値処理の最終的な結

果からしかそれを得ることはできない。また丸め誤差は、消去演算においては、係数行列Aの元数が増えれば増えるほど混入が大であると考えられる。したがって丸め誤差の混入を防ぐためには、計算機の有効桁数をより多くとることが必要であり、一般に倍精度計算を行なうべきであるといえる。また丸め誤差が有限な桁数におけるデータの評価、あるいは処理に起因することより、その丸めの方法もまた重要であるといえる。一般に計算機における有効桁数以下の丸めの方式としては主に2つあり、一方が切り捨て方式であり、他方が四捨五入方式である。これらは、確率的に後者のほうが前者よりも精度の良い解を得ると考えられ(3)、本研究でも追実験を行なってみる。また、上記の考察より、解における丸め誤差は問題の物理特性に依存せず、むしろ計算機の特性、あるいは計算の仕方に依存しているといえる。

3.3 打ち切り誤差

式(1)において、いま係数行列Aの元数が十分に大きいと仮定すると、その時の条件数 cond(A) は式(5)で定義されたように非常に大きくなり、最小固有値あたりの低次の固有値の情報は、計算機の語長が一定で有限なものである限り必然的に解から失われていく。これが打ち切り誤差であり、問題の物理特性に依存しているものである。この解における打ち切り

誤差の上限値は式(1)、(2)より簡単に評価することができる。いま式(2)のdAを、計算機の一定語長計算のため初期の係数行列Aの各々のデータが失なった値であるとする。この時、Royの研究によれば次式[2]が成り立つ。

$$\frac{\|dA\|}{\|A\|} = 10^{-P} \tag{6}$$

ここで、Pは計算機において行列Aの各々の数値が評価される有効桁数を示す。これより、この時の解における有効桁数Sは、次式で与えることができる。

$$S \geq P - \log_{10}(\mathrm{cond}(A)) \tag{7}$$

この式(7)は、解において信頼できる桁数は係数行列Aの条件数に依存していることを示している。しかしながらここで注意しなければならないことは、式(1)で表わされる右辺の定数ベクトルの打ち切り誤差は、式(6)では全く考慮に入れていないということである。その理由は、実際には右辺の打ち切り誤差の影響は左辺の行列の打ち切り誤差とほぼ同じ大きさを有しているのであるが、式(7)は打ち切り誤差に対してかなりの過大評価をしており、この式だけで、解の信頼できる有効桁数を十分に評価していることによる[2]。

またRozanoffによると、この打ち切り誤差を含んでいる解は反復改善を行なっても回復は不可能であり、このことより

も、Roy の言ったように、完全な数値処理のためにはすべてを倍精度演算で行なう必要があるといえる。

いま、もし我々が単精度入力で倍精度消去演算を行なったとすれば、解における有効桁数 S は次式で表わすことができる。

$$\left.\begin{array}{l} \text{if} \quad \log_{10}(\mathrm{cond}(A)) \leq P \quad \text{then} \quad S = P \\ \text{and if} \quad P < \log_{10}(\mathrm{cond}(A)) < 2P \\ \qquad \text{then} \quad S \geq 2P - \log_{10}(\mathrm{cond}(A)) \end{array}\right\} \quad (8)$$

このように、式(7)、(8)で表されるように、打ち切り誤差は条件数 cond(A) に依存しており、初期に係数行列の条件数を求めることによって簡単に評価することができる。しかしながら、これに対して丸め誤差は、各段階ごとで発生するものであり、計算の仕方によっても異なる。すなわち、解の丸め誤差は用いる解法によっても異なるものであり、それ故に、ここでは消去法を基礎とする Band Matrix 法を取り上げて、その時の解に含まれる丸め誤差について調べてみる。

一般に式(1)を Gauss の消去法によって解く時、丸め誤差は入力された係数行列 A を式(9)のように分解することによって発生する。

$$A = LDU \quad (9)$$

ここで、L, D, U はそれぞれ下三角、対角、上三角行列

である。

しかしながらこの分解を計算機で行なうと、丸め誤差によって式(10)、

$$A - LDU \equiv -dA^{*} \tag{10}$$

のようになり、必ずしも0とはならない。ここでdA^{*}は容易に求めることができ、分解誤差e_{0}は次式で表わされる。

$$e_{0} = \frac{\|dA^{*}\|}{\|A\|} \tag{11}$$

よって、式(1)の係数行列Aを LDUに分解した時、解にdxの誤差をひきおこしたとすると次式が成り立つ。

$$(A + dA^{*})(x + dx) = b \tag{12}$$

ここで両辺のノルムをとり、dxが微小であるとして整理すれば、

$$\frac{\|dx\|}{\|x\|} \leq cond(A) \cdot \frac{\|dA^{*}\|}{\|A\|} = cond(A) \cdot e_{0} \tag{13}$$

