第8章　ノルム，条件数，連立一次方程式の誤差解析

ノルムは数値線型代数学に不可欠なツールである。$m \times n$ 行列の mn 個の要素をたった一つの数値にまとめてしまうという能力があってこそ，摂動理論や丸め誤差解析を簡単明瞭な形で表現できるのである。反面，ひどいスケーリングになってしまったり，疎行列のような行列構造を生かしきれないといった問題もあるので，ベクトル・行列の要素単位で数値を見た方がいいケースも多い。それでも誤差解析を行う者にとって，ノルムは価値のある道具であることは間違いないのである。

N.J.Nigham, “Accuracy and Stability of Numerical Algorithms 2nd ed.”, (SIAM)

しかし，ベクトル $\mathbf{x}$ のノルム $\|\mathbf{x}\|$ は，たとえば $= \sqrt{x^2 + y^2}$ とか $= |x| + |y|$ とか $= \max(|x|, |y|)$ なので，たとえ何らかの標準化をしたところで，x が電圧で y が電流であるというような場合，どうして $\|\mathbf{x}\|$ が物理的に意味のある量であると考えられようか。そして，物理的に無意味なものを扱う計算過程を許すような数値計算論が健全であるといえようか。

このような，数値計算の常識以前の常識のことが気にかかるのであるが，著者自身も歯切れのよい解答は持ち合わせていないのが残念である。

伊里正夫・藤野和建「数値計算の常識」(共立出版)

今日の科学技術計算においては大規模な線型計算が多用されるため，理工系初年度で習う線型代数学ではあまり扱わない各種のノルム (norm) という概念が必須のものとなる。ここでは有限次元のベクトルや行列に対するノルムを定義し，その応用として，連立一次方程式の誤差解析とそこで使用する条件数 (condition number) を定義する。

8.1　ベクトルと行列のノルム，条件数

ノルムとは，簡単に言うと $\mathbb{R}$ や $\mathbb{C}$ における絶対値 $|\cdot|$ の拡張概念ということになる。絶対値は $a, b \in \mathbb{C}$ という二数の「距離」を表現するために用いられるものである。つまり，$|a - b|$ が 0 に近ければ，a, b は「近い」と言える。$|a| = |a - 0|$ であるから，これが 0 に近ければ 0 との距離が近いということになる。よって，ノルムも n 次元空間 $\mathbb{C}^n$ もしくは $\mathbb{R}^n$ におけるベクトル間の距離を表

わすものと考えてよい。逆に言えば，距離を表現するに足る性質さえ満たしていれば，それは全てノルムであるということになる。

定義 8.1.1 (ノルムの性質)
$\forall a \in \mathbb{C}^n$(または $\mathbb{R}^n$ 以下同様) に対して定義される $\mathbb{R}$ への写像 $\|\mathbf{a}\| \in \mathbb{R}$ が次の三つの性質を全て満足する時，$\|\cdot\|$ を $\mathbb{C}^n$ におけるノルムと呼ぶ。

1. $\forall \mathbf{a} \in \mathbb{C}^n$(or $\mathbb{R}^n$ 以下同様) に対して，
$$\|\mathbf{a}\| \geq 0$$
となる。

2. $\forall \alpha \in \mathbb{C}$(または $\mathbb{R}$ 以下同様)，$\forall \mathbf{a} \in \mathbb{C}^n$ に対し，
$$\|\alpha \mathbf{a}\| = |\alpha| \cdot \|\mathbf{a}\|$$
を満足する。

3. $\forall \mathbf{a}, \mathbf{b} \in \mathbb{C}^n$ に対し
$$\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\| \quad \text{(三角不等式)}$$
を満足する。

後で行列のノルムも定義するが，上の条件を全て満足する所は全く同じである。
一般的には次の p ノルムがある。

定義 8.1.2 (p ノルム)
$\mathbf{a} = [a_1 \ \cdots \ a_n]^T \in \mathbb{C}^n$ に対し，
$$\|\mathbf{a}\|_p = \left(\sum_{i=1}^{n} |a_i|^p \right)^{1/p}$$
をベクトルの p ノルムと呼ぶ。

