資　　料

単精度データ入力で倍精度演算した逆行列の精度*

平　野　泰　彦**

Abstract

Accuracy of inverse matrix is influenced largely on selecting single preciseon or double precision. A decision must be made on which precision to take in the computer. However, those selections are related to computer time and storage capacity. Matrix inversion is sometimes computed in single precision data input and dcuble precision arithmetic, so accuracy of such inverse matrices was tested from the viewpoints of accumulated rounding errors and matrix singnlarity.

1. まえがき

逆行列計算における精度は，単精度演算と倍精度演算の選択に大きく左右され，使用する計算機で所望の精度を得るには，どちらを用いるかを決定せねばならない．しかし，これらの選択には演算時間やメモリ容量とも関係する．精度を向上するために，単精度のデータ入力で倍精度演算を行なっているのをみかけるので，その有効性について検討した．逆行列計算における精度の低下には，丸め誤差の累積と行列の特異性の二面が考えられる．前者の特徴をもつ Frank 行列と後者の特徴をもつ Hilbert 行列を用いて数値実験を行なった．

逆行列の一般的計算法は非対称行列にも適用できる消去法であって，三角化の手順をとらないので，Jordan 法に属する．この演算には約 n^3 の乗算回数を要し，演算時間はほぼこれに比例する．したがって，行列の次数が2倍になると，演算時間は約8倍に増加する．逆行列により連立方程式の解を求めることができるが，特にその必要がなければ連立方程式として計算する方が有利で，演算時間は 1/3 になる．

対称行列の逆行列計算には，対称部分のみ計算すれば約 1/2 の演算時間でよく，Cholesky 法が用いられている．なお，固有値の計算を必要とするとき，その計算値を利用して逆行列を計算することができる．数実験において，これらの手法をとりあげた．

2. 丸め誤差の累積からみた消去法の精度

逆行列計算には消去法が広く用いられている．そのとき，精度を向上するために，絶対値最大の要素を探して，それが対角項になるように行または列の置換が行なわれ，ピボッティングという．その方法には，完全ピボッティングと部分ピボッティングがある．全要素について絶対値最大の要素を探す完全ピボッティングでは，大きい行列においてかなりの探索時間を要するので，ある行についてのみ最大要素を探す部分ピボッティングを用いた．

丸め誤差の累積から逆行列の精度を検討するために，入力データとして特異性の小さい Frank 行列をとりあげた．すなわち，

$$A=\begin{pmatrix} n & n-1 & n-2 & \cdots & 1 \\ n-1 & n-1 & n-2 & \cdots & 1 \\ n-2 & n-2 & n-2 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 1 & & 1 \end{pmatrix}.$$

この逆行列では三項対角項以外の要素はすべて零である．

$$A^{-1}=\begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix}.$$

* Accuracy of inverse matrix in single precision data input and double precision arithmetic, by Yasuhiko Hirano (Yokosuka Electric Communication Laboratory, N. T. T.)