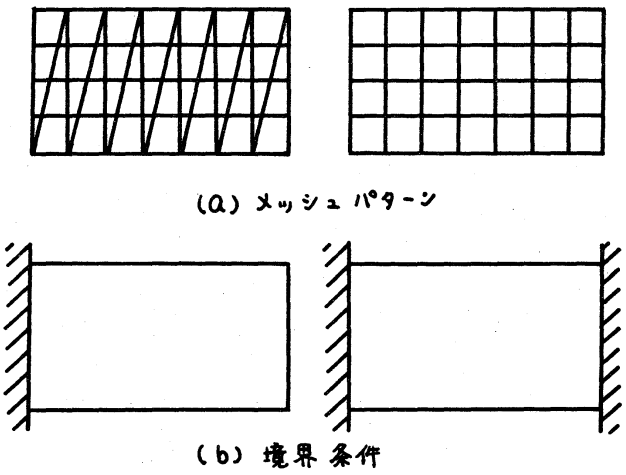
が得られる。この式は、条件数 cond(A)は行列固有の値であることより、対象の行列が変わらなければ一定値であるが、分解の仕方、あるいは計算の仕方によって $\|dA^{*}\| / \|A\|$は変化し、それにつれて解に発生する誤差も変化することになることを示している。

3. 数値実験及び結果

3.1 準備

i) モデル

数値実験に扱ったモデルは図－1に示す。ここでの数値実験の目的は、消去法を基礎とするBand-Matrix法を用いた時の解に発生する数値誤差の規模、及び傾向を知ることであり、そのために、境界条件としては比較的大きな数値誤差を発生すると考えられる2つの種類、すなわちcond(A)が大きくなるような境界を用いた。(図－1参照)係数行列は図－1の2つのメッシュパターンより作成されるため、非対角項の非零要素には-1が、また主対角項a_{ii}には$a_{ii} \geq \sum_{j=1}^{n} |a_{ij}|$で表わされる数値が入っている。このような行列は、工学問題では、一般に問題領域を均質で有限な要素に分割した時、あるいは均質な差分モデルなどによく見られる行列である。また荷重は、問題領域の全節点に単位荷重1.0を載下した。すなわち、式(1)の右辺の定数ベクトルはすべて1.0である。



図－1 数値実験モデル

ii) 誤差評価

発生した解の数値誤差の評価としては、次式によって最大

相対誤差を求め、比較に用いた。

$$\mathrm{Error} = \max_{i=1}^{n} \frac{|X_i(D) - X_i(S)|}{|X_i(D)|} \tag{14}$$

ここで$X_i(D)$と$X_i(S)$は、それぞれ倍精度、及び単精度によるi番目の解を示す。また式(14)において倍精度解を真値としてみなしたのは、次に示す数値実験によって倍精度解は十分に信頼できるということがわかったことによる。

異なる消去順序を用いて、上述したモデルに対し8000元程度までの係数行列を作成し、倍精度消去演算を行なったところ、各々の解は最悪でも約9桁まで一致した。よって、用いた計算機の単精度演算における有効桁数は約8桁であることより、単精度解との比較として倍精度解は十分信頼でき、真値とみなせると判断される。