よく使用されるのは $p = 1, 2, \infty$ の場合である。これをそれぞれ 1 ノルム, 2 ノルム (またはユークリッドノルム), 無限大ノルムと呼ぶ。

$$1 \text{ ノルム} \quad \|\mathbf{a}\|_1 = \sum_{i=1}^{n} |a_i|$$
$$\text{ユークリッドノルム} \quad \|\mathbf{a}\|_2 = \sqrt{\sum_{i=1}^{n} |a_i|^2} \quad = \sqrt{(\mathbf{a}, \mathbf{a})}$$
$$\text{無限大ノルム} \quad \|\mathbf{a}\|_\infty = \max_i |a_i|$$

$\mathbf{x} = [x_1 \ x_2]^T \in \mathbb{R}^2$ の時，それぞれのノルムが 1 になる位置を図示したのが図 8.1 である。もし円の定義を,「原点からの距離 (= ノルム) が一定の点の集まり」とするのであれば，これらは全て円ということになる。

数値計算ではベクトル列 $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k, ...$ を形成するアルゴリズムが多いが，有限次元の線型空間 $\mathbb{C}^n$ や $\mathbb{R}^n$ においては，あるノルムを用いて収束，即ち $\lim_{k\to\infty} \|\mathbf{x}_k\|_p = \|\mathbf{a}\|_p$ となるものが，別のノル

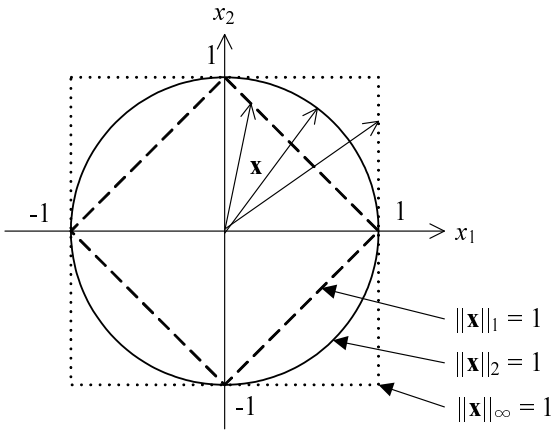

図 8.1: $\|\mathbf{x}\|_1 = \|\mathbf{x}\|_2 = \|\mathbf{x}\|_\infty = 1$ となるベクトルの位置

ムでは発散，即ち $\lim_{k\to\infty} \|\mathbf{x}_k\|_q = \infty$ となってしまう，という事態は起こらない。もちろんその逆も起こらない。実際，任意のベクトル $\|\cdot\|_p$ と $\|\cdot\|_q$ との間には，$\forall \mathbf{x} \in \mathbb{C}^n$ に対して

$$\|\mathbf{x}\|_p \leq \alpha_{pq} \|\mathbf{x}\|_q$$

という定数 $\alpha_{pq} \geq 0$ が存在する。よく用いられる 3 つのノルムにおける α_{pq} は表 8.1 のようになる。

表 8.1: 1 ノルム，ユークリッドノルム，無限大ノルム間における α_{pq}

$p \rightarrow$ $q \downarrow$	1	2	∞
1	1	$\sqrt{n}$	n
2	1	1	$\sqrt{n}$
∞	1	1	1

行列のノルムはベクトルのノルムをベースにして定義されるものが普通である。代表的なものとしては次の p ノルムがある。

定義 8.1.3 (行列の p ノルム)
$\forall A \in M_n(\mathbb{C})$(または $M_n(\mathbb{R})$ 以下同様) に対し，

$$\|A\|_p = \sup_{\mathbf{x}\neq 0} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$$

