数値計算の常識

目　次

Id の欄は URL

http://km.int.oyo.co.jp/ShowDocumentDetailsPage.jsp?DocumentId=＊＊**

の ＊＊** の部分を示す.

数値計算の常識

近年，さまざまな解析ソフトが出回っているので，新たにプログラムを開発する機会は少なくなっているかもしれない．しかし，先端的な研究開発を行うためには，どうしても自分自身でソフトウェアーを作る必要がある．

計算機の目覚しい発達により，計算そのものは容易に行えるようになってきた．しかし，数式どおりにむやみに計算を行っても，正しい結果が得られるとは限らない．正しい結果を得るには問題の性質に応じた計算方法をとらなければならない．既存のアプリケーションを上手に利用するためにも，どのような計算法を取っているかの評価ができなければならない．そこで　この講義では数値「解析」よりも数値「計算」の方法の習得に重点を置く．名づけて「数値計算の常識」とする．

この講義ノートの原型は東工大で行っていた講義「数値解析」用のメモである．十数年以上前のものであるから，時代遅れのところもあるかもしれない．たとえばもとのメモでは例題プログラムを BASIC で書いたものもあった．当時は PC に BASIC が入っているのは当然であったし，電卓に BASIC が組み込まれていた時代でもあった．さすがにこの稿では BASIC は Fortran に書きかえてある．また当時はメインフレーム全盛で，WS や PC で数値計算をすることなど考えもしなかった．そのような時代背景の差はあっても，数値計算の技術自体には普遍的なものがある．この講義で目指しているのはそのようなものである．

1　桁落ちと情報落ち

数値計算において**桁落ち**の危険性については昔からしばしば強調されてきているが，桁落ちとは表裏の関係にある**情報落ち**については触れられることが少ない．代数的には等価な式でも，数値計算では異なる結果を与えることがある．以下にそのような例をいくつか示す．

例 1　$\sqrt{1+x}-1$

$x = 1.23456 \times 10^{-3}$ に対してを有効数字 6 桁で計算する．

$$1+x = 1.00123 \tag{1.1}$$
$$\sqrt{1+x} = 1.00061 \tag{1.2}$$
$$\sqrt{1+x}-1 = 0.00061 \tag{1.3}$$

この計算では (1.3) 式の減算で上位 4 桁が桁落ちした結果，答の有効数字は 2 桁しかない．これまではこの桁落ちの危険性だけが強調されてきた．

しかしよく考えてみると，この桁落ちの伏線は実は (1.1) 式の加算にある．この計算は正確には $1+x = 1.00123456$ となるべきところが，有効数字が 6 桁という制約のために下位 3 桁が切り捨てられている．上位の桁が失われる桁落ちに対して，このように下位の桁が失われることを情報落ちと呼ぶ．桁落ちでは有効数字が失われるが，情報落ちの場合には有効数字が失われることはない．しかしその結果の値を (1.3) 式のような計算に用いると失われた下位の桁が意味をもってくるのである．

桁落ちや情報落ちを防ぐ最も単純な方法はすべての計算を倍精度で行うことである．しかし，倍精度計算を行ったからといって桁落ちや情報落ちがなくなるわけではない．その影響があまり目立たなくなるだけである．

多くの場合，計算法を少し工夫するだけで桁落ちを避けることができる．上の例題の場合には次のようにすればよい．

$$\begin{aligned}\sqrt{1+x}-1 &= \frac{x}{\sqrt{1+x}+1} \\ &= \frac{1.23456 \times 10^{-3}}{2.00061} = 6.17092 \times 10^{-4}\end{aligned}$$

分母の計算で情報落ちが起きているがこれは結果には影響せず，上の結果は最後の桁まで正しい．

似たような例で，$\cos\theta > 0$ のとき $1-\cos\theta$ は

$$1-\cos\theta = \frac{\sin^2\theta}{1+\cos\theta}$$

とすれば桁落ちを避けることができる．

例2 二次方程式

$$x^2 - 6.28318x + 0.123456 = 0$$

の二根 x_1, x_2 を有効数字 6 桁で求める．

通常の計算法の結果は次のようになる．

$$\text{判別式} = 39.4784 - 0.493824 = 38.9846$$
$$x_1,\, x_2 = \frac{6.28318 \pm \sqrt{38.9846}}{2}$$
$$= \frac{6.28318 \pm 6.24376}{2} = 6.26347,\ 0.01971$$

x_1 の方は正しく求められているが，x_2 の計算では 2 桁の桁落ちが生じている．しかしこれも根と係数の関係を用いて

$$x_2 = \frac{0.123456}{x_1} = 0.0197105$$

とすれば最後の桁まで正しく求めることができる．

例3 $x^2 - y^2$

たとえば

$$x = 37507 \qquad y = 37496$$

のとき

$$x^2 - y^2 = 1.40678 \times 10^9 - 1.40595 \times 10^9$$
$$= 8.3 \times 10^5$$

となって 4 桁の桁落ちが生じるが

$$(x-y)(x+y) = 11 \times 75003 = 825033$$

とすれば $x-y$ の計算で 3 桁の桁落ちが生じるが，x, y が誤差を含んでいないとすれば後者の計算の結果は最後の桁まで正しい．この例では $x^2,\ y^2$ が 10 桁以上になり，そこで情報落ちが起きたために後の計算で桁落ちが生じたのである．

例4 $\sin\theta - \sin\theta_0$

このような計算は地球を球としたときの震央距離を計算するときなどによく現れる．これは

$$\sin\theta - \sin\theta_0 = 2\cos\frac{\theta+\theta_0}{2}\sin\frac{\theta-\theta_0}{2} \quad (1.4)$$

によって計算する．たとえば

$$\theta_0 = 35^\circ \qquad \theta = 35^\circ 30'$$

のとき

$$\sin\theta = 0.580703 \qquad \sin\theta_0 = 0.573576$$

であるから式のままに計算すれば 2 桁の桁落ちが生じるが，$\theta - \theta_0$ が正しければ (1.4) 式で計算すれば正しい値が得られる．

例5 標本平均と標本分散

N 個のデータ $x_i\,(1 \le i \le N)$ の標本平均 $\bar{x}$ と標本分散 s^2 の公式として

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (1.5)$$

$$s^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2 \quad (1.6)$$

をあげてある教科書が多い．これは x_i の和と二乗和を一つのループで計算できるためと思われる．この公式を用いて次のようなデータの平均や分散を計算してみる．

i	x_i	i	x_i	i	x_i
1	348.200	5	350.441	9	352.683
2	338.541	6	335.953	10	346.906
3	355.271	7	343.024	11	344.318
4	342.076	8	349.148	12	340.783

この表から次のような値が得られる．

$$\bar{x} = \frac{4147.35}{12} = 345.613$$
$$\frac{1}{N}\sum x_i^2 = 119479$$
$$s^2 = 119479 - (345.613)^2$$
$$= 119479 - 119448 = 31$$

ここでは $\bar{x}$ の計算には問題はないが，s^2 の計算では 4 桁の桁落ちが生じている．

これを避けるためには仮の平均を用いればよい．x_i の仮の平均を $\tilde{x}$ とすれば

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \tilde{x})^2 + \tilde{x} \quad (1.7)$$

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \tilde{x})^2 - (\bar{x} - \tilde{x})^2 \quad (1.8)$$

が成り立つ．前にあげた式 (1.6) は $\tilde{x}=0$ に相当している．$\tilde{x}=\bar{x}$ のときは標本分散の定義そのものである．$\tilde{x}$ はなるべく真の平均 $\bar{x}$ に近い値を選ぶのが望ましいが，厳密な平均値はわからないから適当な値でよい．ここで与えられたデータの範囲が 335 から 355 の間にあることに注目してここではその中央

$$\tilde{x}=345$$

を選ぶと，先の表は次のようになる．

i	$x_i-\tilde{x}$	i	$x_i-\tilde{x}$	i	$x_i-\tilde{x}$
1	3.200	5	5.441	9	7.683
2	−6.459	6	−9.047	10	1.906
3	10.271	7	−1.976	11	−0.682
4	−2.924	8	4.148	12	−4.217

(1.7)，(1.8) 式を用いて計算すれば

$$\bar{x}-\tilde{x}=\frac{1}{N}\sum(x_i-\tilde{x})=0.612$$
$$s^2=\frac{379.474}{12}-0.612^2$$
$$=31.6228-0.374544=31.2483$$

二つの計算を比べればわかるように，第一の計算法では情報落ちが起こっている．$(1/N)\sum x_i$ の主要項は $\tilde{x}$ であり，$(1/N)\sum x_i^2$ の主要項は $\tilde{x}^2$ である．本当に意味のある $\tilde{x}$ からのずれ，すなわち変動分 $x_i-\tilde{x}$ は情報落ちのために失われている．上の例では第一の計算法では平均 $\bar{x}$ は比較的正確に求められていたが，場合によっては平均値さえ求められないことがある．これに対して第二の計算法では変動分だけを対象にしているので精度が上がっている．

> 昔むかし，外国からの研修生に数値計算の実習をしていたときのこと，地震の走時のデータを与えて最小二乗法で走時曲線の勾配を計算させたところ，どうしても勾配が負になるという研修生が現れた．プログラムを調べてみると，走時が 13 時 24 分 49.2 秒とあったとすると，これを全部秒に換算しているからだとわかった．走時は秒以下の部分しか変化しないから，大事な変化分が情報落ちのために失われてしまっていたのである．

例 6　測地緯度 φ における正規 (標準) 重力は

$$\gamma=978.032681(1+0.005278970\sin^2\varphi$$
$$+0.000023461\sin^4\varphi)\quad[\mathrm{gal}]$$

で近似される．この近似式の精度は 4 μgal である．この式は厳密解を近似したものであるから，たとえば理科年表にでている $\sin^8\varphi$ 迄の展開式とは僅かながら係数が異なっている．

この式をそのまま用いて γ を μgal の桁まで計算するためには十進 9 桁が必要である．しかし，実際に重力測定を行う場合には緯度 φ の変化は僅かなものである．そこで，ある基準の緯度 φ_0 における正規重力 γ_0 との差を作ると

$$\gamma-\gamma_0=5.163005(\sin^2\varphi-\sin^2\varphi_0)$$
$$\times[(1+0.0088885\sin^2\varphi_0)$$
$$+0.0044442(\sin^2\varphi-\sin^2\varphi_0)]$$

となる．

$$\sin^2\varphi-\sin^2\varphi_0=\sin(\varphi+\varphi_0)\sin(\varphi-\varphi_0)$$

であるから，$\varphi-\varphi_0=\pm10^\circ$ であったとしても正規重力の変動分 $\gamma-\gamma_0$ はたかだか

$$5.16\sin70^\circ\sin10^\circ=0.84\quad[\mathrm{gal}]$$

である．したがって $\gamma-\gamma_0$ を μgal の桁まで計算するには有効数字 6 桁で十分である．一例として，$\varphi_0=35^\circ$ のとき

$$\sin^2 35^\circ=0.3289899$$

であるから

$$\gamma-\gamma_0=5.178103\sin(\varphi+35^\circ)\sin(\varphi-35^\circ)$$
$$\times[1+0.002257\sin(\varphi+35^\circ)\sin(\varphi-35^\circ)]$$

となる．これも変動分だけに注目した計算法である．

例 7　交代級数

指数関数 e^x はテーラー展開

$$e^x=1+\frac{x}{1!}+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots$$

によって表される．この級数の収束半径は無限大であるから，代数的には x が何であっても収束する．しかし数値的にはそうはならない．たとえば $x=-10$ のとき，次のような結果が得られる．

n	a_n	S_n
15	−7.64717e+2	−2.98494e+2
16	4.77948e+2	1.79454e+2
17	−2.81146e+2	−1.01692e+2
18	1.56192e+2	5.45001e+1
19	−8.22064e+1	−2.77063e+1
20	4.11032e+1	1.33969e+1

ここに a_n はテーラー展開の n 次の項，S_n は n 次までの部分和を表している．また，e+2 は 10^2 を意味している．この表からわかるように，この計算は収束していない．

一方，$x = +10$ のときには

n	a_n	S_n
15	7.64717e+2	2.09529e+4
16	4.77948e+2	2.14308e+4
17	2.81146e+2	2.17120e+4
18	1.56192e+2	2.18682e+4
19	8.22064e+1	2.19504e+4
20	4.11032e+1	2.19915e+4

となって収束している (正しい値は 2.20265e+4 である)．したがって e^{-10} を計算するには e^{10} を求めた後，逆数 $1/e^{10}$ を計算すればよい．

交代級数は桁落ちが激しい上に収束が遅いので，さまざまな加速法が提案されている．

交代級数に限らず，級数を計算するときには絶対値の小さい方から加えていくのが原則である．たとえば，有効数字 6 桁のとき，数値 654321 に 0.1 を何回加えても値は変わらない．しかし 0.1 を 10 回加えた後に 654321 を加えれば結果は 654322 になる．

2 数の表現

計算機内部では数値に限らずすべての情報は 0，1 のビットパターンで表現されている．ビットパターンをそのまま表すと非常に冗長になるしわかりにくいので，8 進数や 16 進数を用いて表すのが普通である．以下に対応表を示しておく．4 ビットをひとまとまりとしたものを**バイト**という．1 バイトで 0 から 15 までの 16 進数 1 桁を表すことができる．16 進数では 10 以上の数をアルファベット a，b，c, d, e, f や, それらの大文字で表すのが習慣である．

10 進	0	1	2	3	4	5	6	7
2 進	0	1	10	11	100	101	110	111
10 進	8	9	10	11	12	13	14	15
2 進	1000	1001	1010	1011	1100	1101	1110	1111
8 進	10	11	12	13	14	15	16	17
16 進	8	9	a	b	c	d	e	f

数値に関していえば，整数と実数は異なった形式で表現するのが普通である．整数はほとんどのシステムで同じ表現が用いられているが，実数（小数点のある数）の表現はシステムによって異なる．

整数の表現　ほとんどのシステムでは整数は二進法そのものを内部表現として用いている．ただし，負数を表現するためには符号と絶対値という方法もあるが，多くの場合には 2 の補数という表現法を用いている．例として 3 ビットの場合には次の表のようになる．