(iii) 計算機の有効桁数以下の数値の取り扱い

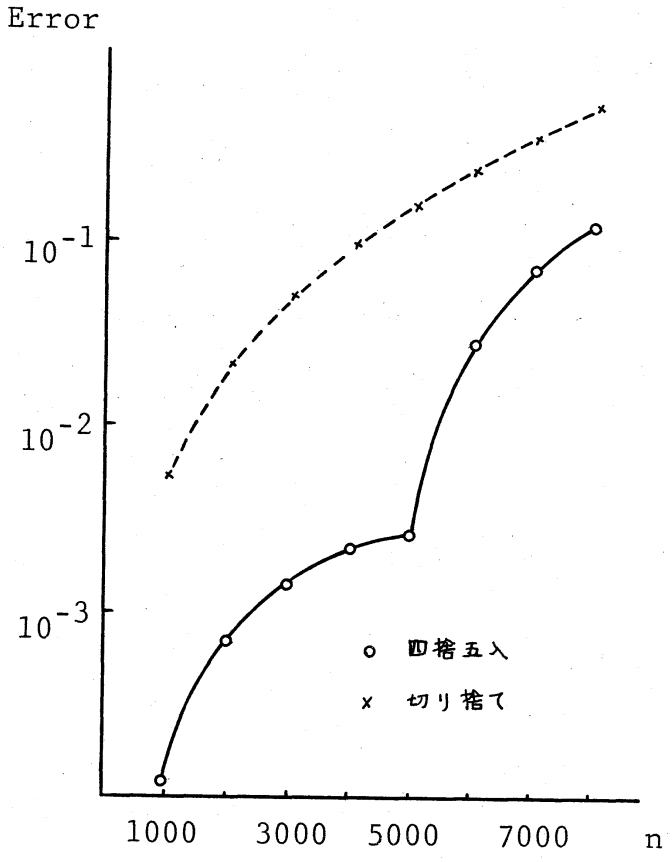


図-2 単精度消去演算における相対誤差

数値誤差の大きさが、計算機における有効桁数以下の数値の取り扱いに依存しているというのは、よく知られた事実である。一般にそれらの処理には、四捨五入方式と切り捨て方式とがある。ForsytheとMolerによれば、四捨五入方式は切り捨て方式よりもより良い結果を与える[3]、ということが言われているが、ここでは彼らの提言を数値実験により確かめてみる。結果を図－2に示す。これより、2つの手法による数値誤差の発生量にはかなりの差があることがわかる。しかも精度は四捨五入のほうが良い。よって消去演算においては、計算機における両者の選択が可能であるならば、四捨五入を用いるべきであるといえる。

以下の数値実験ではすべて四捨五入を用いることにする。

なおここで数値実験に使用した計算機は、ACOSシステム1000モデル20で、単精度演算は36ビットで有効桁数約8桁、倍精度演算は72ビットで有効桁数約18桁である。

3.2　数値実験方法

図－1に示すモデルを対象として、四捨五入方式で、Band-Matrix法を用いてそれぞれ倍精度消去演算、及び単精度消去演算を行ない、それらの解に対し式(14)の誤差評価式を用いて誤差を評価した上で、解の比較・検討を行なった。ここで、数値モデルにおける変数は、次に示す諸要因である。