と定義される $\|\cdot\|_p$ を行列の p ノルムと呼ぶ。

これによって，行列の 1 ノルム，ユークリッドノルム，無限大ノルムも同様に定義されることになる。

$$\text{1 ノルム}\ \|A\|_1 = \sup_{\mathbf{x}\neq 0}\frac{\|A\mathbf{x}\|_1}{\|\mathbf{x}\|_1} = \max_j \sum_{i=1}^{n}|a_{ij}|$$

$$\text{ユークリッドノルム}\ \|A\|_2 = \sup_{\mathbf{x}\neq 0}\frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sqrt{\max_i \lambda_i(A^*A)}\ \text{(ここで } \lambda_i(A) \text{ は } A \text{ の固有値)}$$

$$\text{無限大ノルム}\ \|A\|_\infty = \sup_{\mathbf{x}\neq 0}\frac{\|A\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \max_i \sum_{j=1}^{n}|a_{ij}|$$

なお，行列の p ノルムとベクトルの p ノルムとの間には，$\forall A \in \mathrm{M}_n(\mathbb{C})$，$\forall \mathbf{x} \in \mathbb{C}^n$ に対して

$$\|A\mathbf{x}\|_p \leq \|A\|_p \cdot \|\mathbf{x}\|_p$$

という関係がある。よって，$\forall A, B \in M_n(\mathbb{C})$ に対しても

$$\|AB\|_p \leq \|A\|_p \cdot \|B\|_p$$

が成り立つ。

さて，これでベクトルと行列に絶対値の拡張概念であるノルムを導入できた訳だが，そうすると，ベクトルや行列にも絶対誤差と相対誤差が定義できることになる。今まではなるべく一般的な定義を $\mathbb{C}^n$ や $M_n(\mathbb{C})$ に対して述べてきたが，本書では今のところ複素数を要素とする行列やベクトルを用いたアルゴリズムを扱っていないので，誤差の定義は $\mathbb{R}^n$，$M_n(\mathbb{R})$ 止まりとする。

定義 8.1.4 (ベクトルの誤差)
$\mathbf{a} \in \mathbb{R}^n$ を真の値，$\widetilde{\mathbf{a}} \in \mathbb{R}^n$ をその近似値とする。このとき

$$E(\widetilde{\mathbf{a}}) = \|\mathbf{a} - \widetilde{\mathbf{a}}\| \tag{8.1}$$

を $\widetilde{\mathbf{a}}$ の絶対誤差と言う。更に

$$rE(\widetilde{\mathbf{a}}) = \begin{cases} \dfrac{\|\mathbf{a} - \widetilde{\mathbf{a}}\|}{\|\mathbf{a}\|} = \dfrac{E(\widetilde{\mathbf{a}})}{\|\mathbf{a}\|} & (\mathbf{a} \neq 0) \\ \|\mathbf{a} - \widetilde{\mathbf{a}}\| = E(\widetilde{\mathbf{a}}) & (\mathbf{a} = 0) \end{cases} \tag{8.2}$$

を $\widetilde{\mathbf{a}}$ の相対誤差と言う。特にノルムが $\|\cdot\|_p$ であるときには

$$E_p(\widetilde{\mathbf{a}}) = \|\mathbf{a} - \widetilde{\mathbf{a}}\|_p,\ rE_p(\widetilde{\mathbf{a}}) = \frac{E_p(\widetilde{\mathbf{a}})}{\|\mathbf{a}\|_p}$$

と書くことにする。

行列についても同様に，真値 $A \in M_n(\mathbb{R})$ とその近似値 $\widetilde{A} \in M_n(\mathbb{R})$ に対して，$E(\widetilde{A})$, $E_p(\widetilde{A})$, $rE(\widetilde{A})$, $rE_p(\widetilde{A})$ が定義できる。

最後に，連立一次方程式の誤差解析において重要なファクターである，行列の条件数を定義する。

定義 8.1.5 (行列の条件数)
正則行列 $A \in M_n(\mathbb{R})$ の条件数 $\kappa(A)$ を

$$\kappa(A) = \|A\| \cdot \|A^{-1}\|$$

と定義する。p ノルムを用いたものを $\kappa_p(A)$ と書く。

結論から先に言うと，この $\kappa(A)$ が非常に大きい行列 A を悪条件である (ill-conditioned) と呼び，行列成分，定数ベクトル成分に含まれる初期誤差や，解法の過程で発生した丸め誤差を条件数倍して，数値解に忍び込む可能性が高い，といわれている。次節ではその解説を行う。