この表からわかるように，最高位のビットが 1 のときは負数を意味する．また，許される絶対値の最大が正数と負数とでは異なっていることに注意する．したがって現在広く用いられている 32 ビットで表現できる整数 n は

$$-2^{31} = 2147483648 \le n \le 2^{31} - 1 = 2147483647$$

の範囲である．

10 進表現	内部 2 進表現	8 進数
-4	100	4
-3	101	5
-2	110	6
0	000	0
1	001	1
2	010	2
3	011	3

実数の表現　実数 x には次のような表現を用いる．

$$\begin{aligned} x &= \pm(d_1\beta^{-1} + d_2\beta^{-2} + \cdots + d_t\beta^{-t})\beta^e \\ &= \pm m\beta^e \end{aligned} \tag{2.1}$$

$$0 \le d_i < \beta \quad 0 \le m < 1$$

β は進法の基数を表し，m は仮数部，e は指数部を表す．正負の符号は整数のときと同様に仮数部を補数表現にして表すこともあるが，現在では符号を独立に表すことが多い．

IBM，HITACHI などのいわゆるメインフレームの計算機では $\beta = 16$ の 16 進表現を用いることが多いが，UNIX 系のシステムでは二進法を用いている．

例 1 $x = 0.25$ のとき

$$\beta = 2 \quad x = \frac{1}{2} \cdot 2^{-1} \quad e = -1 \quad d_1 = 1$$
$$m = (0.1000\cdots0)_2$$
$$\beta = 16 \quad x = \frac{4}{16} \cdot 16^{0} \quad e = 0 \quad d_1 = 4$$
$$m = (0.0100\ 00\cdots0)_2$$

と表される．ここに $(\cdots)_2$ は二進数を意味している．これでわかるように，16 進法にくらべて二進法の方がビットを有効に利用している．

例 2 $x = 0.1$ は二進法では無限小数

$$\beta = 2 \quad e = -3 \quad m = (0.11001100110\cdots)_2$$

になる．したがって有限のビット長では 0.1 を正確に表現することはできない．このために，たとえば Fortran プログラム

```
      x=0
    1 x=x+0.1
        ........
      if( x.ne.1 )  goto  1
        ........
```

は無限ループに陥る．

UNIX における実数表現 仮数部，指数部，符号をメモリーの中にどのように配置するかについてはいろいろの方法が考えられる．UNIX 系のシステムでは次のようになっている．これは IEEE の規格に基づいている．なおこれは 1 語が 32 ビットの単精度の場合である．

まず x を二進法で

$$x = \pm(1.d_1d_2\cdots d_{23})_2 \times 2^e$$

と表す．このように最高位の桁が 1 になっている表現を正規化された表現と呼ぶ．このとき，1 語 32 ビットの配置は上位から次のようになる．

符号 (1bit)	指数部 (8bits)	仮数部 (23bits)
$0/1$	$e + 127$	$d_1d_2\cdots d_{23}$

仮数部の最高位は 1 に決まっているので省略してよい．符号は正のときは 0，負のときは 1 である．指数部に $127 = 2^7 - 1$ を加えてあるのは負の指数を正の数で表すためである．指数部が 8 ビットであるから，許される e の範囲は

$$0 \le 2^7 - 1 + e \le 2^8 - 1 \qquad -127 \le e \le 128$$

になる．したがって許される数の絶対値はおよそ

$2^{-127} = 5.9 \times 10^{-39}$ から $2^{128} = 3.4 \times 10^{38}$

までになる．

いくつか例をあげる．

$$2.0 = (1.0)_2 \times 2^1 \to (0100\ 0000\cdots)_2$$
$$= (4000\ 0000)_{16}$$
$$1.5 = (1.1)_2 \to (0011\ 1111\ 1100\ 0000\cdots)_2$$
$$= (3\text{fc0}\ 0000)_{16}$$
$$1.0 = (1.0)_2 \to (0011\ 1111\ 1000\ 0000\cdots)_2$$
$$= (3\text{f80}\ 0000)_{16}$$
$$0.1 = (1.10011001100\cdots)_2 \times 2^{-4}$$
$$\to (0011\ 1101\ 1100\ 1100\cdots)_2$$
$$= (3\text{dcc}\ \text{cccd})_{16}$$

矢印 $\to$ の後ろは内部表現を二進数で表したもので，その次は 16 進数で表したものである．負数のときには最高位のビットを 1 で置き換えればよい．

16 進法による内部実数表現 基数 β が 16 のときの内部表現は

符号 (1bit)	指数部 (7bit)	仮数部 (24bit)
$0/1$	$e + 64$	$d_1d_2\cdots d_{24}$

となっている．仮数部のビット数が 24 と UNIX 系より 1 ビット多くなっているが，UNIX 系では実質的には 24 ビットであるし，16 進法ではもともとビットの使い方が効率的ではないので精度の向上にはあまり役に立っていない．指数部のビット数が減って

いるが 16 進法を用いているために数値の範囲は二進法より広がり，最小と最大の数はおよそ

$$16^{-65} = 5.4 \times 10^{-79} \text{ から } 16^{63} = 7.2 \times 10^{75}$$

と広がっている．

オーバーフローとアンダーフロー　どのような処理系を用いるにせよ，扱うことができる数値の絶対値には下限と上限がある．計算の結果が下限より下回った場合はアンダーフローとして値を 0 としてしまうのが普通である．上限より上回ったときはオーバーフローである．かつてはオーバーフローのときには処理を中止することが多かったが，最近はそのまま処理を続行することが普通である．したがって計算が最後まで行われたからといって正しい結果が得られたとは限らない．

> いまでは考えられないことであるが，計算機の黎明期時代には内部十進法を用いた計算機もあったし，1 語 36 ビットの計算機もあった．十進法を用いれば入出力で十進・二進の変換をする手間を省けるというメリットはある．36 ビットの処理系は数値計算の精度は確かに高かったが，ほかの処理系とのデーターの互換性には問題があった．36 ビットとは 6 ビットで一文字を表した時代の名残で，これはまた 6 孔の紙テープを用いたテレタイプ型の入出力装置に関係している．

例 3　$\sqrt{x^2+y^2}$

上で述べたように，計算機内部では許される数の範囲が決まっている．たとえ x や y が許される数の範囲内であっても，x^2 や y^2 が許される範囲を上に越えたり (オーバーフロー)，下に越えたりすることがあるが (アンダーフロー)，$\sqrt{x^2+y^2}$ 自体は許される範囲内にあることがある．かつては許容範囲を越えると (特にオーバーフローの場合) 計算を強制的に終了してしまうことが多かったが，最近のシステムでは強制終了しないで最後まで計算を続けてしまうことが多い．その結果は NaN(not a number) になる．

このような計算では，x の絶対値が y の絶対値よりも大きいときには

$$\sqrt{x^2+y^2} = |x|\sqrt{1+(y/x)^2}$$

とすればオーバーフローを避けることができる．

例 4　$c/(a \times b)$

先の例と同様であるが，a や b が非常に大きいか小さくても，この式の値自体はオーバーフローあるいはアンダーフローしないことがある．このときには分母を先に計算するとオーバーフローやアンダーフローする危険があるので

$$c/a/b$$

とすればよい．

計算機イプシロン (machin epsilon)　計算機で計算を行う場合，その計算機の有効数字が何桁かを知っておくことは重要なことである．ほとんどの計算機は内部的には二進法ないしは 16 進法を用いているので有効数字も二進ないしは 16 進で考えなければならない．そのためには計算機内部の浮動小数点表示法を知る必要がある．

そこで以下では次のような小さな正数 ε_M を定義する．

$$\varepsilon_M = \mathrm{Min}\{\varepsilon : 1+\varepsilon > 1\}$$

ここで不等式は計算機内部で成り立つことを意味している．つまり，計算機内部で $1+\varepsilon > 1$ が成り立つ最小の ε を計算機イプシロンと呼ぶ．これはある計算機内部における有効数字の大きさを代表する数値になっている．

この数値を求めるためには，たとえば Fortran 言語では次のようなプログラムを走らせてみればよい．

```
      eps = 1
      do  i=1, 64
        eps1 = 1 + eps
        if( eps1.eq.1 ) goto  1
        eps = eps/2
      enddo
      stop
    1 eps = eps*2
      print eps
```

あるシステムの場合，単精度，倍精度について eps はそれぞれ次のようになった．

$$\mathrm{eps} = 5.96046 \times 10^{-8} \qquad 1.387778 \times 10^{-17}$$

もちろんこのプログラムで求められた値は定義通りの計算機イプシロンではないが，有効数字のおおよその大きさを知るには十分な値である．

例 4 $a+(b+c)=(a+b)+c$

これは代数学の基本的な定理であるが，計算機内部では有限長の数値だけを扱っているために，この恒等式は成り立つとは限らない．逆のいい方ををすれば

$$y = x + \Delta x$$

の計算において $\Delta x \neq 0$ であっても $y = x$ が成り立ってしまうのである．このような計算は反復法でよく現れるパターンである．これを逆用して収束の判定に用いることも多い．たとえばプログラム

```
1 deltax = ...
   ......
  if( y.ne.x+deltax )  goto  1
   .....
```

では Δx が 0 にならなくてもループから下に抜け出す．なお，単精度の計算を行っていても，上の比較の判定は倍精度のレジスターで行われることが多い．このときには Δx が単精度で十分に小さくなってもループから抜け出すことができない．先の計算機イプシロンの計算で eps1=1+eps と置き換えているのは単精度同士の比較を行うためである．

丸め誤差のために実数同士の比較には注意が必要である．大小だけの比較なら危険性が少ないが，上の例のように等号 (あるいはその否定) のときには代数的には成り立っていても計算機内では成り立たないことがある．そのため昔の処理系では

```
if(x.eq.y)
```

のような文では，比較が正常に行われない恐れがある，というような警告が出たものである．警告が出ないからといって安心してはいけない．

測定器とのインターフェイス 最近の測定器は A/D コンバーターを通して，あるいはデータロッガーを通して結果がディジタルの情報として出力される．これらの出力は整数値のこともあれば，実数値のこともある．いずれにせよ，これらのディジタルデータをそのまま計算機に取り込んでもうまくいくことはほとんどない．ポイントは以下の点である．

- 1 データのバイト数
- 数値の表現法
  整数値なら負数の表現法，浮動小数点数の表現法 (二進法か 16 進法か)
- バイトの順番

16 ビット，24 ビットの A/D コンバーターからの出力なら 1 語 32 ビットの処理系に合わせるためには残りの 16 ビット，8 ビットに何を詰めるかが問題になる．

それ同時に問題なのはバイトの順番で，通常は下位のバイトから出力されるから，入力の際も下位から詰めていけばよいが，上位から出力される場合もあるので注意が必要である．

いずれにせよ，測定器からの出力はバイト単位で編集をしなければならないので，数値計算を越えた問題である．

3 多項式の零点

以下では多項式の計算法と，多項式の零点を求める方法について述べる．一部，ニュートン法など多項式以外にも適用できる方法についても述べてある．ほとんどの方法は複素係数の多項式に対しても適用可能であるが，実用的に重要な実係数の場合を念頭に置いている．

二次方程式の根　一次方程式の根は自明であるから，二次方程式

$$x^2 + 2bx + c = 0 \tag{3.1}$$

から始める．二次方程式の根の公式はよく知られているように

$$x = -b \pm \sqrt{b^2 - c} \tag{3.2}$$

である．しかしこの公式をそのまま用いると精度が出ないことは既に §1 に述べてある．たとえば，b^2 に比べて c の絶対値が非常に小さいときには複号のどちらかの計算で桁落ちが起こる可能性がある．したがって実係数の場合，精度を上げるためには次のようにした方がよい．

$$\begin{aligned} x_1 &= -b - \sqrt{b^2 - c} \quad x_2 = \frac{c}{x_1} \quad b > 0 \\ x_1 &= -b + \sqrt{b^2 - c} \quad x_2 = \frac{c}{x_1} \quad b < 0 \end{aligned} \tag{3.3}$$

また，b の絶対値が非常に大きい場合には，根そのものは制限の範囲内であるが，b^2 の計算でオーバーフローが起きるという可能性も残されている．これも §1 で述べてある．

三次方程式の根　三次方程式

$$y^3 + a_2 y^2 + a_1 x + a_0 = 0$$

は原点移動

$$y = x - \frac{a_2}{3}$$

によって二次の項を欠く形

$$x^3 + 3px + q = 0 \tag{3.4}$$

に帰着させることができる．ここで

$$\begin{aligned} p &= \frac{1}{3}a_1 - \left(\frac{a_2}{3}\right)^2 \\ q &= \frac{a_2}{3}\left[2\left(\frac{a_2}{3}\right)^2 - a_1\right] + a_0 \end{aligned}$$

である．

この方程式の根として

$$x = u + v \tag{3.5}$$

の形を仮定すると，もとの方程式は

$$u^3 + v^3 + q + 3(u+v)(uv+p) = 0$$

と書き換えられる．これが成り立つためには

$$u^3 + v^3 = -q \qquad uv = -p$$

が成り立てば十分である．したがって二次方程式の根と係数の関係から

$$\begin{aligned} u^3 &= \frac{1}{2}[-q \pm \sqrt{q^2 + 4p^3}] \\ v^3 &= -\frac{p^3}{u^3} \end{aligned} \tag{3.6}$$

が得られる．ここで u^3 は二根のうちのどちらか一つでよいが，二次方程式のときと同様に桁落ちの起きない方を選ぶものとする．

実係数のときには，判別式

$$D = q^2 + 4p^3 \tag{3.7}$$

の正負によって根が実数になったり複素数になったりする．判別式 D が正のときには u^3，v^3 が実数になるから，その三乗根 u，v は 1 実根，2 複素根になる．したがって (3.5) 式から，三次方程式 (3.4) の根は 1 実根，2 複素根になる．

判別式が負のときには u^3，v^3 が複素数になるが，これらは

$$\begin{aligned} u^3 &= \sqrt{-p^3}e^{i\varphi} \quad v^3 = \sqrt{-p^3}e^{-i\varphi} \\ \tan\varphi &= \frac{\sqrt{-D}}{-q} \quad D < 0 \end{aligned} \tag{3.8}$$

と書くことができる．$D < 0$ であるから p はかならず負である．これらの三乗根は

$$\begin{aligned} u_1 &= \sqrt{-p}e^{i\varphi/3} & v_1 &= \sqrt{-p}e^{-i\varphi/3} \\ u_2 &= u_1 e^{2\pi i/3} & v_2 &= v_1 e^{-2\pi i/3} \\ u_3 &= u_1 e^{-2\pi i/3} & v_3 &= v_1 e^{2\pi i/3} \end{aligned} \tag{3.9}$$