i) 全節点数　n

ii) 構造系モデルの横幅　a

iii) 構造系モデルの縦幅　b

iv) 境界固定節点数

v) メッシュパターン

vi) 境界条件

vii) 消去順序

ここでは、これらの諸要因の組み合せをいろいろ変化させて数値実験を行なった。その結果、及び考察については次節で要約している。

3.3　実験結果及び考察

ここでの数値実験の主目的は、入力データに全く誤差を有しない時、対象とする線形方程式の消去演算過程で発生する丸め誤差の発達傾向を知ること、及びその発生要因

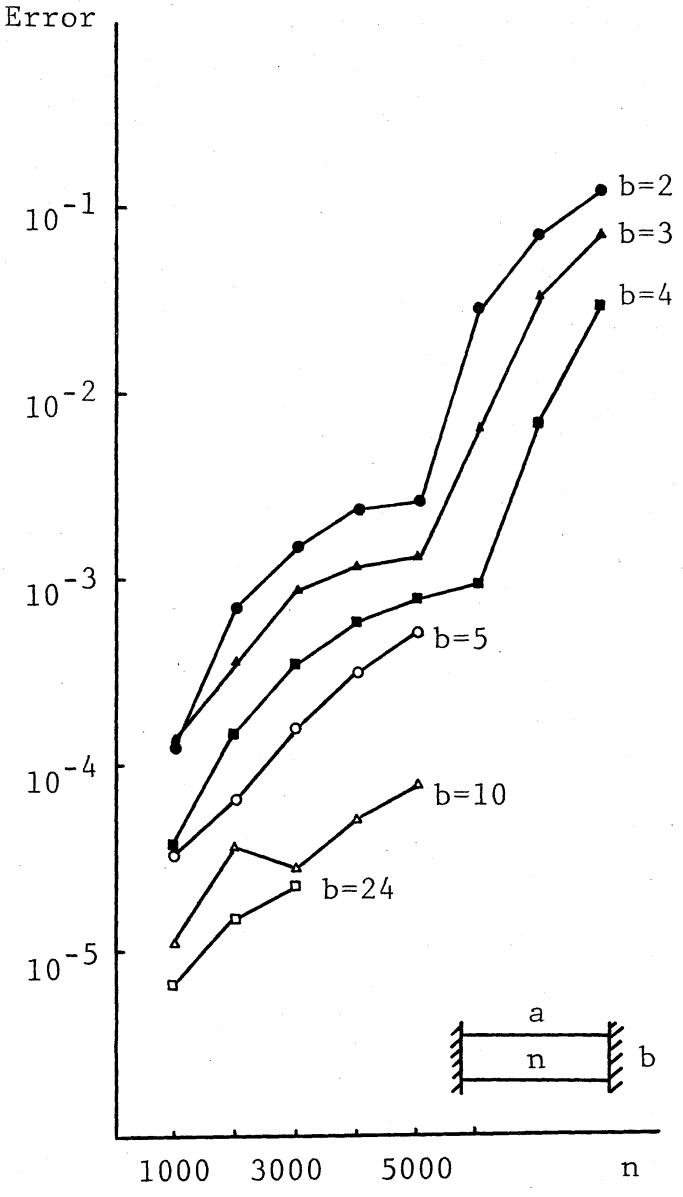
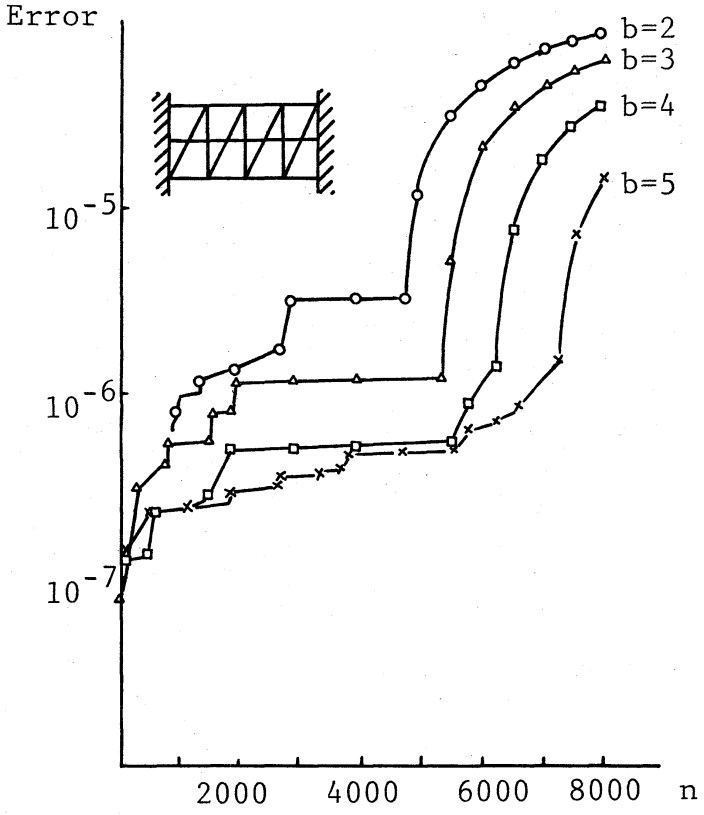


図-3　両端固定ばりの数値誤差①

に検討を加えることにある。そこで、まず図－3は横幅aと縦幅bを変化させることによって元数を変えた時の解の数値誤差の傾向を求めたものである。ただし境界条件は図－1の両端固定を用いている。これより誤差の傾向として、ある元数に達すると急激に誤差が増加することがわかる。



図－4　主対角項の数値誤差

さらに、この時の消去演算過程の主対角項の数値誤差を求めたものが図－4であるが。興味深いことに、解の誤差と主対角項の誤差の発達傾向がほとんど同一の傾向を示している。これは、用いたモデルによって作成された係数行列の半Band幅が非常に小さくて、非対角項よりも主対角項の影響のほうが上まわっているためと考えられる。さらにこの時の主対角項の値に注目してみたところ、ある元数以上になると主対角項は消去演算をされているにもかかわらずその値は変化しな

くなる、すなわち、計算機の分解機能が停止してしまい、行列特性を示す条件数に関係なく、誤差は累積していき、このような解の誤差が急激に増加する現象が見られると考えられる。