問題 8.1.1
次の行列 $A \in M_2(\mathbb{R})$ の条件数 $\kappa_1(A)$ 及び $\kappa_\infty(A)$ を求めよ。

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

8.2 連立一次方程式の誤差解析

連立一次方程式 (7.1) を実際にコンピュータで計算しようとすると，一般的には行列，ベクトル成分は丸められて誤差が混入していると考えられる。また，解法の過程でも丸め誤差が混入するのが普通である。よって，これらを全て，行列あるいは定数ベクトルに忍び込んだ初期誤差として解釈することにすれば，解こうとする (7.1) は

$$\widetilde{A}\,\widetilde{\mathbf{x}} = \widetilde{\mathbf{b}}$$

という形に化けているということが出来る。では最終的に得られる $\widetilde{\mathbf{x}}$ に含まれる誤差 $E(\widetilde{\mathbf{x}})$ はどうなっているのだろうか。

それを解析する基になる補題，定理を示すことにする。

補題 8.2.1 (定数項 b に誤差がある場合)
連立一次方程式

$$A(\mathbf{x} + E(\widetilde{\mathbf{x}})) = \mathbf{b} + E(\widetilde{\mathbf{b}})$$

において

$$rE(\widetilde{\mathbf{x}}) \leq \kappa(A) \cdot rE(\widetilde{\mathbf{b}})$$

である。

(証)

$E(\widetilde{\mathbf{x}}) = A^{-1}E(\widetilde{\mathbf{b}})$ から

$$\|E(\widetilde{\mathbf{x}})\| \leq \|A^{-1}\|\,\|E(\widetilde{\mathbf{b}})\|$$

を得る。更に，$\|\mathbf{b}\| \leq \|A\|\,\|\mathbf{x}\|$ から与式を得る。

(証明終)

補題 8.2.2 (係数行列 A に誤差がある場合)

$$(A + E(\widetilde{A}))(\mathbf{x} + E(\widetilde{\mathbf{x}})) = \mathbf{b}$$

において

$$\frac{\|E(\widetilde{\mathbf{x}})\|}{\|\mathbf{x} + E(\widetilde{\mathbf{x}})\|} \leq \kappa(A) \cdot rE(\widetilde{A})$$

である。

(証)

$$\begin{aligned} \mathbf{x} &= A^{-1}\mathbf{b} \\ &= A^{-1}(A + E(\widetilde{A}))(\mathbf{x} + E(\widetilde{\mathbf{x}})) \\ &= \mathbf{x} + E(\widetilde{\mathbf{x}}) + A^{-1}E(\widetilde{A})(\mathbf{x} + E(\widetilde{\mathbf{x}})) \end{aligned}$$

から

$$-E(\widetilde{\mathbf{x}}) = A^{-1}E(\widetilde{A})(\mathbf{x} + E(\widetilde{\mathbf{x}}))。$$

従って，

$$\begin{aligned} \|E(\widetilde{\mathbf{x}})\| &\leq \|A^{-1}\|\,\|E(\widetilde{A})\| \cdot \|\mathbf{x} + E(\widetilde{\mathbf{x}})\| \\ &= \|A^{-1}\|\,\|A^{-1}\| \cdot \frac{\|E(\widetilde{A})\|}{\|A\|} \cdot \|\mathbf{x} + E(\widetilde{\mathbf{x}})\| \end{aligned}$$

より，与式を得る。

(証明終)

定理 8.2.3 (係数行列，定数項共に誤差を含んでいるとき)

$$(A + E(\widetilde{A}))(\mathbf{x} + E(\widetilde{\mathbf{x}})) = \mathbf{b} + E(\widetilde{\mathbf{b}})$$

なるとき $\|A^{-1}E(\widetilde{A})\| < 1$ ならば，

$$rE(\widetilde{\mathbf{x}}) \leq \frac{\kappa(A)}{1 - \|A^{-1}E(\widetilde{A})\|}\left(rE(\widetilde{\mathbf{b}}) + rE(\widetilde{A})\right)$$