となり複素数であるが，u と v は互いに複素共役になっているので，(3.5) 式に代入すると三根ともに実数になる．

四次方程式の根 四次方程式

$$y^4 + a_3y^3 + a_2y^2 + a_1y + a_0 = 0$$

も原点移動

$$y = x - \frac{a_3}{4}$$

によって三次の項を欠く形

$$x^4 + px^2 + qx + r = 0 \tag{3.10}$$

に直すことができる．ここに

$$p = a_2 - 6\left(\frac{a_3}{4}\right)^2$$
$$q = a_1 - 2a_2\left(\frac{a_3}{4}\right) + 8\left(\frac{a_3}{4}\right)^3$$
$$r = a_0 - a_1\left(\frac{a_3}{4}\right) + a_2\left(\frac{a_3}{4}\right)^2 - 3\left(\frac{a_3}{4}\right)^4$$

である．(3.10) 式を二乗の差

$$(x^2 + \alpha)^2 - (\beta x - \frac{q}{2\beta})^2 = 0$$

に書き換えたとする．上式と (3.10) 式の係数を比較すれば

$$2\alpha - \beta^2 = p \qquad \alpha^2 - \frac{q^2}{4\beta^2} = r \tag{3.11}$$

が得られる．これらの式から β を消去すれば

$$4(\alpha^2 - r)(2\alpha - p) = q^2$$

が得られる．これは α に関する三次方程式である．実係数の場合，この方程式は必ず実根を持つ．この実根を改めて α とすればそれに応じた β が (3.11) 式から得られる．これらの α, β を用いて二組の二次方程式

$$x^2 + \beta x + \alpha - \frac{q}{2\beta} = 0$$
$$x^2 - \beta x + \alpha + \frac{q}{2\beta} = 0 \tag{3.12}$$

を解けば四次方程式 (3.10) の根が得られる．

組立除法（ホーナーの方法） 五次以上の代数方程式には根の公式がないことはよく知られている．したがって五次以上の高次方程式の根を求めるにはなんらかの形の反復法を用いなければならない．その際には多項式の値やその微分を計算する必要がある．

以下では n 次の多項式を

$$p_n(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots$$
$$\cdots + a_1x + a_0 \tag{3.13}$$

と書く．$p_n(x)$ の値を計算するのに，x の k 乗に a_k を掛けて加えるというやりかたはしない．上式を

$$p_n(\mu) = (\cdots((a_n\mu + a_{n-1})\mu + a_{n-2})\mu$$
$$+ \cdots + a_1)\mu + a_0$$

と書き換えて最も内側の括弧の中から計算すれば

$$b_{n+1} = 0$$
$$b_k = a_k + \mu b_{k+1} \tag{3.14}$$
$$k = n, n-1, \cdots, 0$$

という漸化式が得られる．ここで最後に得られた b_0 が $p_n(\mu)$ の値

$$p_n(\mu) = b_0$$

である．

この計算の途中で得られた b_k も意味のある量である．$p_n(x)$ を $x - \mu$ で割り算して

$$p_n(x) = (x - \mu)(b_nx^{n-1} + b_{n-1}x^{n-2} +$$
$$\cdots + b_2x + b_1) + b_0 \tag{3.15}$$

と置く．(3.13)，(3.15) 式の係数を比べれば，上式の b_k が全く同じ漸化式 (3.14) を満たしていることがわかる．すなわち，漸化式 (3.14) で得られる量 b_k は $p_n(x)$ を $x - \mu$ で割ったときの商の多項式の係数と余りを意味している．

この商の多項式

$$b_nx^{n-1} + \cdots + b_2x + b_1$$

をもう一度 $x - \mu$ で割るともとの多項式 $p_n(x)$ が

$$p_n(x) = (x - \mu)[(x - \mu)(b_nx^{n-2} + \cdots$$

という形に表されることになる．これを $n - 1$ 回繰り返すと

$$p_n(x) = d_n(x - \mu)^n + d_{n-1}(x - \mu)^{n-1} +$$
$$+ \cdots + d_0$$

の形の式の係数 d_k が得られる．ここで $x = w + \mu$ と置けば上式は w の多項式になる．いいかえれば，多項式の原点移動したときの係数が求められることになる．この方法は後でも用いられる．

二次式による割り算もよく用いられる．いま改めて

$$p_n(x) = (x^2 - ux - v)(b_n x^{n-2} + b_{n-1}x^{n-3} + \cdots + b_3 x + b_2) + b_1(x - u) + b_0 \tag{3.16}$$

と置いて両辺の係数を比較すれば

$$\begin{aligned} b_{n+1} &= b_{n+2} = 0 \\ b_k &= a_k + ub_{k+1} + vb_{k+2} \\ k &= n, n-1, \cdots, 0 \end{aligned} \tag{3.17}$$

が得られる．特に

$$u = 2\mu \qquad v = -\mu^2$$

と置けば

$$p_n(x) = (x - \mu)^2(b_n x^{n-2} + b_{n-1}x^{n-3} + \cdots + b_3 x + b_2) + b_1(x - 2\mu) + b_0$$

であるから

$$p_n(\mu) = b_0 - \mu b_1 \qquad p_n'(\mu) = b_1 \tag{3.18}$$

となる．この関係は後で述べるニュートン法で高次方程式の根を計算するときに役に立つ．

挟み打ち法（線型内挿） 代数方程式に限らず，実数方程式

$$y = f(x) = 0 \tag{3.19}$$

の根を求めることを考える．$f(x)$ の符号の変化を調べて根が (x_1, x_2) の間にあることがわかったとする．この範囲で $f(x)$ を直線で近似すれば

$$f(x) \sim y_1 + \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$$

となる．$f(x) = 0$ の次の近似根は

$$x_3 = x_1 - \frac{x_2 - x_1}{y_2 - y_1}y_1 \tag{3.20}$$

である．通分した形 $(x_1y_2 - x_2y_1)/(y_2 - y_1)$ を用いないのは丸め誤差小さくするためである．反復法では上のように，もとの値，足す補正量という形に表すのが定石である．

こうして求められた $y_3 = f(x_3)$ と y_1 または y_2 の中で y_3 と符号の異なるものを用いて次の根を求めるという手順を繰り返せば根が求められる．

根の近似値 x_k の真の根 x からのずれを ε_k とし

$$x_k = x + \varepsilon_k \tag{3.21}$$

とする．このとき $f(x) = 0$ であるから

$$y_1 = f(x_1) = f(x + \varepsilon_1) = f'(x)\varepsilon_1 + \frac{1}{2}f''(x)\varepsilon_1^2$$

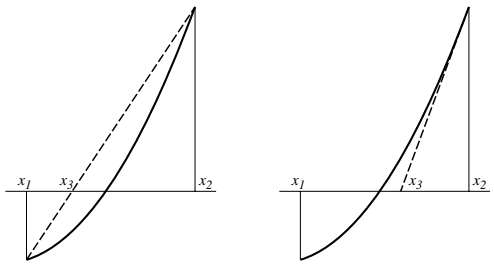
等が成り立つ．これらを (3.20) 式に代入すれば x_3 の誤差は

$$\varepsilon_3 = \frac{f''(x)}{2f'(x)}\varepsilon_1\varepsilon_2 \tag{3.22}$$

と表わされる．グラフを書いてみればわかるように，何回か反復を行うと近似根が一方向から根に近づく．下図の左のような例では点 x_2 は変わらないので，ε_2 は定数である．したがって一回の反復で誤差は前回の誤差 ε_1 の

$$\frac{f''(x)}{2f'(x)}\varepsilon_2$$

倍になる．このような収束の仕方を一次収束という．



この方法は $f(x)$ の関数値のみを用いていること，内挿が単純であるためにローバスト (robust，頑丈) な方法であるが収束は遅い．また，上で述べたように一方向から根に近づくために真の根から離れたところで収束と判定してしまうことがある．これを避けるためには，途中で二分法等を併用すればよい．たとえば先の図の例で四番目の近似根を一次近似で求める代わりに

$$x_4 = \frac{x_2 + x_3}{2}$$

とすれば，一次近似だけをするよりも狭い範囲に根を追い込むことができる．

逆内挿法 内挿の近似を高めれば収束が速くなることが期待される．そこで一次近似ではなく，二次近似を行うことを考える．このときに重要なことは，$f(x)$ を x の二次式で近似するのではなく，$f(x)$ の逆関数 $x = f^{-1}(y)$ を y の二次式で近似することである．いま，三点 (x_1, x_2, x_3) における関数値 (y_1, y_2, y_3) が与えられているとき，次のようなラグランジェの補間公式が成り立つ．

$$x \sim \frac{(y-y_2)(y-y_3)}{(y_1-y_2)(y_1-y_3)}x_1 + \frac{(y-y_3)(y-y_1)}{(y_2-y_3)(y_2-y_1)}x_2 + \frac{(y-y_1)(y-y_2)}{(y_3-y_1)(y_3-y_2)}x_3 \tag{3.23}$$

この式が $y = y_i$ のとき $x = x_i$ となることは明らかである．われわれが必要なのは $y = f(x) = 0$ になるときの x の値であるから，上式で $y = 0$ と置いて

$$x_4 = \frac{y_2y_3}{(y_1-y_2)(y_1-y_3)}x_1 + \frac{y_3y_1}{(y_2-y_3)(y_2-y_1)}x_2 + \frac{y_1y_2}{(y_3-y_1)(y_3-y_2)}x_3 \tag{3.24}$$

が得られる．しかしこの式で計算すると桁落ちが生じるので，次のような計算法をとる方がよいし，計算も簡単である．

いま，二点 x_1, x_j 間の一次近似で求めた根の近似値を

$$x_{[1j]} \equiv x_1 - \frac{x_j - x_1}{y_j - y_1}y_1 \quad j = 2, 3 \tag{3.25}$$

と表すことにする．(3.20) 式の x_3 はこの表記法では $x_{[12]}$ である．任意の x_3 を用いて逆内挿法で求めた近似値 x_4 は

$$x_4 = x_{[123]} \equiv x_{[12]} - \frac{x_{[13]} - x_{[12]}}{y_3 - y_2}y_2 \tag{3.26}$$

と書くことができる．これは点の順序にはよらない．(3.24) 式から (3.26) 式を導くのは大変であるが，(3.26) 式に (3.25) 式を代入して x_1, x_2, x_3 の係数を計算すれば (3.24) 式になることが確かめれれる．

先と同じようにして誤差を定義すると x_4 の誤差は

$$\varepsilon_4 = \frac{f''}{2f'}\left(\frac{f''}{f'} - \frac{f'''}{3f''}\right)\varepsilon_1\varepsilon_2\varepsilon_3 \tag{3.27}$$

と表わされる．$f(x)$ などの引数 x は省略してある．上式では x_1, x_2, x_3 は任意であるが，x_3 として特に x_1 と x_2 を内挿した値

$$x_3 = x_{[12]}$$

に選ぶと，その誤差は (3.22) 式で与えられる．したがってこのときには

$$\varepsilon_4 = \left(\frac{f''}{2f'}\right)^2\left(\frac{f''}{f'} - \frac{f'''}{3f''}\right)(\varepsilon_1\varepsilon_2)^2 \tag{3.28}$$

になる．

4 点目 x_4 が与えられたとき $x_{[1234]}$ を計算する式は (3.26) 式から容易に想像できるであろう．しかしあまり高次の内挿式を用いると内挿値が飛んでしまうことがあるので，最新の 3 点を用いて (3.26) 式にとどめておくのが安全である．

ニュートン法 上では関数値だけを用いて根を求める方法であったが，微係数を計算することができれば，もっと能率のよい方法がある．

いま k 番目の根の近似値を x_k とし，それに対する補正値を Δx とする．これは

$$f(x_k + \Delta x) = 0$$

を満たさなければならない．上式を展開すれば

$$f(x_k) + f'(x_k)\Delta x + \cdots = 0$$

となる．二次以上の項を無視すれば次の近似値として

$$x_{k+1} = x_k + \Delta x = x_k - \frac{f(x_k)}{f'(x_k)} \tag{3.29}$$

が得られる．この式の幾何学的な意味は先の図の右図から明らかである．

k 番目の近似根の誤差を ε_k とすると

$$\varepsilon_{k+1} = \frac{f''(x)}{2f'(x)}\varepsilon_k^2 \tag{3.30}$$

が得られる．すなわち，誤差は前回の誤差の二乗に比例する．このような収束を二次の収束という．これは一回反復を行うと有効数字の桁数が約二倍になることを意味している．重根のところでは上式の分母が 0 になるのでこの式は成り立たない．このときには ε_{k+1} は ε_k に比例，すなわち一次の収束になる．

ニュートン法が収束するためには初期値をうまく選ぶ必要がある．このことは方程式のグラフを書いてみれば理解できる．証明は省略するが収束の十分条件は

$$\left|\frac{f''(x)f(x)}{[f'(x)]^2}\right| < 1$$

で与えられる．ただし，ここでの x は真の根という意味ではなく，根を捜している領域の x という意味である．

平方根　ニュートン法の例として平方根の計算をとりあげる．a の平方根は

$$f(x) = x^2 - a = 0$$

の根である．(3.29) 式から反復法の公式は

$$x_{k+1} = x_k - \frac{x_k^2 - a}{2x_k} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right)$$

となる．このような計算法は Fortran や C などのコンパイラーの数学ライブラリーでも用いられている．

数値例として $\sqrt{2}$，$\sqrt{3}$，$\sqrt{10}$ を計算してみる．

k	$\sqrt{2}$	$\sqrt{3}$	$\sqrt{10}$
0	1.0	1.5	5.0
1	1.5	1.75	3.5
2	1.41666666	1.73214286	3.17857143
3	1.41421569	1.73205081	3.16231942
4	1.41421356		3.16227766

有効桁数が急速に増加しているのがわかる．

立方根　a の立方根は二つの方程式

$$f_1(x) = x^3 - a = 0$$
$$f_2(x) = x^2 - \frac{a}{x}$$

のどちらの根としても求めることができる．それぞれ対応する反復公式は

$$f_1 : x_{k+1} = \frac{1}{3}\left(2x_k + \frac{a}{x_k^2}\right)$$
$$f_2 : x_{k+1} = \frac{x_k(1 + 2a/x_k^3)}{2 + a/x_k^3}$$