** 横須賀電気通信研究所応用プログラム研究室

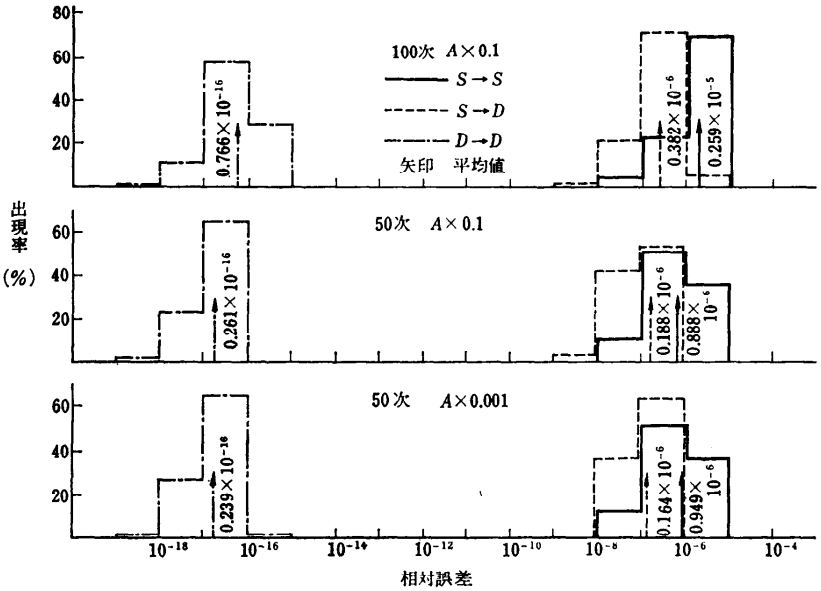

図 1　Frank 行列の消去法による逆行列の精度

Fig. 1　Accuracy of Frank matrix inversion by eleimination method

ここに，行列 A の各要素は整数であるから，計算機内で有限ビットの2進数で表わされ，単精度と倍精度の表現は同一である．小数を2進数に変換すると，0.5，0.25 などの特殊な場合を除いて循環小数になり，計算機内では一般に下位けたが切捨てられて1ワードまたは2ワードに格納される．したがって，入力データとして行列 A の各要素に 0.1 を掛けた値を用い，逆行列として A^{-1} の各要素の 10 倍の値を得ることにした．

まず，行列の大きさを50次にとり，FACOM 230-60 を用いて計算した．零でない各要素の相対誤差の絶対値の平均値としてつぎの値を得た．

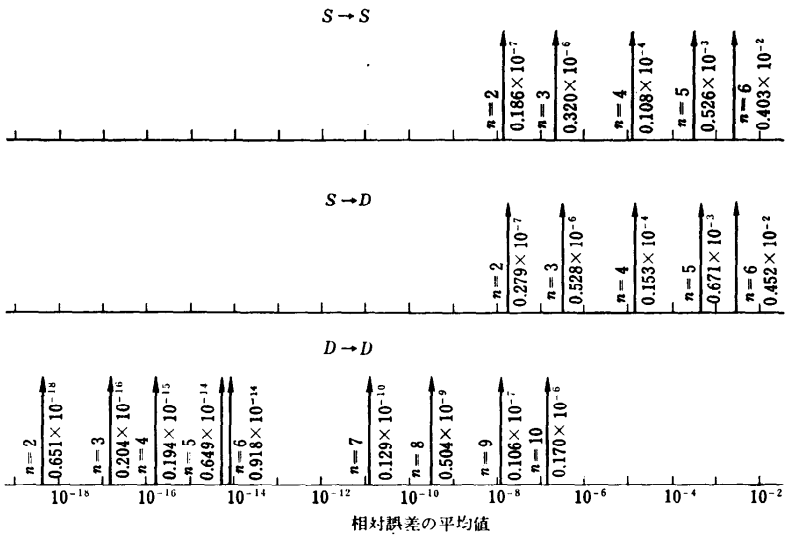

図 2　Hilbert 行列の消去法による逆行列の精度

Fig. 2　Accuracy of Hilbert matrix inversion by elimination method

単精度入力単精度演算 ($S\rightarrow S$) 0.888×10^{-6}

単精度入力倍精度演算 ($S\rightarrow D$) 0.188×10^{-6}

倍精度入力倍精度演算 ($D\rightarrow D$) 0.261×10^{-16}

$D\rightarrow D$ は $S\rightarrow S$ に比し著るしく精度が向上する．しかし，$S\rightarrow D$ では，$S\rightarrow S$ の1けたにも足りないわずかな改善に過ぎない．その理由はつぎのように考えられる．$S\rightarrow D$ の倍精度演算の効果は，$S\rightarrow S$ の個々の演算で生ずる丸め誤差の累積が除去されることにある．念のため，100 次の行列について同様に計算したところ，次数が大きいだけ丸め誤差の累積が大きく，$S\rightarrow D$ の効果は 50 次のときよりも若干増加した．なお，上記の各要素を 0.1 倍する代りに 0.001 倍してみたが，精度はほとんど変らなかった．これらの数値実験の結果を図1に示す．

3. 行列の特異性からみた消去法の精度

与えられた行列の行列式の値が零であるとき，その行列を特異行列または非正則行列といい，逆行列は存在しない．行列式の値が零に近い特異性をもつことは，ベクトル空間において行列の表わすベクトルが完全に重なってはいないが，近似していることを意味する．したがって，このような行列の逆行列計算において，入力データの誤差が拡大されて精度の低下することが考えられる．零に近い値ということを，数値計算上は相対的に表現する必要があり，最小最大固有値の比または NORM 行列[1]などが用いられる．