である。

(証)

$I + A^{-1}E(\widetilde{A})$ の固有値は $1 + \lambda(A^{-1}E(\widetilde{A}))$ だから，$\lambda(A^{-1}E(\widetilde{A}))$ によらず，正則になる。従って，

$$(I + A^{-1}E(\widetilde{A}))^{-1} = I - A^{-1}E(\widetilde{A})(I + A^{-1}E(\widetilde{A}))^{-1}$$

が成立するから，

$$\|(I + A^{-1}E(\widetilde{A}))^{-1}\| \leq 1 + \|A^{-1}E(\widetilde{A})\|\,\|(I + A^{-1}E(\widetilde{A}))^{-1}\|$$

よって，

$$\|(I + A^{-1}E(\widetilde{A}))^{-1}\|\,(1 - \|A^{-1}E(\widetilde{A})\|) \leq 1$$

から

$$\|(I + A^{-1}E(\widetilde{A}))^{-1}\| \leq \frac{1}{1 - \|A^{-1}E(\widetilde{A})\|}$$

を得る。

ここで $(A + E(\widetilde{A}))(\mathbf{x} + E(\widetilde{\mathbf{x}})) = \mathbf{b} + E(\widetilde{\mathbf{b}})$ と $A\mathbf{x} = \mathbf{b}$ より

$$E(\widetilde{A})\mathbf{x} + (A + E(\widetilde{A}))E(\widetilde{\mathbf{x}}) = E(\widetilde{\mathbf{b}})。$$

これに，左から A^{-1} をかけると

$$A^{-1}E(\widetilde{A})\mathbf{x} + (I + A^{-1}E(\widetilde{A}))E(\widetilde{\mathbf{x}}) = A^{-1}E(\widetilde{\mathbf{b}})$$

これを $E(\widetilde{\mathbf{x}})$ について解くと，

$$E(\widetilde{\mathbf{x}}) = (I + A^{-1}E(\widetilde{A}))^{-1}A^{-1}(E(\widetilde{A})\mathbf{x} - E(\widetilde{\mathbf{b}}))。$$

よって，

$$\begin{aligned}\frac{\|E(\widetilde{\mathbf{x}})\|}{\|\mathbf{x}\|} &\leq \|(I + A^{-1}E(\widetilde{A}))^{-1}\|\|A^{-1}\|\left(\|E(\widetilde{A})\| + \frac{\|E(\widetilde{\mathbf{b}})\|}{\|\mathbf{x}\|}\right)\\ &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}E(\widetilde{A})\|}\left(\|E(\widetilde{A})\| + \frac{\|E(\widetilde{\mathbf{b}})\|}{\|\mathbf{b}\|}\right)。\end{aligned}$$

ここで $\|\mathbf{x}\| \leq \|A\|\|\mathbf{b}\|$ より与式を得る。

(証明終)

この不等式は $\|A^{-1}\| \geq \frac{1}{\|A\|}$ を用いて，

$$rE(\widetilde{\mathbf{x}}) \leq \frac{\kappa(A)}{1 - \dfrac{\|E(\widetilde{A})\|}{\|A\|}} \cdot \left(rE(\widetilde{A}) + rE(\widetilde{\mathbf{b}})\right) \tag{8.3}$$

という形にしたものが主に用いられる。

今まで見てきた不等式がいずれも条件数をファクターとして持つ不等式になっていることが分かった。以上をまとめると

1. 定数項のみに誤差がある場合 —

$$rE(\widetilde{\mathbf{x}}) \leq \kappa(A) \cdot rE(\widetilde{\mathbf{b}})$$

2. 係数行列のみに誤差がある場合 —

$$\frac{\|E(\widetilde{\mathbf{x}})\|}{\|\mathbf{x} + E(\widetilde{\mathbf{x}})\|} \leq \kappa(A) \cdot rE(\widetilde{A})$$

3. 双方に誤差がある場合 —

$$rE(\widetilde{\mathbf{x}}) \leq \frac{\kappa(A)}{1 - \|A^{-1}E(\widetilde{A})\|}\left(rE(\widetilde{\mathbf{b}}) + rE(\widetilde{A})\right)$$