となる．$\sqrt[3]{10}$ の計算は次のようになる．

k	f_1	f_2
0	3.0	3.0
1	2.37037037	2.20312500
2	2.17350863	2.15445072
3	2.15460159	2.15443469
4	2.15443470	

おなじ初期値を用いると f_2 の反復の方が速く収束している．これは f_2 の二階微分

$$f_2''(x) = \frac{2(x^3 - a)}{x^3}$$

が根 $x = \sqrt[3]{a}$ で 0 になるために，誤差が二次よりも速く 0 に収束するからである．ただし，f_2 の反復公式の方が f_1 のそれよりも複雑になっている．

逐次代入法　この方法もニュートン法と同様に多項式でないときにも利用できる方法である．解くべき方程式が

$$x = \varphi(x) \tag{3.31}$$

の形をしているとき，根の近似値 x_k から次の近似値を

$$x_{k+1} = \varphi(x_k) \tag{3.32}$$

のように逐次代入によって求めることができる場合がある．この方法が収束するためには

$$|\varphi(x) - \varphi(y)| \le q|x - y| \quad 0 < q < 1 \tag{3.33}$$

が成り立てば十分である．上式が成り立てば

$$|x_{k+1} - x_k| \le q|x_k - x_{k-1}|$$

になるからである．

例としてケプラー方程式

$$u = nt + e\sin u$$

をとりあげる．この方程式は惑星の軌道上の位置を決めるためのもので，n は平均角速度，t は時刻，e は軌道の離心率で，これらが与えられたときに u を解くのが問題である．u が求められれば時刻 t における惑星の黄経が求められる．$\varphi(u) = nt + e\sin u$ とすれば，惑星の離心率 e は 1 よりも小さいのでこの $\varphi(u)$ が条件 (3.33) 式を満たしていることは簡単な計算で明らかである．

そこで火星 ($e = 0.0934$) の場合を例にとって $nt = \pi/6$ のときの根を手計算で計算してみると次のようになった．

k	u_k
0	0.5235987
1	0.5702987
2	0.5488112
3	0.5479608
4	0.5479269
5	0.5479256

収束の速さはニュートン法に比べると遅いが，収束しているのはたしかである．

なお逐次代入法は積分方程式を解くときによく用いられ，その第一近似がボルン近似である．

ベアストウ・ヒッチコックの方法 実係数の代数方程式の根はニュートン法で求めることができるが，複素根のときには計算を複素数で行わなければならない．実係数の代数方程式のすべての根を実数計算だけで求める方法としてベアストウ・ヒッチコックの方法がある．この方法は二変数のニュートン法の応用例である．もちろんこの方法は複素係数の代数方程式にも適用できる．

n 次多項式 $p_n(x)$ を二次因子 x^2-ux-v で割り算して

$$p_n(x)=(x^2-ux-v)(b_nx^{n-2}+b_{n-1}x^{n-3}+\cdots+b_2)+b_1(x-u)+b_0 \tag{3.34}$$

と置く．係数 b_k は漸化式 (3.17) から求められる．もし余り b_1,b_0 が 0 であれば x^2-ux-v が $p_n(x)$ の因数である．これから $p_n(x)$ の零点が二個求められる．剰余は u,v の関数であることを考えると，割り切れるための条件は

$$b_0(u,v)=0 \qquad b_1(u,v)=0$$

でなければならない．ある (u,v) について割り算をしたときにはこの式が成り立たない．そこで (u,v) に加えるべき補正量を $(\Delta u,\Delta v)$ とするとこれらは

$$b_0(u+\Delta u,v+\Delta v)=0$$
$$b_1(u+\Delta u,v+\Delta v)=0$$

を満たしていなければならない．一変数のニュートン法と同様にして二次以上を無視すると

$$\frac{\partial b_0}{\partial u}\Delta u+\frac{\partial b_0}{\partial v}\Delta v=-b_0(u,v)$$
$$\frac{\partial b_1}{\partial u}\Delta u+\frac{\partial b_1}{\partial v}\Delta v=-b_1(u,v)$$

が得られる．これを解けば補正量が求められる．

補正量を求めるためには $\partial b_0/\partial u$ などの量を知らなければならない．そこでいま

$$c_k=\frac{\partial b_k}{\partial u} \qquad d_k=\frac{\partial b_{k-1}}{\partial v} \tag{3.35}$$

と置く．(3.17) 式を u で偏微分すれば

$$\begin{aligned} &c_{n+1}=c_{n+2}=0\\ &c_k=b_k+uc_{k+1}+vc_{k+2}\\ &\quad k=n,n-1,\cdots,0 \end{aligned} \tag{3.36}$$

が得られる．同様に v で偏微分すれば

$$\begin{aligned} &d_{n+1}=d_{n+2}=0\\ &d_k=b_k+ud_{k+1}+vd_{k+2}\\ &\quad k=n,n-1,\cdots,0 \end{aligned}$$

が得られる．二つの漸化式は全く同じであるから $d_k=c_k$ である．よって補正量を決める方程式は

$$\begin{aligned} &c_0\Delta u+c_1\Delta v=-b_0\\ &c_1\Delta u+c_2\Delta v=-b_1 \end{aligned} \tag{3.37}$$

である．この方程式から補正量 Δu，Δv が求められる．この方法はニュートン法であるから収束は二次である．

一例として，6 次多項式

$$p_6(x)=7x^6+6x^5+5x^4+4x^3+3x^2+2x+1 \tag{3.38}$$

の因数を求める．結果は次のようになった．長たらしい表を掲げるのは反復の初めには b_0 や b_1 が非常に大きな値になるが，一旦収束が始まると収束が非常に速いことを示したかったからである．

k	u	v	b_0	b_1
0	1.00000e+0	2.00000e+0	5.14000e+2	2.55000e+2
1	−8.17127e−1	5.12230e+0	1.46922e+3	−3.31247e+2
2	−6.30386e−1	3.54868e+0	4.55822e+2	−9.81376e+1
3	−5.28856e−1	2.36285e+0	1.38267e+2	−2.84699e+1
4	−4.78485e−1	1.51738e+0	4.15141e+1	−7.84127e+0
	···	···	···	···
10	−1.22970e+0	−4.14370e−1	6.34216e−1	−5.71124e−1
11	−1.16159e+0	−4.79867e−1	1.88651e−3	1.71509e−2
12	−1.26827e+0	−4.84868e−1	2.24267e−4	−4.07016e−4
13	−1.26822e+0	−4.84843e−1	1.83115e−8	−3.86946e−8

この手続きを繰り返して二次因子が求まれば $n-2$ 次式に対して同様な方法で二次因子を求めることができる．この方法は実係数のときには実数計算だけで複素根まで求めることができるが，問題点もある．一つは (u, v) の初期値をどのように選ぶかということである．ニュートン法は初期値が真の根に近くなければ収束しない．したがって (u, v) をうまく選ばないと根が得られないことがある．

もうひとつの問題は誤差の累積である．二次因子で割り算していくたびに誤差が累積して解くべき方程式が歪んでしまい，根の精度が低下してしまうからである．

DKA 法 最近，n 次多項式の零点を同時に求める方法が開発された．$p_n(z)$ の零点の近似値を $z_i (i = 1, 2, \cdots, n)$ とし，それらに加えるべき補正量を Δz_i とする．このとき

$$p_n(z) = a_n(z - z_1 - \Delta z_1)(z - z_2 - \Delta z_2) \cdots (z - z_n - \Delta z_n) \tag{3.39}$$

でなければならない．例によって二次以上の項を無視すれば

$$p_n(z) = a_n \prod_{i=1}^{n}(z - z_i) - a_n \sum_{i=1}^{n} \Delta z_i \prod_{j \neq i}(z - z_j)$$

が成り立つ．ここで $z = z_i$ と置けば右辺の第一項が 0 になるから

$$p_n(z_i) = -\Delta z_i a_n \prod_{j \neq i}(z_i - z_j)$$

したがって

$$\Delta z_i = -\frac{p_n(z_i)}{a_n \prod_{j \neq i}(z_i - z_j)} \qquad i = 1, 2, \cdots, n \tag{3.40}$$

が得られる．ニュートン法の補正量は (3.29) 式から

$$\Delta z_i = -\frac{p_n(z_i)}{p_n'(z_i)}$$

であった．(3.39) 式から微分 $p_n'(z)$ を計算して $z = z_i$ と置くと (3.40) 式の分母が得られる．したがってこれも一種のニュートン法であるから収束は二次である．単純なニュートン法との違いは，(3.40) 式の分母には z_i 以外の根の情報が入っていることである．このために根相互が干渉し合って一根だけが発散するようなことがなく，全体としてつじつまの合った根が求められるのである．

実は収束がもっと速い方法がある．(3.39) 式の両辺の対数微分をとると

$$\frac{p_n'(z)}{p_n(z)} = \sum_j \frac{1}{z - z_j - \Delta z_j}$$

となる．ここで $z = z_i$ と置けば

$$\frac{1}{\Delta z_i} = \sum_{j \neq i} \frac{1}{z_i - z_j} - \frac{p_n'(z_i)}{p_n(z_i)} \tag{3.41}$$

が得られる．この補正量 Δz_i を求めるためには関数値 $p_n(z_i)$ だけでなく，その微分 $p_n'(z_i)$ が必要になる．これは組み立て除法 (3.18) 式を用いれば $p_n(z_i)$ と同時に計算することができるから，(3.40) 式に比べて計算量が大きく増加することはない．なお，この方法は三次の収束である．

初期値の選択 ニュートン法では初期値の選択が重要になるが，次のような方法がある．原点を移動して $n-1$ 次の係数を 0 にした多項式

$$P(w) = \frac{1}{a_n} p_n\left(w - \frac{a_{n-1}}{n a_n}\right) = w^n + c_{n-2}w^{n-2} + \cdots + c_0 \qquad (3.42)$$

を作る．係数 c_k は組立除法を $n-1$ 回繰り返すことによって求められる．次にこれらの係数を用いた実係数多項式を

$$S(w) = w^n - |c_{n-2}|w^{n-2} - |c_{n-3}|w^{n-3} - \cdots - |c_1|w - |c_0| \qquad (3.43)$$

と置く．このときすべての係数が 0 という特別な場合を除いては，$S(w)$ は正の実軸上にはただ一つの零点をもつ．$w > 0$ に零点をもつことは

$$S(0) < 0 \quad S(\infty) > 0$$

から明らかである．$S(w)$ が $w = r_1, r_2$ に零点をもつとすれば

$$r_2^{-n}S(r_2) - r_1^{-n}S(r_1) = |c_{n-2}|(r_1^{-2} - r_2^{-2}) + \cdots |c_0|(r_1^{-n} - r_2^{-n}) = 0$$

が成り立たなければならない．しかし $r_1 \neq r_2$ のとき各項の符号は同じであるから，左辺が 0 になることはない．したがって $S(r)$ の零点は $w > 0$ には 1 個しかない．

そこでこの正の零点を

$$S(r_0) = 0 \qquad r_0 > 0 \qquad (3.44)$$

とする．このとき

$$S(r) > 0 \qquad r > r_0$$

であることを注意しておく．さて，$P(w)$ のすべての零点は半径 r_0 の円の内部

$$|w| \leq r_0 \qquad (3.45)$$

にある．これは次のようにして証明することができる．

いま，半径が r_0 の円の外に零点があったとしてこれを

$$w = re^{i\varphi} \qquad r > r_0$$

と置く．このとき

$$P(re^{i\varphi}) = r^n e^{in\varphi} + c_{n-2}r^{n-2}e^{(n-2)i\varphi} + \cdots + c_0 = 0$$

が成り立つ．全体を $e^{in\varphi}$ で割れば

$$r^n + c_{n-2}r^{n-2}e^{-2i\varphi} + \cdots + c_0 e^{-ni\varphi} = 0$$

が得られる．一方，$r > r_0$ であるから

$$S(r) = r^n - |c_{n-2}|r^{n-2} - \cdots - |c_0| \equiv \delta > 0$$

が成り立っている．二つの式の差を作れば

$$(|c_{n-2}| + c_{n-2}e^{-2i\varphi})r^{n-2} + (|c_{n-3}| + c_{n-3}e^{-3i\varphi})r^{n-3} + \cdots + (|c_0| + c_0 e^{-ni\varphi}) = -\delta$$

が導かれる．この式の右辺は負である．ところが左辺の各項は一般に複素数であり，しかも実数部は負になることはあり得ない．よって $P(w)$ が $|w| > r_0$ に零点を持つという仮定が誤りであったことが示された．

以上のことから初期値として次のようなものを選べばよいことがわかる．いま

$$S(r) > 0 \quad r > r_0$$

となるような r が求められたとする．r はできるだけ r_0 に近い方が後の反復が効率的であるが，それほど厳密に r_0 に近くなくてもよい．$S(r) > 0$ となる r が見つかった後，ニュートン法を二，三度繰り返せば十分である．このとき $p_n(z)$ の零点の近似値として，半径 r の円周上に等間隔に並んだ

$$z_j = -\frac{a_{n-1}}{n a_n} + r\exp\Big\{i\Big[\frac{2(j-1)\pi}{n} + \frac{\pi}{2n}\Big]\Big\} \qquad j = 1, 2, \cdots, n \qquad (3.46)$$

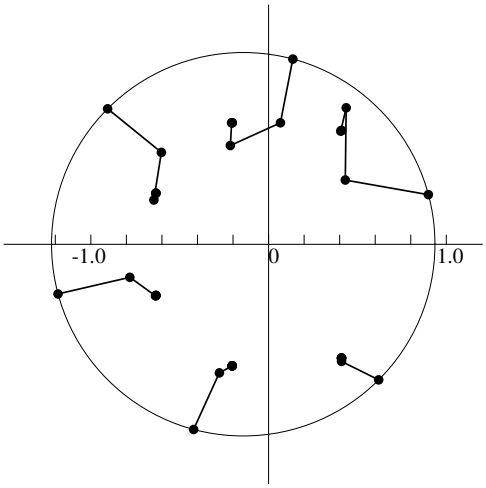
とすればよい．初期値が常に実軸から離れるようにしてあるのは反復 (3.40) が実軸に関する対称性を保存しているために，実根が奇数個あるときにも常に収束するようにするためである．なお，原点移動をするのは，ひとつには零点の存在範囲を狭くするためであるが，もうひとつには反復 (3.40) が

$$\sum z_i$$

を保存しているためである．

ニュートン反復 (3.40) は Durand と Kerner によって独立に提案された．初期値の選定法 (3.46) は Aberth によって提案された．よってここで述べた方法は DKA 法と呼ばれている．