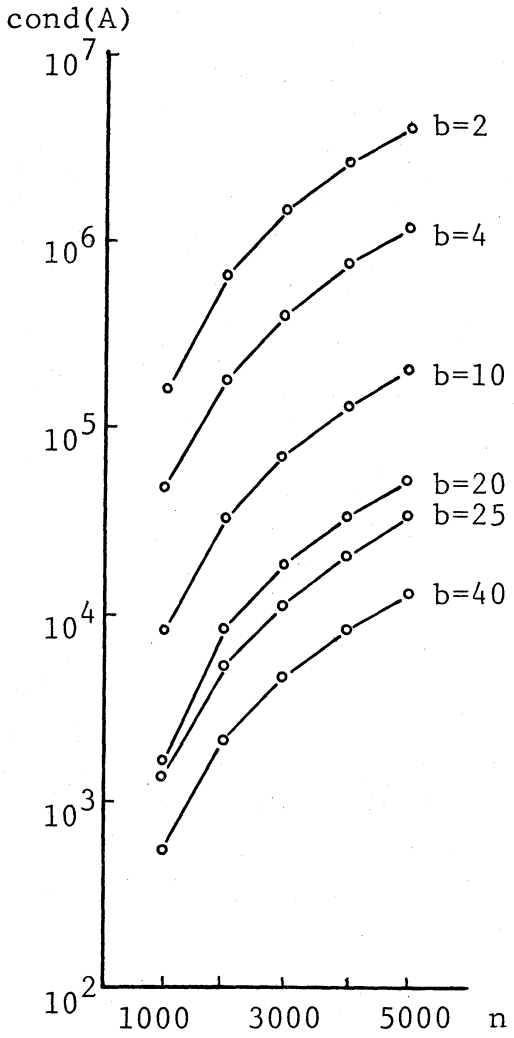


図-5　条件数

すなわちこれは、単精度消去演算の限界を示しており、それ以上の元数において消去演算を行なうことは無意味計算になりかねないことを意味している。また、今、この限界点をN_0とすれば、数値実験により最悪の場合を考えてみると、それは4000元あたりであるといえる。

よって次の結果を得る。

〔結果－1〕

単精度消去演算には計算機の機能上の面から限界があり、最悪の場合、それは4000元あたりでおとずれ、それ以上は無意味計算になりかねない。

次にこのモデルにおける条件数 cond(A)を求めてみた。条件数は式(5)で示されるように最大固有値と最小固有値の比で表わされるので、ここでは、λ_{max}は係数行列Aの行和の最大値で近似し、λ_{min}はベクトル逆反復法[4]を用いて求めた。

この手法は出発反復ベクトルx_1を仮定し、後は各反復ステップ$k=1,2,\cdots$で次式

$$\left.\begin{aligned} A\bar{x}_{k+1} &= x_k \\ x_{k+1} &= \frac{\bar{x}_{k+1}}{(\bar{x}_{k+1}^T\ \bar{x}_{k+1})^{1/2}} \end{aligned}\right\} \quad (15)$$

を計算していくことによって、固有ベクトルϕ_1の近似値を求め、さらにRayleigh商

$$\rho(\bar{x}_{k+1}) = \frac{\bar{x}_{k+1}^T\ x_k}{\bar{x}_{k+1}^T\ \bar{x}_{k+1}} \quad (16)$$

を用いて最小固有値λ_{min}の近似値を求める解法である。

これらの結果、得られた条件数をプロットしたのが図－5である。条件数とここで用いたモデルとの対比を行なってみると、元数nを一定にした時、

縦幅bの増加 ⟶ cond(A)は減少

縦幅bの減少 ⟶ cond(A)は増加

となることがわかる。これより、先程の単精度消去演算の限界N_{cr}に対して次に示す結果を得る。

[結果 - 2]

単精度消去演算における計算機の機能上の面から与えられる限界 N_{cr} は、条件数 cond(A) にも依存しており、N_{cr} の値は条件数が減少すれば、大きくなる。

次に元数が N_{cr} 以下の時の丸め誤差の傾向を求めてみた。結果は図-6,7に示す。図-6は両端固定ばりにおいて、横幅 a と縦幅 b を変化させた時の解に発生する誤差であり、図-7は一片固定の片持ちばりについて、消去順序を変えて、固