となる。

このうち，2. では $\|E(\widetilde{\mathbf{x}})\|$ が $\|\mathbf{x}\|$ より小さいことが前提となる。$\|E(\widetilde{\mathbf{x}})\| > \|\mathbf{x}\|$ では

$$\frac{\|E(\widetilde{\mathbf{x}})\|}{\|\mathbf{x} + E(\widetilde{\mathbf{x}})\|} \approx 1$$

となり，評価式の意味がなくなってしまう。

3. での $\|E(\widetilde{A})\|/\|A\| < 1$ という条件は，たちの悪い問題では満たされないこともあり，注意する必要がある。

上の不等式はいずれも，行列・ベクトルノルムは同じものであるが，有限次元のノルムの同値性から，都合の良いものを組み合わせて使うこともできる。

これらの不等式を利用して，初期誤差の摂動の限界を調べる。よって計算に必要な桁数は

$$\text{計算に必要な桁数} \geq \log_{10}(\kappa(A)) + \max(\widetilde{A}\text{の精度桁}, \widetilde{\mathbf{b}}\text{の精度桁}) \tag{8.4}$$

となる。

ここで言う，$\widetilde{A}$ の精度桁，$\widetilde{\mathbf{b}}$ の精度桁とは，それぞれの成分の中で最も有効桁の少ない成分のものを意味する。

この式は大雑把なもので，完璧を期すならば，丸め誤差の累積の効果を見込んで $+\log_{10} n$ を加えると良い。ただ，次数が大きくなったとき，この 3 倍になる可能性がある事を Wilkinson が指摘していることを述べておく。

8.3 Hilbert 行列の数値例

行列に誤差が混入した場合の，その固有値の摂動については，Wilkinson[42] が理論的な考察を数多く行っている。しかし実際，どのような場合に，どの程度の影響が与えられるのかを，彼の得た結果から予測することは難しい。その原因には，具体例が乏しいことも挙げられる。

本稿では，それを実対称行列の場合に限定することで，行列の要素に誤差が混入することにより起こる固有値の変動を一種の摂動と考え，それに伴う連立一次方程式への影響について，具体例を基に説明する。

なお，以下の計算は SparcStation IPX と Sun Fortran(単精度 2 進 24 ビット) で行なった。

8.3.1 固有値の摂動

Courant-Fischer(Min-Max) の定理から，次の Weyl の定理が容易に導きだされる。

定理 8.3.1 (Weyl の定理)

$A,\ B : n \times n$ Hermite 行列，　$\lambda_i(A) : A$ の i 番目の固有値 $(\lambda_1(A) \leq \lambda_2(A) \leq \cdots \leq \lambda_n(A))$

とするとき，

$$\lambda_i(A) + \lambda_1(B) \leq \lambda_i(A+B) \leq \lambda_i(A) + \lambda_n(B)\ (i = 1, 2, \cdots, n)。 \tag{8.5}$$

Wilkinson はこれとほぼ同じことを [42](p.101〜p.102) で証明している。

この B が行列 A の入力誤差行列で，しかも対称であるとすると，もし

$$|b_{ij}| \leq \varepsilon\ (i, j = 1, 2, \cdots, n) \tag{8.6}$$

であれば

$$|\lambda_i(B)| \leq n\varepsilon\ (i = 1, 2, \cdots, n) \tag{8.7}$$

であることは Rayleigh-Ritz の定理から明らかである。よって

$$|\lambda_i(A+B) - \lambda_i(A)| \leq n\varepsilon \ (i = 1, 2, \cdots, n) \tag{8.8}$$

を得る。即ち，固有値の絶対誤差は入力誤差の最大値のせいぜい n 倍で抑えることができる。

現在広く利用されている対称行列用の固有値ルーチンは，もととなる行列の上三角（もしくは下三角）成分のみ用いて計算する。従って，初期誤差行列の対称性は理論的にも厳密に保たれていると考えてよい。