数値実験のために，特異性の行列として Hilbert 行列[2] を用いた．行列 A の各要素は

$$a_{ij}^{(n)}=\frac{1}{i+j-1}\qquad i,j=1,2,\cdots,n$$

であって，行列 A はつぎのように表わされる．

$$A=\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{n+2} \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1} \end{pmatrix}.$$

この逆行列の各要素はつぎの値で表わされることがわかっている．

$$A^{-1}=\{b_{ij}^{(n)}\},$$

$$b_{ij}^{(n)}=\frac{(-1)^{i+j}(n+i-1)!(n+j-1)!}{(i+j-1)[(i-1)!(j-1)!]^2(n-i)!(n-j)!}.$$

次数が大きくなるにしたがってその特異性は大きくなり，最小最大固有値の比はつぎの値である．

$n=3$ 0.19082×10^{-2}

$n=4$ 0.64459×10^{-4}

$n=5$ 0.20982×10^{-5}

$n=6$ 0.66885×10^{-7}

$n=7$ 0.21036×10^{-8}

$n=8$ 0.65541×10^{-10}

$n=9$ 0.20278×10^{-11}

$n=10$ 0.62397×10^{-13}.

これらの行列の逆行列を $S\rightarrow S$, $S\rightarrow D$ および $D\rightarrow D$ により消去法を用いて計算した．各要素の相対誤差の絶対値の平均は図2に示すようになり，次数の増加とともに精度が低下する．そのとき，$S\rightarrow D$ と $S\rightarrow S$ による精度の差異はほとんどなく，倍精度演算の効果は全くみられない．入力データの2進化による丸め誤差は両者で全く同一で，計算途中で生ずる丸め誤差は非常に小さいからである．しかも，$n=7$ 以上ではともに真値とかけ離れた結果になり，そのときの最小最大固有値の比は1ワードに格納しうる仮数部の範囲を越えている．本来，ピボットの制限値を 10^{-6} 程度にとって計算不能とすべきものである．なお，$D\rightarrow D$ により計算した逆行列の精度のよいことはいうまでもない．

3. Cholesky 法

対称行列の逆行列計算には，Cholesky 法が有利である．Cholesky 平方根法と呼ばれて，$A=LL^T$ に三角分解していたが，$A=LDL^T$ に分解することにより，対角要素の n 個の平方根の計算を避けることができる．統計問題における行列は一般に対称行列であって，しかも対角要素は非対角要素より大きいので，Cholesky 法を適用するのに都合よい．

まず，消去法の場合と同様に，丸め誤差の累積の見地から，50 次の Frank 行列について数値実験を行なった．これを消去法と比較して図3に示す．

Cholesky 法は積和の計算であって，ピボッティングを行なわないので，消去法に比し $S\rightarrow S$ の精度が多少劣っているが，$S\rightarrow D$ の改善もそれだけ大きくなっている．

つぎに，行列の特異性の見地から，Hilbert 行列を用いて前記の数値実験を行なった．各次数について全要素の相対誤差の平均値を図4に示す．この場合は $S\rightarrow S$ の演算中の丸め誤差が大きいため，$S\rightarrow D$ により