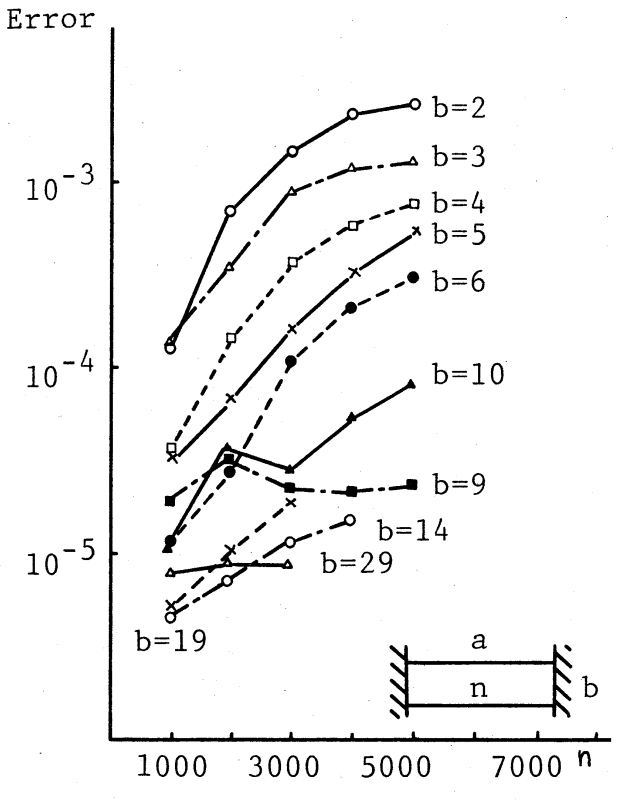


図-6 両端固定ばりの数値誤差 ②

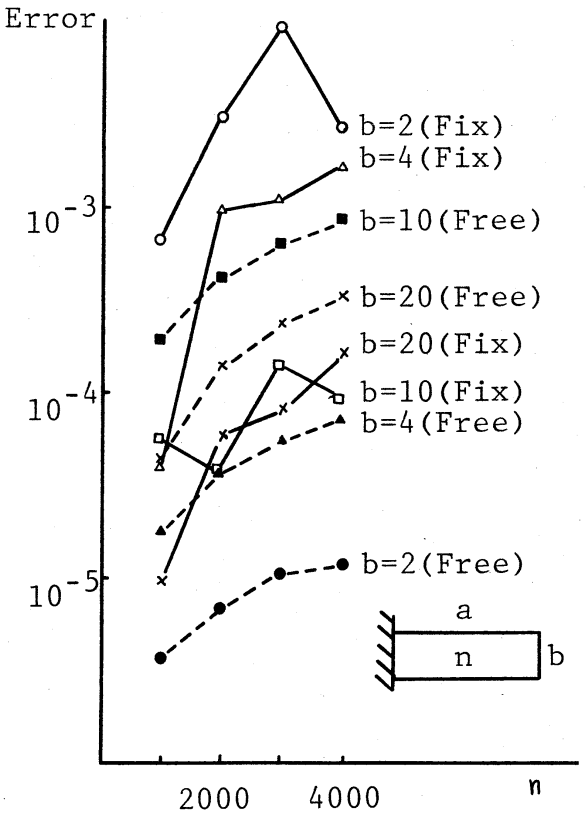
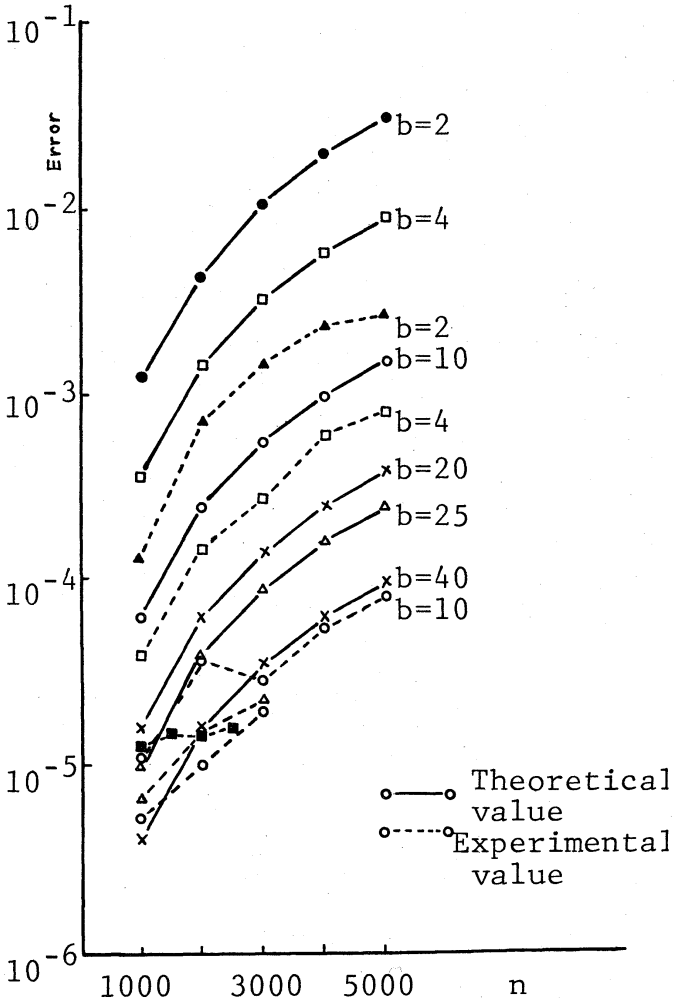