6 次式 (3.38) に対して初期値を (3.46) 式で推定し，三次の反復法 (3.41) 式で計算した近似根の軌跡を下に示す．この図からわかるように，この計算は 4 回の反復で収束している．これはベアストウ・ヒッチコックの方法に比べてはるかに速いし，精度も高い．



多項式の零点の誤差　これまで高次代数方程式の根の計算法について述べてきたが，実はこの計算は次に述べるような意味で非常に不安定である．たとえば n 次多項式 $p_n(x)$ の係数 a_k に δ_k の誤差があったとする．このときに本来の零点 x_0 にどれくらいの誤差が生じるかを推定してみる．

$$(a_n+\delta_n)x^n+(a_{n-1}+\delta_{n-1})x^{n-1}+\cdots+(a_0+\delta_0)=0$$

の根を $x=x_0+\varepsilon$ 置き，ε,δ_k の一次までとると

$$p_n(x_0+\varepsilon)+\sum_{k=1}^{n}\delta_k x_0^k=0$$

$$p_n(x_0)+p_n'(x_0)\varepsilon+\sum_{k}\delta_k x_0^k=0$$

$$\varepsilon=-\frac{\sum_k \delta_k x_0^k}{p_n'(x_0)} \tag{3.47}$$

が得られる．高次方程式のときには δ_k が小さくてもこの値は非常に大きくなることがある．

このことは逆に考えるとよくわかる．n 次多項式の零点 x_j がたとえば M 桁までわかっていたとしよう．このとき

$$p_n(x)=a_n\prod_{j=1}^{n}(x-x_j)$$

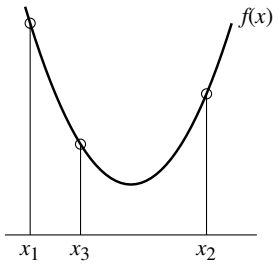
の x^k の係数を正しく知るためには $n\times M$ 桁が必要になる．一般には M 桁の計算においては，$p_n(x)$ の係数は M 桁までしか知ることはできない．したがって多項式の係数の精度と，その零点の精度には非常に大きな隔たりがある．

以上の考察から，多項式の零点を求める計算は可能ならばできるだけ避けた方がよいという結論が得られる．代表的な例は固有値問題である．行列の固有値は代数方程式の根の問題に帰着させることができる．理論的にはその通りであるが，数値的には上に述べた理由によりこれは最も精度が悪い方法である．後でも示すように，行列の要素が M 桁の精度で知られていれば，代数方程式を用いなくても M 桁の精度で固有値を計算する方法がある．

極小値の探索 (黄金分割)　実関数の根は正と負の両側から挟んでいけばよいが，極小値を求めるには最低 3 点が必要である．3 点が $x_1<x_3<x_2$ の順に並んでいるとき，実関数 $f(x)$ が $x_1<x<x_2$ に極小値をもつためには

$$f(x_3)<f(x_1)\qquad f(x_3)<f(x_2)$$

でなければならない．次に点 x_4 を選んで極小値を狭い範囲に追い込みたい．点 x_3 の左右のうち広い方に x_4 を選ぶのが常識的だろう．



そこで図のように x_3 の右側の区間の方が広いときには x_3 の右側に x_4 を選ぶことにし

$$\frac{x_3 - x_1}{x_2 - x_1} = u \qquad \frac{x_4 - x_3}{x_2 - x_1} = v$$

と置く．次の極小値は x_1 と x_4 の間か，x_3 と x_2 の間である．どちらの範囲に極小値が入るかはわからないので，この範囲を等しくなるようにとることにする．したがって

$$u + v = 1 - u$$

とする．このような点の選び方を続けてきたとすれば，点 x_3 は実は前回の x_4 であるから

$$u = \frac{v}{1 - u}$$

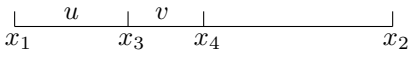
でなければならない．これら二つの式から

$$\begin{aligned} u^2 - 3u + 1 &= 0 \\ u &= \frac{3 - \sqrt{5}}{2} = 0.381966 \\ v &= \sqrt{5} - 2 = 0.236068 \end{aligned} \tag{3.48}$$

が得られる．すなわち x_3 は全区間を左から

$$0.382 : 0.618 = 1 : 1.618$$

の比で内分している．この比は黄金比として古くから知られた値である．x_3 の左側の区間の方が広いときには，x_3 から左側に v のところが x_4 になる．



以上をまとめると次のような手順になる．はじめに黄金比になっている三点 (x_1, x_3, x_2) で極小値を挟み込む．つぎに中点 x_3 の左右の広い方の区間を $0.382 : 0.618$ に内分する点を求める．これが x_4 である．どちらの方向に進むにしろ，x_3 に近い方が狭い区間になる．ここで $f(x_4)$ の値を計算する．もし右側に進んだときには，$f(x_4) > f(x_3)$ なら (x_1, x_3, x_4) が新しい極小値の範囲である．反対に $f(x_4) < f(x_3)$ なら (x_3, x_4, x_2) が新しい範囲である．x_4 が左側にきたときにも同様にして新しい範囲を決めることができる．

一回の反復で極小の範囲は 0.62 倍になる．これは二分法に比べてやや分が悪い．十分に狭い範囲に追い込んだ後は，2 次式などを用いた近似法を用いた方がよい．

> ギリシャ時代から黄金比 1：1.62 は縦横比として最も美しいとされてきたが，数学的には美しいが実用的にはどうかという問題もある．われわれが日常的に用いている紙の規格，A 版，B 版は半分に折ったときにも同じ縦横比になるように $1 : \sqrt{2}$ になっている．欧米の紙の規格がどうなっているかは知らないが，よく用いられているレターサイズは 1：1.29，そのほかにエクゼクティブは 1：1.44，レーガルは 1：1.65 でこれが黄金比に最も近い．

4 連立一次方程式

ガウスの消去法 特別な対称性のない連立一次方程式なら，ガウスの消去法による解法が最も効率が良く，また精度も高い．

理解しやすいようにはじめは三元連立方程式

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \quad \text{(a)}$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 \quad \text{(b)}$$

$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3 \quad \text{(c)}$$

を解くことを考える．まず (a)，(b) 式から x_1 を消去する．そのためには (a) 式に a_{21}/a_{11} を掛けて (b) 式から引けばよい．その結果

$$a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 = b_2^{(2)} \quad \text{(d)}$$

が得られる．ここに

$$\begin{aligned} l_{21} &= \frac{a_{21}}{a_{11}} \\ a_{22}^{(2)} &= a_{22} - l_{21}a_{12} \\ a_{23}^{(2)} &= a_{23} - l_{21}a_{13} \\ b_2^{(2)} &= b_2 - l_{21}b_1 \end{aligned}$$

である．つぎに (c) 式から x_1 を消去したい．そのためには (b) 式と (c) 式を利用してもよいが，ここでは再び (a) 式を利用する．すなわち，(a) 式に a_{31}/a_{11} を掛けて (c) 式から引く．

$$\begin{aligned} l_{31} &= \frac{a_{31}}{a_{11}} \\ a_{32}^{(2)} &= a_{32} - l_{31}a_{12} \\ a_{33}^{(2)} &= a_{33} - l_{31}a_{13} \\ b_3^{(2)} &= b_3 - l_{31}b_1 \end{aligned}$$

とすれば

$$a_{32}^{(2)}x_2 + a_{33}^{(2)}x_3 = b_3^{(2)} \quad \text{(e)}$$

が得られる．

このようにして x_1 を消去して得られるふたつの式 (d)，(e) からさらに x_2 を消去すれば

$$\begin{aligned} l_{32} &= \frac{a_{23}^{(2)}}{a_{22}^{(2)}} \\ a_{33}^{(3)} &= a_{33}^{(2)} - l_{32}^{(2)}a_{23}^{(2)} \\ b_3^{(3)} &= b_3^{(2)} - l_{32}^{(2)}b_2^{(2)} \end{aligned}$$

より

$$a_{33}^{(3)}x_3 = b_3^{(3)} \quad \text{(f)}$$

が得られる．これらをまとめれば

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \quad \text{(a)} \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 &= b_2^{(2)} \quad \text{(d)} \\ a_{33}^{(3)}x_3 &= b_3^{(3)} \quad \text{(f)} \end{aligned}$$

となる．以上が**前進消去**と呼ばれる過程である．

上式 (a)，(d)，(f) 式から未知数を求めるには最後の式から逆に解いていけばよい．

$$\begin{aligned} x_3 &= \frac{b_3^{(3)}}{a_{33}^{(3)}} \\ x_2 &= \frac{1}{a_{22}^{(2)}}\left(b_2^{(2)} - a_{23}^{(2)}x_3\right) \quad \text{(g)} \\ x_1 &= \frac{1}{a_{11}}\left(b_1 - a_{12}x_2 - a_{13}x_3\right) \end{aligned}$$

この過程を**後退代入**という．

このやり方を一般化するのは容易である．解くべき n 元連立方程式を

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = b_i \quad (4.1)$$
$$i = 1, 2, \cdots, n$$

とする．$x_1, x_2 \cdots$ を順次消去していくと，x_{k-1} まで消去した段階で，方程式は

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \cdots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + \cdots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\ \ddots \qquad\qquad \vdots \qquad\qquad & \\ a_{kk}^{(k)}x_k + \cdots + a_{kn}^{(k)}x_n &= b_k^{(k)} \\ a_{k+1k}^{(k)}x_k + \cdots + a_{k+1n}^{(k)}x_n &= b_{k+1}^{(k)} \\ \vdots \qquad\qquad \vdots & \\ a_{nk}^{(k)}x_k + \cdots + a_{nn}^{(k)}x_n &= b_n^{(k)} \end{aligned} \quad (4.2)$$

の形になる．ここで $a_{ij}^{(1)} = a_{ij}$，$b_i^{(1)} = b_i$ である．次のステップは k 行を用いて $k+1$ 行から n 行までの x_k を消去する過程である．これは公式

$$l_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \quad (4.3)$$

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik}a_{kj}^{(k)} \quad (4.4)$$

$$b_i^{(k+1)} = b_i^{(k)} - l_{ik}b_k^{(k)} \quad (4.5)$$

$$k < i \le j \le n \qquad 0 \le k < n$$

で計算する．添字の動き方を明確にするために，上の計算を Fortran 形式で書けば

```
do  k=1, n-1
  do  i=k+1, n
    lik=a(i,k)/a(k,k)
    do  j=k+1, n
      a(i,j)=a(i,j)-lik*a(k,J)
    enddo
    b(i)=b(i)-lik*b(k)
  enddo
enddo
```

となる．

消去を続行すると最後には次のような上三角方程式になる．

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \cdots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\ a_{33}^{(3)}x_3 + \cdots + a_{3n}^{(3)}x_n &= b_3^{(3)} \\ \ddots \qquad\qquad \vdots & \\ a_{nn}^{(n)}x_n &= b_n^{(n)} \end{aligned} \tag{4.6}$$

よって解は

$$\begin{aligned} x_n &= \frac{b_n^{(n)}}{a_{nn}^{(n)}} \\ x_i &= \frac{1}{a_{ii}^{(i)}}\Big(b_i^{(i)} - \sum_{j=i+1}^{n} a_{ij}^{(i)}x_j\Big) \qquad (4.7) \\ & \quad i = n-1,\, n-2,\, \cdots,\, 1 \end{aligned}$$

によって計算できる．

(4.4) 式の右辺に現れる $a_{ij}^{(k)}$ はこの計算が終われば二度と用いられることはない．したがって新たに計算された $a_{ij}^{(k+1)}$ をもと $a_{ij}^{(k)}$ があった場所に保存しても問題は起こらない．いいかえれば，一般に $a_{ij}^{(k)}$ はもとの係数行列 a_{ij} のあった場所に計算することができる．同様に (4.5) 式の $b_i^{(k)}$ は b_i の場所に計算することができる．(4.3) 式の右辺の $a_{ik}^{(k)}$ もこの計算の後は不要になるので，l_{ik} をこの場所に保存しておくことができる．要するに，もとの係数行列，右辺を保存しておく必要がないならば，すべての計算結果は a_{ij}，b_i の上に上書きすることによってメモリーを節約することができる．

係数行列が同じで，右辺だけが異なる方程式を何組も解かなければならないことも多い．このときには左辺の計算結果を残しておけば，右辺の異なるときの解は，(4.5) 式と (4.7) 式の計算だけで済むので非常に能率的である．特に，係数行列の逆行列を計算するときには右辺のある行だけが 1 でほかは 0 という方程式を n 回解くことになるが，このときには (4.4) 式は一回だけ，(4.5) 式と (4.7) 式を n 回解けばよい．

> アルゴリズムは数式で書くよりもプログラムで書いた方がわかりやすい．そのためにアルゴルなどという言語もあるが，この講義では Fortran という少し古いというか，数値計算用として最初に作られた言語を用いる．

消去法の計算量　消去法で一番計算量が多いのは (4.4) 式である．ある k に対してこの式は $(n-k)^2$ 個の新しい要素を求めるために計算される．この式の計算には一回の積と一回の差 (和) が含まれているが，これを計算量の単位とする．このように定義すると，(4.4) 式を $k=1$ から $n-1$ まで計算するのに必要な計算量は，n が非常に大きいとすれば

$$(4.4)\text{ 式} \qquad \sum_{k=1}^{n-1}(n-k)^2 \sim \frac{1}{3}n^3$$

になる．(4.3)，(4.5)，(4.7) 式の計算量は n が非常に大きいときにはすべて等しくなり

$$(4.3),\,(4.5),\,(4.7)\text{ 式} \qquad \frac{1}{2}n^2$$

になる．したがって (4.1) 式を一回だけ解くのに必要な計算量は

$$\frac{1}{3}n^3 + O(n^2) \tag{4.8}$$

である．

未知数の数 n が小さいときには計算量は問題にならないが，最近は $n=100$ や $n=1000$ の問題を解くことも多くなったので，計算量の少ないアルゴリズムを選択することが重要である．(4.8) 式の係数 1/3 が 2/3 になれば悪いアルゴリズムといわなければならない．計算回数が少ないということは丸め誤差が少ないということにも通じている．