以下，この定理を具体例で確認する。

例題 8.3.2
悪条件で名高い，Hilbert 行列

$$\begin{aligned} A &= \begin{bmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n+1} \\ \vdots & \vdots & & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \cdots & \frac{1}{2n-1} \end{bmatrix} \\ &= [a_{ij}] \end{aligned}$$

の固有値を求める。その際，行列の成分 $a_{ij} = \frac{1}{i+j-1}$ を

1. 単精度 — 2 進仮数部 24 ビット
2. 倍精度 — 2 進仮数部 53 ビット

で計算し，それらを 4 倍精度の固有値ルーチンで解いたのが，下表の第 1 列，第 2 列である。第 3 列は 1 列，2 列の差であり，単精度行列には入力誤差が入るので，この誤差による固有値の摂動量を表わしている。

Table 1 Hilbert 行列の固有値 (Dimension:10)

	単精度	倍精度	単 – 倍
1	0.1751920E + 01	0.17519196702651775E + 01	0.945007773E – 07
2	0.3429295E + 00	0.34292954848350908E + 00	0.364589606E – 07
3	0.3574183E – 01	0.35741816271639246E – 01	0.183592214E – 07
4	0.2530897E – 02	0.25308907686700395E – 02	0.645588386E – 08
5	0.1287263E – 03	0.12874961427636959E – 03	0.233577239E – 07
6	0.4739574E – 05	0.47296892931844192E – 05	0.988459762E – 08
7	0.1260786E – 06	0.12289677387895489E – 06	0.318180201E – 08
8	–0.2991611E – 07	0.21474388118628759E – 08	0.320635455E – 07
9	0.3893632E – 07	0.22667468977874495E – 10	0.389136475E – 07
10	0.2016819E – 07	0.10930958726131586E – 12	0.201680775E – 07

Hilbert 行列では，成分の計算で丸め誤差が発生する。これを単精度計算で行なったから，(8.6) における ε の値はこの場合

$$|\varepsilon| \leq 0.39736 \times 10^{-7} \tag{8.9}$$

で抑えられる。従って，$n|\varepsilon| \leq 0.39736 \times 10^{-6}$ となり，(8.8) の評価式が成立していることがわかる。

次に，単精度行列の固有値を単精度，倍精度，4 倍精度の QR 分解法 (第 11 章参照) で求めてみると，

Table 2 入力誤差を伴った Hilbert 行列の固有値 (Dimension:10)

	単精度	倍精度	4 倍精度
1	.1751903E + 1	.17519195757643918E + 1	.17519195757644001152650466989724 05E + 1
2	.3429263E + 0	.34292951202454800E + 0	.34292951202454838313324344914622 69E + 0
3	.3574016E − 1	.35741834630860338E − 1	.35741834630860684139288790393098 98E − 1
4	.2530411E − 2	.25308972245537668E − 2	.25308972245539076395248624783899 67E − 2
5	.1286942E − 3	.12872625655238820E − 3	.12872625655240671113946815495498 03E − 3
6	.4697030E − 5	.47395738908145271E − 5	.47395738908125403187192270822335 44E − 5
7	.1872392E − 6	.12607857589827829E − 6	.12607857589213146943228308504062 64E − 6
8	−.4769931E − 7	−.29916106699545878E − 7	−.29916106701330251733819361448393 10E − 7
9	.1464610E − 7	.38936315021525424E − 7	.38936315025379488897249169001616 44E − 7
10	.3607411E − 7	.20168186899426792E − 7	.20168186894434676728414764239058 32E − 7

となり，固有値は殆ど変化していない。

これは最小固有値が成分計算で発生した丸め誤差により摂動を起こし，本来もっと小さくなるはずの order が停留しているためである。

8.3.2 連立一次方程式への影響

前節で見たように，行列の固有値は悪条件性が高ければ高いほど摂動の量が大きい。しかも Hilbert 行列の場合，次数を上げても最小固有値の order はせいぜい入力誤差のそれぐらいにしかならない。従って，条件数もあまり変化しなくなる。試みに，各 Dimension の，単精度・倍精度・4 倍精度の Hilbert 行列の条件数を求めてみると，