図-7 片持ちばりの数値誤差

定端側からの消去と、自由端側からの消去を行なった時の解に発生する数値誤差である。ここで式(13)の丸め誤差による係数行列Aの分解誤差について考えてみる。いま、図－5の条件数を式(7)により解の有効桁数に変換し、図－6と同一紙面上にプロットすると図－8のようになる。この結果は、式(13)によい評価を与えていることがわかる。またここで、縦幅bの増加は条件数の低下を示しているにもかかわらず、実験値でb＝9～29あたりで逆転するのは、bの増加は半Band幅の増加、すなわち演算回数の増加をもたらし、これが、式(13)の右辺の分解誤差 $\|dA^*\| / \|A\|$ の増加を促がしていると考えられる。また、b＝41で条件数より求められた誤差を上まわっているが、これは、ここでの誤差が行列のみについて考えていることより、荷重項の影響が現われてきて



図－8　条件数と数値誤差

いるのではないかと思われる。しかし、工学的判断によれば、この誤差は十分に小さいので、ほとんど無視できるであろう。

次に図－7について考えてみる。ここで縦幅bの値が同じものは、同一係数行列であることを示しており、条件数も同じである。ところが実験値をみれば、固定端側からの消去と自由端側からの消去では、解に発生する誤差にかなりの差がみられる。これは丸め誤差の発生がある範囲を有しており、消去の仕方によってそれは変動することを示している。またbの値が大きくなればこの誤差範囲がせばめられているのは、条件数に関係しており、条件数が小さくなることによって、その丸め誤差の拡大率というものが小さくなっていることによると考えられる。このことより次のことがいえる。

[結果－3]

消去順序をかえると、解に発生する数値誤差も変わる。これは、丸め誤差が計算の仕方に依存していることを示している。

[結果－4]

消去順序をかえた時の解に発生する数値誤差の範囲は、条件数 cond(A) に依存しており、解の数値誤差は次式

$$\frac{\|dx\|}{\|x\|} \leq \mathrm{cond}(A) \cdot \frac{\|dA^*\|}{\|A\|}$$

で与えられることより、丸め誤差をおさえるために計算を工夫（例えば、消去順序など）することは、条件数が大きい場合には誤差改良に関して有意であるが、条件数が小さい場合には、その効果は少ない。

また、丸め誤差は、計算機の有効桁数以下の数値処理の仕方にも大いに依存しており、この方法として四捨五入を用いるべきであると先にも述べたが、このことは、丸め誤差の発生をより確率的にしている。