枢軸要素の選択　(4.3) 式の要素 $a_{kk}^{(k)}$ は**枢軸要素** (**ピボット**) と呼ばれる．これはステップ k で分母に現れるので，これが 0 になればそれ以降の消去が不可

能になる．完全に 0 にならなくても絶対値が小さくなると丸め誤差が大きくなる．

そこで枢軸要素の絶対値ができるだけ大きくなるように，方程式の行や列を入れかえる方法がとられる．最も簡単なのは行を入れかえる方法で，ステップ k のとき k 列の要素

$$|a_{ik}^{(k)}| \quad k \le i \le n$$

が最大になる行 i を見つけて，行 i と行 k を入れかえる方法である．行を入れかえても解 x_i の順番は変わらないので，計算は簡単である．これを枢軸要素の**部分選択法**という．

ひとつの右辺に対する解だけを求めるのであれば部分選択法でどの行とどの行を入れかえたかの記録は必要ない．しかしいくつもの右辺に対する解を求めるときには，右辺を入れかえるためにステップ k でどの行との入れかえを行ったかの記録が必要になる．

もっと念入りにしたいときには

$$|a_{ij}^{(k)}| \quad k \le i, j \le n$$

が最大になる i, j を求め，i 行と k 行，j 列と k 列を入れかえる．このときには解の順番が入れかわるので，計算はやや複雑になる．これを**完全選択法**という．

なお，枢軸要素の積

$$\det[a_{ij}] = a_{11}^{(1)} \cdot a_{22}^{(2)} \cdots a_{kk}^{(k)} \cdots a_{nn}^{(n)} \qquad (4.9)$$

は係数行列の行列式の値である．部分選択法をとったときには行の入れかえが奇数回のときには上式の符号を反転させなければならない．

不定，不能方程式　行や列の入れかえを行っても枢軸要素が 0 になってしまうことがある．簡単な例として $n = 3$ で x_1 を消去したときに

$$a_{22}^{(2)} = a_{32}^{(2)} = 0$$

になってしまったとする．このときには第二，三行目の式は

$$a_{23}^{(2)} x_3 = b_2^{(2)} \qquad a_{33}^{(2)} x_3 = b_3^{(2)}$$

であるから，

$$\frac{b_2^{(2)}}{a_{23}^{(2)}} = \frac{b_3^{(2)}}{a_{33}^{(2)}} \qquad (4.10)$$

が成り立てば上式の比が x_3 にほかならない．しかし x_3 が決まっても，残るのは第一式だけであるから，x_1 と x_2 は独立には決まらない．x_1 と x_2 は x_1–x_2 平面上の，連立方程式の第一式から決まる直線上の任意の点である．すなわち (4.10) 式が成り立つときは方程式は**不定**である．

一方，(4.10) 式が成り立たないときには解は存在しない．すなわち方程式は**不能**である．これは方程式系に互いに矛盾する式が含まれていることを意味している．不定，不能どちらの場合にも，係数行列の行列式の値は 0，すなわち係数行列は特異行列になる．要するに係数行列が特異なときには解が一意的には得られない．不定になるか不能になるかは方程式の右辺によって決まる．

***LU* 分解**　ガウスの消去法で得られる係数 $a_{ij}^{(k)}$ を要素とする $n \times n$ の正方行列を $\boldsymbol{A}^{(k)}$ とする．$\boldsymbol{A}^{(k)}$ から $\boldsymbol{A}^{(k+1)}$ を求める計算は行列の積

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{M}_k \boldsymbol{A}^{(k)}$$

の形に表すことができる．$\boldsymbol{M}_k$ は $n \times n$ の正方行列である．$k = 1$ のときにはこの行列 $\boldsymbol{M}_1$ は

$$\boldsymbol{M}_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -l_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -l_{n1} & 0 & \cdots & 1 \end{bmatrix}$$

である．一般に $\boldsymbol{M}_k$ は単位行列の (k, k) 成分の下に列ベクトル $[-l_{k+1,k}, -l_{k+2,k}, \cdots, -l_{n,k}]^T$ を詰めたものである (^{T} は転置行列を表す)．ガウスの消去法ではこのような行列を $n-1$ 回掛けて，もとの行列 $\boldsymbol{A}$ を上三角行列 $\boldsymbol{U}$ に変換する操作

$$\boldsymbol{M}_{n-1} \cdots \boldsymbol{M}_2 \boldsymbol{M}_1 \boldsymbol{A} = \boldsymbol{U} \qquad (4.11)$$

であるといえる．ここに $\boldsymbol{U}$ は (4.6) 式の左辺の三角行列である．いま

$$\boldsymbol{L} = \boldsymbol{M}_1^{-1} \boldsymbol{M}_2^{-1} \cdots \boldsymbol{M}_{n-1}^{-1} \qquad (4.12)$$

と置けば，(4.11) 式は $\boldsymbol{A}$ が

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U} \qquad (4.13)$$

と表されることを意味している．これを行列 $\boldsymbol{A}$ の LU 分解という．

$\boldsymbol{L}$ を求めるためには $\boldsymbol{M}_k^{-1}$ を計算しなければならないが，これは簡単で $\boldsymbol{M}_k$ の要素である $-l_{i,k}$ の符号を正にすればよい．またこれらの積も簡単に計算できて

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ l_{21} & 1 & \cdots & \cdots & 0 \\ l_{31} & l_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & 1 \end{bmatrix} \tag{4.14}$$

となる．すなわち $\boldsymbol{L}$ は下三角行列である．ガウスの消去法は行列 $\boldsymbol{A}$ を (4.13) 式のように LU 分解して解を求める方法といえる．下三角行列の要素 l_{ij} は (4.3) 式で計算してある．なお行列の積の行列式は行列式の積

$$\det \boldsymbol{A} = \det \boldsymbol{L} \det \boldsymbol{U}$$

であるから，$\det \boldsymbol{L} = 1$ より $\boldsymbol{A}$ の行列式の値が (4.9) で表されることもわかる．

ガウス・ジョルダンの方法　ガウスの消去法の変形であるガウス・ジョルダンの方法は，一時期には最良の方法として推奨されていたが，計算量がガウス法の約二倍になるのであまり薦められない．しかし逆行列の計算には便利であるので簡単に触れておく．

前と同様に $n = 3$ の場合で考えてみる．まず一行目を a_{11} で割った式

$$x_1 + a_{12}x_2 + a_{13}x_3 = b_1$$

を作る．上付き添字の煩雑さを避けるために，割り算されたものも同じ記号で表している．この式に a_{21} を掛けて二行目から引き，a_{31} を掛けて三行目から引くと x_1 が消去できる．ここで新たに計算された二行目を枢軸要素 a_{22} で割ると，次のような形になる．

$$\begin{aligned} x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ x_2 + a_{23}x_3 &= b_2 \\ a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned}$$

この二行目に a_{32} を掛けて三行目から引くと，三行目から x_2 が消去できる．ここまではガウスの方法と基本的には同じである．ガウス・ジョルダン法が違うのは，ここで二行目に a_{12} を掛けて一行目からも x_2 を消去することである．最後に x_3 が残るが，これは三行目を正規化して，1 行目，二行目から消去する．最終的には

$$\begin{aligned} x_1 &= b_1 \\ x_2 &= b_2 \\ x_3 &= b_3 \end{aligned}$$

の形になって，左辺の係数行列は単位行列に，右辺は解になる．もし右辺に列ベクトルでなく単位行列を入れておけばそこに $\boldsymbol{A}$ の逆行列が得られることになる．

一般的な n 次元の場合の式を書くことはしないが，Fortran プログラムを示しておく．ここでは $\boldsymbol{A}$ の $n+1$ 列目に右辺 b_i が入っているとしている．また，ガウス・ジョルダン法ではもとの係数行列が破壊されてしまうので，不要になったところへ単位行列の列を入れて計算を進めている．結果として a(i,j) には逆行列が，a(i,n+1) には解が得られる．

```
det=1
do  k=1, n
  akk=a(k,k)
  det=det*akk
  a(k,k)=1
  do  j=1, n+1
    a(k,j)=a(k,j)/akk
  enddo
  do  i=1, n
    if( i.ne.k )  then
      aik=a(i,k)
      a(i,k)=0
      do  j=1, n+1
        a(i,j)=a(i,j)-aik*a(k,j)
      enddo
    endif
  enddo
enddo
```

なお，このプログラムではピボットの選択は行っていない．

クラウトの方法（LU 分解） ガウスの消去法は最終的には三角方程式 (4.6) を導くものであったが，今度は積極的に $\boldsymbol{A}$ が下三角行列と上三角行列の積

$$\boldsymbol{A} = \boldsymbol{L} \cdot \boldsymbol{U}$$

$$\boldsymbol{U} = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & & u_{nn} \end{bmatrix} \tag{4.15}$$

に分解できたとする．$\boldsymbol{L}$ は (4.14) 式と同じ形である．

$\boldsymbol{L}$ と $\boldsymbol{U}$ を求めるために $\boldsymbol{L}\cdot\boldsymbol{U} = \boldsymbol{A}$ の両辺の (1,1) 要素を比較すれば

$$u_{11} = a_{11}$$

が求められる．これが求められれば両辺の一列目を比較することによって

$$l_{i1} = \frac{a_{i1}}{u_{11}} \qquad i = 2, 3, \cdots, n$$

が得られる．これで $\boldsymbol{L}$ と $\boldsymbol{U}$ の一列目が求められたことになる．次に (1,2) 成分と (2,2) 成分から

$$u_{12} = a_{12} \qquad u_{22} = a_{22} - l_{21}u_{12}$$

が成り立つ．第二式の l_{21} は既に求められているから，これで u_{12}，u_{22} が決まった．これらを用いて二列目の残りの要素から

$$l_{i2} = \frac{1}{u_{22}}\Big(a_{i2} - l_{i1}u_{12}\Big) \qquad i = 3, 4, \cdots, n$$

が求まる．これで二列目まで求められた．

これを一般化するのは容易である．$j = 1, 2, \cdots, n$ に対して次の二組の計算を行なう．

$$l_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj} \qquad i = 1, 2, \cdots, j \tag{4.16}$$

$$l_{ij} = \frac{1}{u_{jj}}\Big(a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj}\Big) \qquad i = j{+}1, j{+}2, \cdots, n \tag{4.17}$$

この方法を列ごとのクラウトの方法という．単に LU 分解するだけなら行と列を交互に計算していく方法もあるが，上の列ごとの計算法は行の入れかえに便利だからである．

> 学生時代にこの方法を習ったときに，これは玄人 (クロート) の方法だからよく覚えておくように，といわれた記憶がある．

三重対角方程式 係数行列が対角線とその上下だけが 0 でない方程式は，三重対角方程式と呼ばれ，いろいろな場面によく現れる．これは次のように書くことができる．

$$\begin{aligned} &\gamma_i x_{i-1} + \alpha_i x_i + \beta_i x_{i+1} = b_i \\ &\qquad i = 1, 2, 3, \cdots, n \\ &\qquad \gamma_1 = \beta_n = 0 \end{aligned} \tag{4.18}$$

α_i, β_i, γ_i が係数 a_{ij} に相当するものである．この方程式にガウスの消去法を施すと

$$a_i x_i + \beta_i x_{i+1} = c_i \qquad i = 1, 2, \cdots, n \tag{4.19}$$

が得られる．a_i, c_i は次式で計算される．

$$\begin{aligned} &a_1 = \alpha_1 \qquad c_1 = \beta_1 \\ &a_i = \alpha_i - \frac{\gamma_i}{a_{i-1}}\beta_{i-1} \\ &c_i = b_i - \frac{\gamma_i}{a_{i-1}}c_{i-1} \\ &\qquad i = 2, 3, \cdots, n \end{aligned} \tag{4.20}$$

a_i, c_i がわかると (4.12) 式から x_i が

$$\begin{aligned} x_i = \frac{1}{a_i}(c_i - \beta_i x_{i+1}) \\ i = n, n{-}1, \cdots, 1 \end{aligned} \tag{4.21}$$

によって計算される．

三重対角行列の場合，ピボットの部分選択の対象となる行は k 行と $k+1$ 行だけであるから計算は簡単である．しかし三重対角方程式が表れるような問題のほとんどの場合，係数行列は優対角の条件，すなわち

$$|\alpha_i| > |\beta_i| + |\gamma_i| \tag{4.22}$$

が成り立っているので，ピボットの選択をしなくても正確な解が求められることが多い．

逆行列 (4.1) 式の係数から作られる $n \times n$ の正方行列を $\boldsymbol{A} = [a_{ij}]$，右辺 b_i を要素とする列ベクトルを $\boldsymbol{b} = [b_1, b_2, \cdots, b_n]^T$ (^{T} は転置) 解ベクトルを $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]^T$ とすると (4.1) 式は

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{4.23}$$

と書くことができる．この方程式の形式的な解は，逆行列を $\boldsymbol{A}^{-1}$ とすると

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} \tag{4.24}$$

と書くことができる．表現が簡単であるからこの書き方はよく用いられるが，これは逆行列を作ってから $\boldsymbol{b}$ との積を計算することを意味しているわけではない．むしろ，逆行列を掛けて解を求めると丸め誤差が大きくなるので，逆行列を用いることはできるだけ避けた方がよい．逆行列の要素そのものが必要な場合を除いては逆行列を用いるべきではない．

ベクトルのノルム　ベクトル $\boldsymbol{x}$ のノルム $\|\boldsymbol{x}\|$ は次のような条件を満たす．

$$\begin{aligned} &\|\boldsymbol{x}\| \geq 0 \\ &\|\boldsymbol{x}\| = 0 \quad \text{なら} \quad \boldsymbol{x} = \boldsymbol{0} \\ &\|\alpha\boldsymbol{x}\| = |\alpha|\|\boldsymbol{x}\| \\ &\|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\| \end{aligned} \tag{4.25}$$

α はスカラーである．これらの条件を満足するノルムは実は一つではない．ベクトルで最もよく用いられるのは L_2 ノルムで

$$\|\boldsymbol{x}\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2} \qquad L_2 \text{ ノルム} \tag{4.26}$$