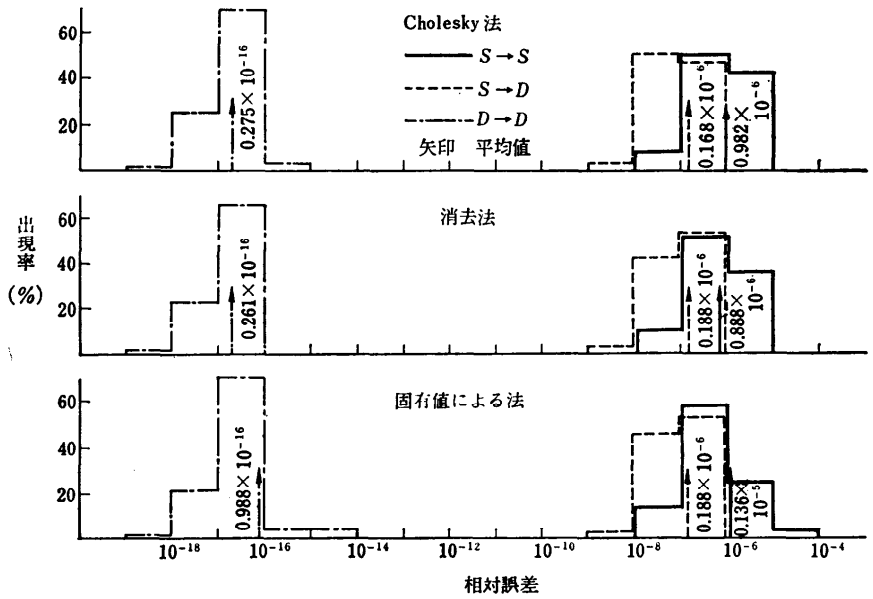

図 3 Frank 行列の各手法による逆行列の精度

Fig. 3 Accuracy of Frank matrix inversion by each methods

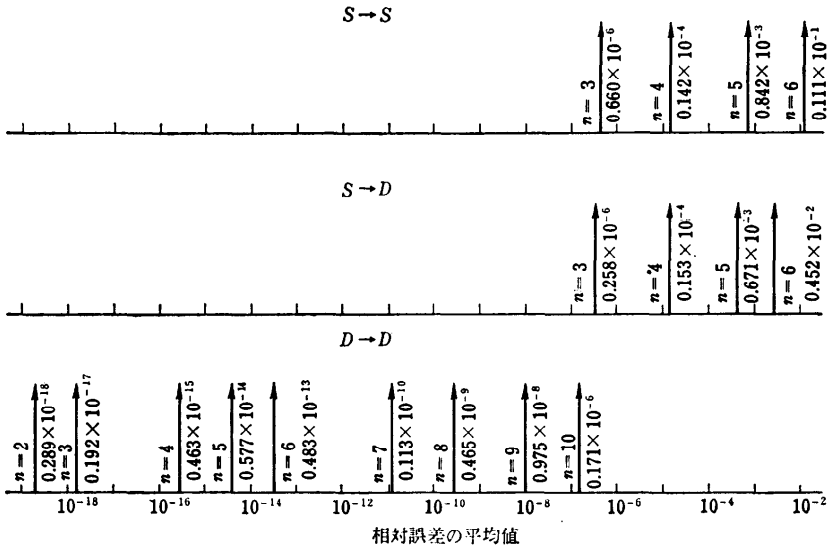

図 4 Hilbert 行列の Cholesky 法による逆行列の精度

Fig. 4 Accuracy of Hilbert matrix inversion by Cholesky method

精度が多少改善され，特に $n=6$ で大きい．

Cholesky 法を消去法と比較すると，精度は劣っているが，演算時間は 1/2 以下である．その上，メモリ容量も約 1/2 でよい．しかし，三角形の一次元配列で演算処理を行なうので，ほかの部分とのインタフェースを考慮させねばならない．

4. 固有値による逆行列計算

統計問題などにおいて，固有値と固有ベクトルの計算を必要とするときは，その計算値を用いて逆行列を

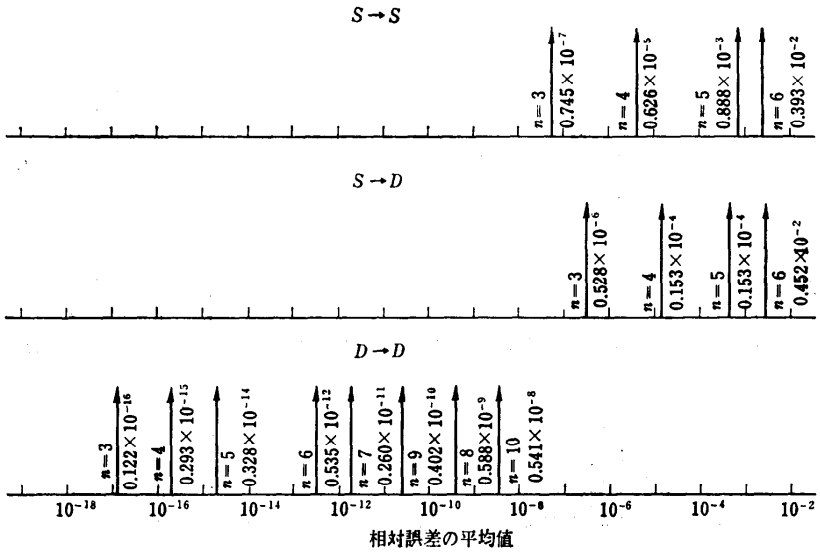

図 5 Hilbert 行列の固有値法による逆行列の精度

Fig. 5 Accuracy of Hilbert matrix inversion by eigen-value method

有利に計算することができる．対称行列の固有値計算法として Jacobi 法がよく知られているが，その適用範囲は 20 次以下である．大きい行列に対しては，Householder 法が有利であって，しかも固有ベクトルを容易に計算することができる．