以上の諸結果を要約すると、消去演算過程

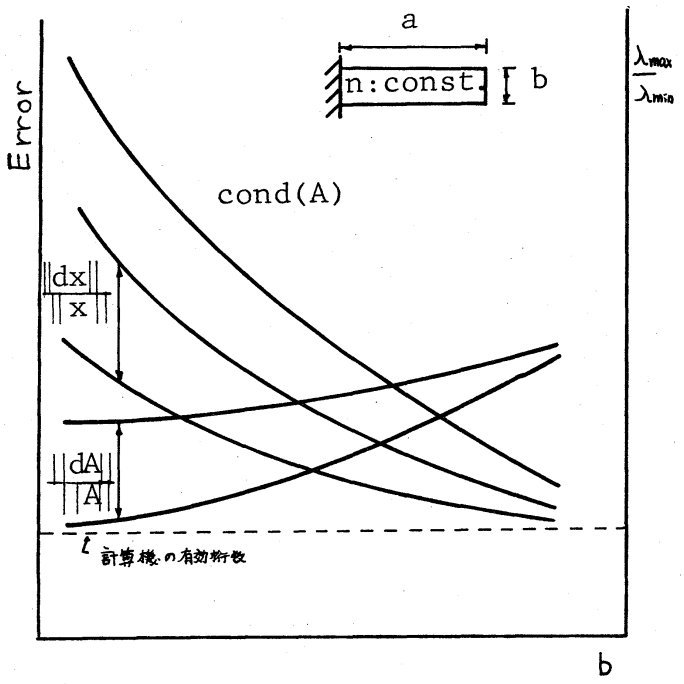


図-9　誤差モデル ①

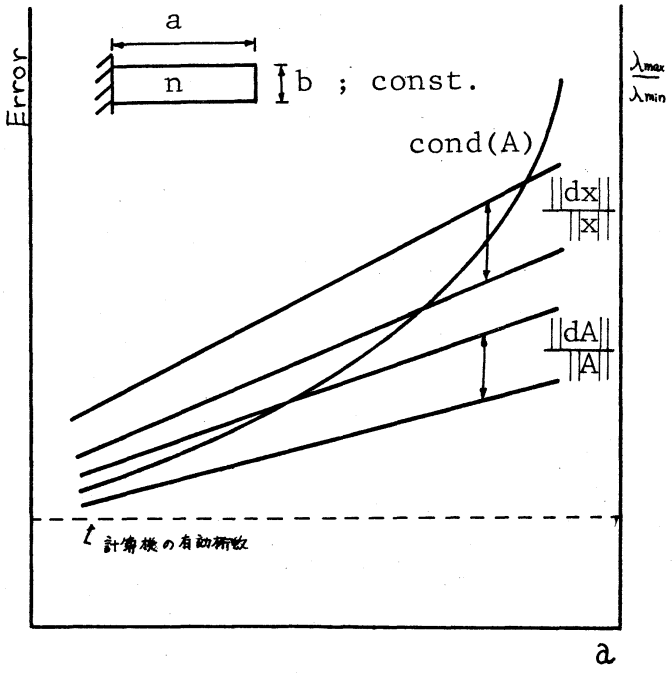


図-10　誤差モデル ②

中に発生する丸め誤差の解に与える影響は、図－9,10に示す2つのモデルで表わすことができる。

図－9は、元数nを一定にしてbを変化させた時の誤差モデルである。これは、丸め誤差による分解誤差はbの増加による演算回数の増加によって次第に大となるが、逆に条件数は減少していくため、結果として解に発生する誤差は減少方向をとることを示している。

図－10は、縦幅bを一定にしてαを変化させた時の誤差モデルである。これは、丸め誤差による分解誤差は、やはり演算回数の増加により次第に大となり、この時条件数もまた増加するため解に発生する誤差は増加することを示している。

また、この時分解誤差は消去順序を考慮することによって減少させることができるが、解の誤差改良の意味では、条件数が大であればこのことは有意であるが、小さい場合にはその効果は少ないといえる。

4. 結論

本研究で得られた結果を要約すれば、次のようになる。

i) 消去演算においては、四捨五入のほうが切り捨てに比べて解の精度は良く、四捨五入を用いるべきである。

ii) 単精度消去演算はある元数N_0以上になると、計算機の機能

面において無意味な計算になりかねない。また、この元数N_{cr}は条件数に依存している。

iii) 丸め誤差をおさえるために計算の仕方を工夫（例えば、消去順序など）することは、与えられた係数行列の条件数が小さい場合には効果が少ないが、大きい場合には有意である。

実問題をBand Matrix法を用いて単精度消去演算を行なうと想定して、最低3桁の精度保障のある解を得ようとするのであれば、最悪の場合で1000元程度まで、またそれ以上の元数については、消去順序を考慮すれば、およそ3000元くらいまでなら適用できると考えられる。

参考文献

(1) J.H Wilkinson 'Rounding Errors in Algebraic Processes', Her Britannic Majesty's Stationary Office (1963)

(2) J.R. Roy, "Numerical Error in Structural Solutions", Journal of Structural Division, ASCE (1971), ST4, pp. 1039 - 1054

(3) G.E. Forsythe and C.B. Moler, 'Computer Solution of Linear Algebraic Systems', Prentice - Hall, Inc (1967)

(4) K.J. Bathe / E.L. Wilson 著, 菊地 文雄 訳 「有限要素法の数値計算」