である．そのほかに

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i| \qquad L_1 \text{ ノルム} \tag{4.27}$$

$$\|\boldsymbol{x}\|_\infty = \max_{1\leq i\leq n} |x_i| \qquad L_\infty \text{ ノルム} \tag{4.28}$$

もよく用いられる．

行列のノルム　ベクトルのノルムに比べて行列のノルムはなじみが薄い．行列のノルムは

$$\|\boldsymbol{A}\| = \sup_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \tag{4.29}$$

で定義される．このように定義されたノルムは，(4.25) 式のベクトルを行列で置き換えた式を満たしている．さらに

$$\|\boldsymbol{A}\boldsymbol{B}\| \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\| \tag{4.30}$$

$$\|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{x}\| \tag{4.31}$$

を満たしている．(4.31) 式は定義 (4.29) 式と同値であるが，等号が成り立つ $\boldsymbol{x}$ が存在する，すなわち

$$\begin{aligned} &\|\boldsymbol{A}\boldsymbol{x}\| = \|\boldsymbol{A}\| \cdot \|\boldsymbol{x}\| \quad \text{となる} \\ &\|\boldsymbol{x}\| \neq 0 \text{ が存在する} \end{aligned} \tag{4.32}$$

が成り立つ．

定義式 (4.29) 式の右辺に現れるベクトルのノルムとしてなにを選ぶかによって，行列のノルムにもいろいろなものが考えられる．特に，L_1，L_∞ ノルムは行列要素を用いて

$$\begin{aligned} \|\boldsymbol{A}\|_1 &= \max_j \sum_i |a_{ij}| \\ \|\boldsymbol{A}\|_\infty &= \max_i \sum_j |a_{ij}| \end{aligned} \tag{4.33}$$

と表すことができる．

L_2 ノルムを L_1，L_∞ ノルムのように行列の要素を用いて表すことは困難である．しかし L_2 ノルムについては次のような性質がある．行列 $\boldsymbol{A}$ の固有値を λ_i とするとき

$$\rho(\boldsymbol{A}) = \max_i |\lambda_i| \tag{4.34}$$

を行列 $\boldsymbol{A}$ のスペクトル半径という．$\boldsymbol{A}$ が実対称行列のときには L_2 ノルムは

$$\|\boldsymbol{A}\|_2 = \rho(\boldsymbol{A}) \tag{4.35}$$

で与えられる．

条件数　物理的な問題で連立方程式

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

を解くときには，係数 $\boldsymbol{A}$ や $\boldsymbol{b}$ に誤差が含まれている．これらの誤差によって解がどれだけ変化するかを知っておくことは重要である．

右辺 $\boldsymbol{b}$ に $\delta\boldsymbol{b}$ の誤差が含まれているときの解を $\boldsymbol{x}+\delta\boldsymbol{x}$ とすれば

$$\boldsymbol{A}(\boldsymbol{x}+\delta\boldsymbol{x}) = \boldsymbol{b} + \delta\boldsymbol{b} \qquad \delta\boldsymbol{x} = \boldsymbol{A}^{-1}\delta b$$

であるから，ノルムの性質を用いれば

$$\begin{aligned} &\|\boldsymbol{b}\| = \|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{x}\| \\ &\|\delta\boldsymbol{x}\| = \|\boldsymbol{A}^{-1}\delta\boldsymbol{b}\| \leq \|\boldsymbol{A}^{-1}\| \cdot \|\delta\boldsymbol{b}\| \end{aligned}$$

が成り立つから

$$\frac{\|\delta \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \kappa(\boldsymbol{A}) \frac{\|\delta \boldsymbol{b}\|}{\|\boldsymbol{b}\|} \tag{4.36}$$

$$\kappa(\boldsymbol{A}) = \|\boldsymbol{A}\| \cdot \|\boldsymbol{A}^{-1}\| \tag{4.37}$$

が得られた．すなわち，解の相対誤差は右辺の相対誤差の $\kappa(\boldsymbol{A})$ 倍になる．$\kappa(\boldsymbol{A})$ を行列 $\boldsymbol{A}$ の**条件数**という．κ が大きいほど方程式 (4.1) は条件が悪い．すなわち，右辺の僅かな変化によって解が大きく変化する．

前とは反対に係数行列に誤差 $\delta\boldsymbol{A}$ が含まれているときの解を $\boldsymbol{x} + \delta\boldsymbol{x}$ とすれば

$$(\boldsymbol{A} + \delta\boldsymbol{A})(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{b}$$

$$\delta\boldsymbol{x} = -\boldsymbol{A}^{-1}\delta\boldsymbol{A}(\boldsymbol{x} + \delta\boldsymbol{x})$$

が成り立たなければならないから，先と同じような計算を行えば

$$\frac{\|\delta \boldsymbol{x}\|}{\|\boldsymbol{x} + \delta\boldsymbol{x}\|} \leq \kappa(\boldsymbol{A}) \frac{\|\delta \boldsymbol{A}\|}{\|\boldsymbol{A}\|} \tag{4.38}$$

が得られる．この式も係数行列の相対誤差 $\|\delta\boldsymbol{A}\|/\|\boldsymbol{A}\|$ が小さくても，条件数が大きければ解の誤差が大きくなることを示している．

$\boldsymbol{A}$ が実対称行列のとき，行列のノルムとして L_2 を用いることにすれば，$\boldsymbol{A}^{-1}$ の固有値は $\boldsymbol{A}$ の固有値の逆数であるから，(4.35) 式より

$$\kappa_2(\boldsymbol{A}) = \frac{\max_i |\lambda_i|}{\min_i |\lambda_i|} \tag{4.39}$$

で表される．$\boldsymbol{A}$ が特異行列であれば条件数は無限大になる．$\boldsymbol{A}$ が対称行列でないときにはこの関係は成り立たず，絶対値最大と最小の固有値の比が小さくても条件数が大きくなることがあり得る．一般の行列の L_2 条件数は特異値分解のところで述べる．

条件数の推定　条件数を求めるには $\boldsymbol{A}$ と $\boldsymbol{A}^{-1}$ のノルムが必要になる．いま L_1 ノルムを用いることにすると，$\|\boldsymbol{A}\|_1$ は (4.33) 式の第一式から簡単に求めることができる．問題は $\|\boldsymbol{A}^{-1}\|_1$ である．逆行列 $\boldsymbol{A}^{-1}$ を計算してしまえば問題はないが，大きな行列のときには大変だし，条件数は正確に求めても意味がないので簡便法をとる．

はじめに (4.33) 式から任意の $\boldsymbol{A}$ に対して $\|\boldsymbol{A}\|_1 = \|\boldsymbol{A}^T\|_\infty$ が成り立つことを注意しておく．したがって (3.29) 式から

$$\|\boldsymbol{A}^{-1}\|_1 = \|\boldsymbol{A}^{-T}\|_\infty = \sup \frac{\|\boldsymbol{A}^{-T}\boldsymbol{e}\|_\infty}{\|\boldsymbol{e}\|_\infty}$$

である．$\boldsymbol{A}^{-1}$ の転置行列を簡単に $\boldsymbol{A}^{-T}$ と書いてある．いま

$$\boldsymbol{A}^{-T}\boldsymbol{e} = \boldsymbol{v}$$

と置くと，$\|\boldsymbol{A}^{-1}\|_1$ は

$$\|\boldsymbol{e}\|_\infty = \max_i |e_i| = 1 \quad \text{の条件下で}$$
$$\|\boldsymbol{v}\|_\infty = \max_i |v_i| \quad \text{を最大にする}$$

という問題に書きかえられる．

ガウスの消去法などによって $\boldsymbol{A}$ が LU 分解されているとき，$\boldsymbol{v}$ は

$$\boldsymbol{U}^T\boldsymbol{z} = \boldsymbol{e} \qquad \boldsymbol{L}^T\boldsymbol{v} = \boldsymbol{z} \tag{4.40}$$

から計算される．$\boldsymbol{U}^T$ は下三角行列であるから $\boldsymbol{z}$ の成分は上から求められ，この $\boldsymbol{z}$ を用いて第二式から後退代入によって $\boldsymbol{v}$ が求められる．ノルムの定義は許されるすべての $\boldsymbol{e}$ に関する $\|\boldsymbol{v}\|_\infty$ の最大値であるが，そのような計算はできないので次のような便法をとる．

$\boldsymbol{e}$ の成分の絶対値の最大値は 1 でなければならないので，ここではすべての成分の絶対値を 1 にとり

$$\boldsymbol{e}^T = [1, \pm 1, \pm 1, \cdots]$$

とする．第一成分だけを 1 にしたのは便宜上で，第二成分以下は以下のように決める．(4.40) の第一式は

$$\begin{aligned} z_1 &= \frac{e_1}{u_{11}} = \frac{1}{u_{11}} \\ z_k &= \frac{1}{u_{kk}} \left(e_k - \sum_{i=1}^{k-1} u_{ik} z_i \right) \\ & \quad k = 2, 3, \cdots, n \end{aligned} \tag{4.41}$$

と書けるが，ここで $e_k = \pm 1$ の符号を z_k の絶対値が大きくなるように選ぶ．すなわち，符号関数を $\mathrm{sign}(x)$ として

$$e_k = \mathrm{sign}\left(-\sum_{i=1}^{k-1} u_{ik} v_i \right)$$

とする．こうして決められた $\boldsymbol{z}$ を用いて (4.40) の第二式から

$$v_n = z_n$$

$$v_k = z_k - \sum_{i=k+1}^{n} l_{ik} z_i \qquad (4.42)$$
$$k = n-1, n-2, \cdots, 1$$

によって $\boldsymbol{v}$ を求めれば

$$\|\boldsymbol{A}^{-1}\|_1 \doteq \max_i |v_i| \qquad (4.43)$$
$$\kappa_1(\boldsymbol{A}) = \|\boldsymbol{A}\|_1 \cdot \|\boldsymbol{A}^{-1}\|_1$$

である．このやり方では $\boldsymbol{v}$ を大きくしているわけではないが，数値実験の結果では $\|\boldsymbol{A}^{-1}\|_1$ の良い近似を与える．

例として対称行列

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 2 & 3 & 4 & 5 & 6 \\ 3 & 3 & 3 & 4 & 5 & 6 \\ 4 & 4 & 4 & 4 & 5 & 6 \\ 5 & 5 & 5 & 5 & 5 & 6 \\ 6 & 6 & 6 & 6 & 6 & 6 \end{bmatrix}$$

について上の方法で求めた L_1 ノルムに基づいた条件数の推定値は

$$\kappa_1 \sim 144$$

であった．この行列の絶対値最大の固有値は 27.72, 絶対値最小の固有値は -0.2682 であるから，L_2 ノルムに基づいた条件数は

$$\kappa_2 = 103$$

である．ノルムの定義の違いを考えれば，ここで述べた簡便法は条件数の推定には十分な確度を持っていると考えてよいだろう．

反復改良法　方程式のサイズが大きいときには，係数行列 $\boldsymbol{A}$ の条件数がそれほど大きくなくても，丸め誤差の累積のために正しい解が求められないことが多い．連立方程式

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

を解いたときの解 $\boldsymbol{x}$ は正確に上式を満たしておらず，誤差 $\delta\boldsymbol{x}$ を含んでいる．$\boldsymbol{x} + \delta\boldsymbol{x}$ がもとの方程式を満たすためには

$$\boldsymbol{A}(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{b}$$

を満足しなければならない．したがって $\delta\boldsymbol{x}$ は

$$\boldsymbol{A}\delta\boldsymbol{x} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} = \boldsymbol{r} \qquad (4.44)$$

を解くことによって求められる．右辺に現れる $\boldsymbol{x}$ はもとの方程式を解いて得られた誤差を含んだ解である．なお，残差 $\boldsymbol{r}$ の計算は倍精度で行わなければならない．$\boldsymbol{x} + \delta\boldsymbol{x}$ が改良された解である．この解を用いてさらに改良を加えることもできるが，通常は一回の反復で十分である．

(4.43) 式をガウスの方法で解くためには，方程式のサイズを n とすれば，$n^3/3$ の計算量が必要である．しかし，最初に $\boldsymbol{x}$ を解いたときに求まる LU 分解の表を残しておけば，(4.43) 式の解は右辺の計算と後退代入だけによって求めることができる．この計算量は n^2 のオーダーであるから，反復改良の手間はとるにたらない．

もともと $\boldsymbol{A}$ の条件数が非常に大きいときには，反復改良を行っても効果はない．また，$n < 10$ 程度の小さなサイズの方程式のときには，条件数相応の解が求まってしまうので，この場合にも反復改良の効果はほとんどない．効果があるのは n は大きいが，条件数がそこそこの大きさの方程式に対してである．

5 内挿と関数近似

ラグランジュの補間公式 $x_0, x_1, \cdots, x_n$ における値が $f(x)$ の値に一致するような近似式 $f_n(x)$ のことを $f(x)$ の補間式という．すなわち

$$f_n(x_k) = f(x_k) \qquad k = 0,\, 1,\, \cdots,\, n \quad (5.1)$$

を満足する関数 $f_n(x)$ である．

このような関数を x の多項式で作るのは容易である．既に §3 で用いた式を一般化すれば次のようになる．まず

$$\begin{aligned}\Pi_n(x) &= (x-x_0)(x-x_1)\cdots(x-x_n)\\ &= \prod_{i=0}^{n}(x-x_i)\end{aligned}$$

と置けば

$$f_n(x) = \sum_{k=0}^{n} \frac{\Pi_n(x)}{(x-x_k)\Pi_n'(x_k)} f(x_k) \qquad (5.2)$$

が条件 (5.1) 式を満たしていることは明らかである．この式をラグランジュの補間公式という．

ニュートンの補間公式 ラグランジュの公式 (5.2) をそのまま用いるのは計算に不便である．(5.2) 式の $f_n(x)$ は x の n 次式であるから

$$\begin{aligned}f_n(x) &= a_0 + a_1(x-x_0)\\ &\quad + a_2(x-x_0)(x-x_1) + \cdots\\ &\quad + a_n(x-x_0)(x-x_1)\cdots(x-x_n)\end{aligned}$$

の形に展開することができる．明らかに

$$f_n(x_0) = a_0$$

である．また

$$f_n[x_0, x] \equiv \frac{f_n(x) - f_n(x_0)}{x - x_0}$$

と定義すれば

$$\begin{aligned}f_n[x_0, x] &= a_1 + a_2(x-x_1) + \cdots\\ &\quad + a_n(x-x_1)\cdots(x-x_n)\end{aligned}$$