いま，行列 A の固有値からなる対角行列を Λ，固有ベクトルをならべた行列を U とすると，

$$A=U\Lambda U^{-1}$$

で表される．したがって，

$$A^{-1}=(U\Lambda U^{-1})^{-1}$$
$$=U\Lambda^{-1}U^{-1}.$$

Λ^{-1} は各固有値の逆数を対角要素とする行列である．また，固有ベクトルを正規化すれば $U^{-1}=U^T$ であって，逆行列を計算する必要はない．したがって，行列の乗算により逆行列を計算することができる．

これまでと同様に，Frank 行列と Hilbert 行列を用いて，精度の実験を行なった．それらの結果を図3の下図と図5に示す．この方法の精度は多少わるいが，計算した固有値，特に最小固有値の精度がわるく，これが影響している．これを改善するには，逆行列 Λ^{-1} の代りに，二分法を適用する三項対角行列の逆行列を計算することも考えられる．しかし，精度のわるい固有値に対する固有ベクトルの精度は低下しているので，大きく改善することはできないだろう．$S\rightarrow D$ のときの精度は，図5に示すように $n=3,4$ では $S\rightarrow S$ に比し劣っており，最小固有値の精度が大きく影響している．その理由はつぎのように説明せられる．行列の次数が小さいので，三項対角化における丸め誤差の累積は小さく，$S\rightarrow D$ の最小固有値の精度が $S\rightarrow S$ よりわるく現れたためである．

以上の数値実験に FACOM 230-60 を用いたが，この計算機は丸め誤差に一般と異なる処理をしている．1ワードを越えるビットの丸め誤差は一般に切捨てであって，たとえば 0.01 を 100 回加算すると，丁度 1,0 にならないで 0.999… になることはよく知られている．すなわち，1.0 を越えることは絶対にない．ところが，FACOM 230-60 では，1ワード以下のビットを0捨1入で丸めるという方法をとっている．このように丸めると，加算の結果は1.0以上か，以下か，あるいは丁度1.0になるのか予想できない．したがって，これらを比較するときに注意せねばならない．

一例を示すと，一般的な切捨ての計算機では

```
IF(1.0−X.LT.H) GO TO 10
```

でよい．しかし，丸めた計算機では

```
IF(1.0−X.LT.H∗0.5) GO TO 10
```

のようにせねばならない．しかも，これは単精度演算のときのみであって，倍精度演算では一般と同様の切捨てである．丸めることにより丸め誤差が多少小さくなるけれども，そのための処理を必要とし，しかも上

記の欠点をもっていて，一長一短の論議の余地が残されている．

5. むすび

逆行列計算の精度を調べるために，丸め誤差の累積については大きい Frank 行列を用い，行列の特異性については小さい行列をとりあげて数値実験を行なった．しかし，実際問題としては両性格のもっと混在した行列になるだろう．また，ほかの計算機を使用すると，たとえば，1ワード 32 ビットのバイトマシンでは丸め誤差が大きく，多少相違した結果になるかも知れない．

以上の結論として，統計計算などの実際問題において，逆行列計算を単精度のデータ入力で倍精度演算することは，プログラミングが煩雑であるわりに，その効果はあまり期待できない．したがって，単精度演算で所望の精度が得られないときは，データ入力からファイル記録，演算処理などすべて倍精度にすることが安全である．しかも，大形計算機では両者の演算速度の差異はほとんどないものが多い．しかしながら，メモリ容量の増加については考慮せねばならない．なお，逆行列の場合から一般的に類推して，単精度のデータ入力で倍精度演算することは，丸め誤差の累積に対してのみ有効であって，その利用には十分注意せねばならない．

終りに，固有値計算については富士通提供の HOUS 2S, HOUS 2D を用い，精度の検討について同社のご協力を得たことに感謝いたします．

参考文献

1) 大原正志，石原和夫：条件の悪い連立1次方程式の誤差解析における数値実験，情報処理，Vol. 14, No. 2, pp. 135～142 (1973).
2) R. T. Gregory and D. L. Karney: A Collection of Matrices for Testing Computational Algorithms, Wiley-Interscience.

(昭和48年7月28日受　付)
(昭和48年10月22日再受付)