Table 3 Hilbert 行列の条件数

Dimension	κ_2(単精度)	κ_2(倍精度)	κ_2(4 倍精度)
5	0.4737947E+06	0.4766073E+06	0.4766073E+06
10	0.8686551E+08	0.1602714E+14	0.1602629E+14
15	0.3260381E+10	0.1246449E+20	0.6116566E+21
20	0.8902582E+10	0.4053651E+19	0.2452156E+29

となり，成分計算の精度が少ないと，行列の Dimension を上げたとき，真の最小固有値が ε 以下になっても，条件数はほとんど変化しない。このようなとき，連立一次方程式の数値解にはどのような影響があるのだろうか。次の例でそれを示す。

Table 4 連立一次方程式の数値解の精度桁

定数項	単精度	倍精度	4 倍精度
1	1.796	11.766	26.732
2	0.391	10.436	25.603
3	−0.450	9.728	25.105
4	−0.889	9.733	25.279
5	−0.856	9.106	25.311
6	−0.619	8.687	26.768
7	−1.010	8.860	25.305
8	0.109	9.331	25.485
9	−0.789	9.083	27.335
10	−0.441	9.654	25.467

右の表は，次のような条件下で求めた連立一次方程式の数値解の $-\log_{10}$(相対誤差) である。

係数行列 全て単精度の Hilbert 行列。

数値解法 4 倍精度の共役勾配法 (第 10 章参照)。

真の解 $x_i = \sqrt{i}\ (i = 1, 2, \cdots, 10)$。

定数項 $b_i = \sum_{j=1}^{10} a_{ij}\sqrt{j}$ を単精度，倍精度，4 倍精度で計算した。

上の表から，それぞれの計算桁からの損失桁は，ほぼ条件数 (10^{+8}) 以下で抑えられていることが示された。

以上の数値例から次の結論を得る。

1. 固有値を求めるときに，行列が良条件であれば，計算精度を初期誤差の order よりも少し多めにとることで，丸め誤差の影響を避けることができる。
2. 行列が悪条件の時は，初期誤差による摂動が大きく，計算の精度を上げても，特に最小固有値が化けてしまい，真の固有値を求めることが不可能になることがある。
3. 逆に，連立一次方程式を解く際には，条件数が大きくならないことから，計算による丸め誤差の影響は少なくなる。但し，初期誤差による摂動のため，精度の良い解は望めない。

演習問題

1. 次の正則行列 A について，以下の問いに答えよ。

$$A = \begin{bmatrix} 3 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 3 \end{bmatrix}$$

   (a) A の逆行列は

$$A^{-1} = \frac{1}{21}\begin{bmatrix} \boxed{(1)} & 3 & 1 \\ 3 & \boxed{(2)} & 3 \\ 1 & 3 & \boxed{(3)} \end{bmatrix}$$

   となる。空欄を埋め，逆行列を完成させよ。

   (b) $\kappa_1(A), \kappa_\infty(A)$ をそれぞれ求めよ。

2. $\forall a, b \in \mathbb{C}$ に対して，絶対値 $|\cdot|$ はノルムの性質を全て満足することを示せ。

3. 1 ノルム, ユークリッドノルム, 無限大ノルムがノルムの性質を満足していることを示せ。また，それぞれのノルムの値を計算するのに必要な演算 (加減算，乗算，除算，平方根等) の回数を求めよ。

4. 正則行列 $A \in M_n(\mathbb{R})$ に対して，$\|A^{-1}\| \geq 1/\|A\|$ であることを示せ。

5. 特に A が実対称行列であれば
$$\kappa_2(A) = \frac{\max_i |\lambda_i(A)|}{\min_i |\lambda_i(A)|}$$
であることを示せ。ここで $\lambda_i(A)$ $(i = 1, 2, ..., n)$ は A の固有値である。

6. $A \in M_n(\mathbb{R})$ における Frobenius ノルム $\|A\|_F$ は
$$\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2} \tag{8.10}$$
と定義される。これがノルムの条件を満足することを示せ。