であるから

$$f_n[x_0, x_1] = a_1$$

が成り立つ．一般に

$$\begin{aligned}&f[x] = f(x)\\ &f[x_0, x_1, \cdots, x_k, x] =\\ &\quad \frac{f[x_0, x_1, \cdots, x_{k-1}, x] - f[x_0, x_1, \cdots, x_{k-1}, x_k]}{x - x_k}\end{aligned} \qquad (5.3)$$

と定義すれば

$$a_k = f[x_0, x_1, \cdots, x_k]$$

が成り立つから

$$\begin{aligned}f_n(x) &= f[x_0] + f[x_0, x_1](x-x_0)\\ &\quad + f[x_0, x_1, x_2](x-x_0)(x-x_1) + \cdots\\ &\quad + f[x_0, x_1, \cdots, x_n](x-x_0)(x-x_1)\cdots(x-x_n)\end{aligned} \qquad (5.4)$$

が得られる．

エイトケンの補間公式 上式を計算するのは面倒であるが，次のようにすれば (5.3) 式と (5.4) 式を同時に機械的に計算を進めることができる．

$$I_{0i}(x) = \frac{1}{x_i - x_0} \begin{vmatrix} f(x_0) & x_0 - x \\ f(x_i) & x_i - x \end{vmatrix}$$
$$i = 1, 2, \cdots, n$$
$$I_{01i}(x) = \frac{1}{x_i - x_1} \begin{vmatrix} I_{01}(x) & x_1 - x \\ I_{0i}(x) & x_i - x \end{vmatrix}$$
$$i = 2, 3, \cdots, n$$

$$\begin{aligned}I_{012\cdots k-1i}(x) &= \frac{1}{x_i - x_{k-1}}\\ &\quad \times \begin{vmatrix} I_{012\cdots k-1}(x) & x_{k-1} - x \\ I_{012\cdots k-2i}(x) & x_i - x \end{vmatrix}\end{aligned} \qquad (5.5)$$
$$i = k, k+1, \cdots, n$$

上式の計算は数行のプログラムですむ．配列 x(i) に座標 x_i が，配列 y(i) に対応する関数値 $f(x_i)$ が与えられているとき，$x = x_c$ における内挿値 y_c を求めるプログラムは次のようになる．

```
real  x(0:n), y(0:n), f(0:n)
do  i=0, n
  f(i)=y(i)
enddo
do  k=0, n-1
```

```
      do  i=k+1, n
        f(i)=f(k)+(f(k)-f(i))
     *         *((x(k)-xc)/(x(i)-x(k)))
      enddo
    enddo
    yc=f(n)
```

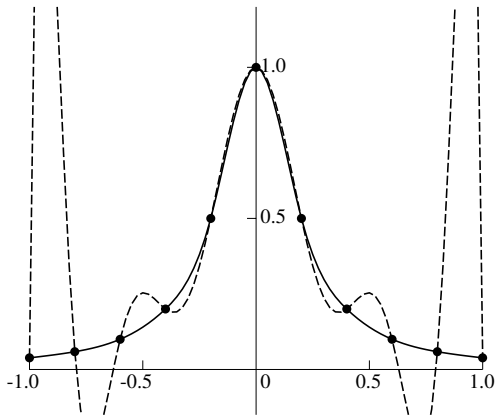
f(i) は作業用である．

ラグランジュの補間公式 (5.2)，ニュートンの補間公式 (5.4)，エイトキンの補間公式 (5.5) はすべて x の n 次多項式であるから，表現は異なっても等価なものである．いずれの多項式も点 $x_0, x_1, \cdots, x_n$ の順番にはよらない．もちろん同じ点があってはならない．これらの補間公式では分点と分点の間の値についてはなんの制約も置いていないので，ときに激しく振動することがある．

下に示す図はよく用いられる例題で，関数

$$f(x) = \frac{1}{1+25x^2}$$

を $x = -1.0, -0.8, \cdots, 0.8, 1.0$ の 11 点で与えて，上のプログラムを用いてグラフを描いたものである．黒丸が与えた点，破線が補間値である．補間値は与えられた点を通ってはいるが，それ以外では極端に振動している．これは全体を 10 次の多項式で近似しようとしているからである．



スプライン補間　前項のように全体を一つの多項式で近似するのではなく，区分的に低次の多項式で近似する方法もある．ここでは分点が $x_0 < x_1 < \cdots < x_n$ の順に並んでいるものとする．よく用いられるのは三次の自然スプライン関数である．これは分点で二次の微係数までが連続になるように係数が決められている．いま，区間 $x_{j-1} \le x \le x_j$ で三次式

$$\begin{aligned} S_j(x) &= \frac{h_j^2}{6} M_{j-1} \left[\left(\frac{x_j - x}{h_j} \right)^3 - \left(\frac{x_j - x}{h_j} \right) \right] \\ &\quad + \frac{h_j^2}{6} M_j \left[\left(\frac{x - x_{j-1}}{h_j} \right)^3 - \left(\frac{x - x_{j-1}}{h_j} \right) \right] \\ &\quad + \frac{x_j - x}{h_j} f_{j-1} + \frac{x - x_{j-1}}{h_j} f_j \quad 1 \le j \le n \end{aligned} \tag{5.6}$$

$$x_{j-1} < x < x_j \qquad h_j = x_j - x_{j-1} \qquad f_j = f(x_j)$$

を定義する．これが与えられた点を通る，すなわち

$$S_j(x_j) = f_j \qquad S_j(x_{j-1}) = f_{j-1}$$

が成り立っていることは明らかである．一階微分，二階微分を計算すると

$$\begin{aligned} S_j'(x) &= -\frac{h_j}{6} M_{j-1} \left[3 \left(\frac{x_j - x}{h_j} \right)^2 + 1 \right] \\ &\quad + \frac{h_j}{6} M_j \left[3 \left(\frac{x - x_{j-1}}{h_j} \right)^2 - 1 \right] + \frac{f_j - f_{j-1}}{h_j} \\ S_j''(x) &= M_{j-1} \frac{x_j - x}{h_j} + M_j \frac{x - x_{j-1}}{h_j} \end{aligned}$$

となる．最後の式から M_j は分点 x_j における二次の微係数を意味していることがわかる．分点で関数値，二次微分が連続という条件は満足されているから，残るのは一階微分の連続性で，分点 x_j における連続条件

$$S_j'(x_j) = S_{j+1}'(x_j)$$

から

$$\begin{aligned} &\frac{h_j}{6} M_{j-1} + \frac{h_j + h_{j+1}}{3} M_j + \frac{h_{j+1}}{6} M_{j+1} \\ &\qquad = \frac{f_{j+1} - f_j}{h_{j+1}} - \frac{f_j - f_{j-1}}{h_j} \\ &\qquad M_0 = M_n = 0 \qquad j = 1, 2, \cdots, n-1 \end{aligned} \tag{5.7}$$

が導かれる．この式は M_1 から M_{n-1} を未知数とする三重対角方程式にほかならない．しかも明らかに優対角であるから，§4 で述べた方法で簡単に解くことができる．

方程式 (5.7) は近似曲線の曲率の二乗積分が最小という条件からも導くことができる．したがってスプライン補間では分点間で内挿値が暴れるという心

配はあまりない．なお，上では端点で二次微係数が 0 と仮定したが，その他の条件，たとえば端点で一次微係数が与えられたとき，周期条件なども考えられる．

先の図の実線は (5.7) 式を解いて得られた M_j を用いて (5.6) 式を用いて内挿したものである．この図のスケールでは真の値 $f(x)$ との差はわからない．

実関数の内積とノルム　区間 $[a,b]$ で定義された実関数 $u(x)$，$v(x)$ の内積と L_2 ノルムはそれぞれ

$$(u,v)=\int_a^b u(x)v(x)dx \tag{5.8}$$

$$\|u\|_2^2=(u,u) \tag{5.9}$$

で定義される．$\|u\|_2$ はノルムであるから (4.25) 式の関係などを満たしている．以下では L_2 ノルムだけを用いることにし，下の添字 2 は省略する．

最小二乗近似　これからは関数を近似することを考える．そのために補間公式を用いてもいいが，補間公式では分点で与えられた関数値に等しくなるという強い制約条件があるために，逆に分点間では補間値が暴れるという欠点がある．そこで近似区間で平均的に近似するために，誤差の二乗平均が最小になるような近似式を求める．

関数 $f(x)$ を区間 $[a,b]$ で近似するものとし，この区間で独立な基底関数 $\varphi_i(x)$ を用いて n 次の近似部分和として

$$f_n(x)=\sum_{i=0}^{n} a_i\varphi_i(x) \tag{5.10}$$

を用いる．この近似関数の誤差の二乗積分

$$S_n=\int_a^b |f(x)-\sum_{i=0}^{n} a_i\varphi_i(x)|^2dx$$

が a_k について最小になるためには

$$\int_a^b [f(x)-\sum_{i=1}^{n} a_i\varphi_i(x)]\varphi_k(x)dx=0$$

が成り立たなければならない．したがって，係数 a_i を決める式として

$$\sum_{i=0}^{n} a_i(\varphi_i,\,\varphi_k)=(f,\,\varphi_k) \tag{5.11}$$

$$k=0,1,\cdots,n$$

が得られた．これは a_i に関する $n+1$ 元の連立一次方程式である．a_i がこの式を満足したときの誤差の二乗積分は

$$S_n=\|f\|^2-\sum_{i=0}^{n} a_i(f,\varphi_i)=\|f\|^2-\|f_n\|^2$$

で表される．

内積が計算できればこの方法は明快であるが，問題がある．一つには，$\varphi_i(x)$ として勝手なものを選ぶと，連立方程式 (5.11) の条件数が非常に大きくなることが多いということである．$\varphi_i(x)$ としてすぐに思いつくのは冪 x^i であるが，このときには (5.11) 式の条件は非常に悪くなる．二つには，係数 a_i が最大次数 n が変わると変わってしまうことである．たとえば $n=10$ の計算を行なった後，$n=11$ の計算をするときには (5.11) 式を改めて解きなおさなければならず，過去の計算が無駄になってしまうからである．

フーリエ展開　上で述べた欠点を避けるために基底として

$$(\varphi_i,\,\varphi_j)=\delta_{ij}=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases} \tag{5.12}$$

を満たす $\varphi_i(x)$ を考える．これを正規直交系という．δ_{ij} はクロネッカーの記号である．

いま，実関数 $f(x)$ が与えられたとき，正規直交系を用いて (5.10) 式で近似関数を定義すると，(5.11) 式の左辺は $i=k$ の項だけが残ることになる．したがって

$$a_i=(f,\,\varphi_i) \tag{5.13}$$

となり，これは n によらない．また誤差は

$$S_n=\|f\|^2-\sum_{i=0}^{n} a_i^2$$

と表されるから，n が増えれば S_n は減少する (増加しない)．したがって $n\to\infty$ では

$$\lim_{n\to\infty}\|f-f_n\|=0 \tag{5.14}$$

が成り立つ．そこで形式的に

$$f(x)\sim\sum_{i=0}^{\infty}(f,\varphi_i)\varphi_i(x) \tag{5.15}$$

と書き，これを $f(x)$ のフーリエ展開という．(f, φ_i) はフーリエ係数である．形式的にという意味は，(5.15) の右辺が x の各点ごとに $f(x)$ に収束するわけではないからである．

通常のフーリエ展開　基底関数 φ_i として三角関数 $\cos kx, \sin kx$ を選べば，$[-\pi, \pi]$ で定義された実関数 $f(x)$ に対するフーリエ展開は

$$f(x) \sim \frac{1}{2}a_0 + \sum_{k=1}^{\infty}(a_k \cos kx + b_k \sin kx) \qquad (5.16)$$

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kx dx$$

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin kx dx$$

である．係数 $1/\pi$ が付いているのは，$\cos kx, \sin kx$ が正規化されていないからである．

$f(x)$ は $|x| \le \pi$ だけで定義されているが，基底関数が周期関数であるから，部分和 $f_n(x)$ は必然的に周期関数 $f_n(x+\pi) = f_n(x)$ になる．

ギブスの振動　部分和 $f_n(x)$ は定義域のすべての x で一様に $f(x)$ に収束するわけではない．例として不連続な関数

$$f(x) = \begin{cases} 1 & 0 < x \le \pi \\ 1/2 & x = 0 \\ 0 & -\pi \le x < 0 \end{cases} \qquad (5.17)$$

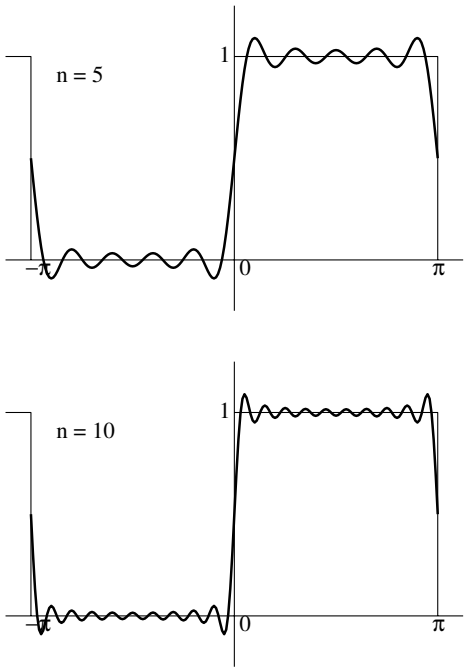
をとりあげる．この関数のフーリエ係数は

$$a_0 = 1 \qquad b_{2k-1} = \frac{2}{(2k-1)\pi}$$

になる．a_0 以外の a_k，偶数次の b_k は 0 である．したがって部分和は

$$f_{2n-1}(x) = \frac{1}{2} + \frac{2}{\pi}\sum_{k=1}^{n}\frac{\sin(2k-1)x}{2k-1}$$

である．$n = 5, 10$ のときのグラフを下に示す．



上式のままでは収束の様子がわかりにくいので，少し持って回ったやりかたをする．部分和は b_k の積分表現を用いて

$$f_n(x) = \frac{1}{2} + \sum_{k=1}^{n}\left[\frac{1}{\pi}\int_0^{\pi} f(y)\sin kydy\right]\sin kx$$

$$= \frac{1}{2} + \frac{1}{2\pi}\int_0^{\pi}\sum_{k=1}^{n}[\cos k(x-y) - \cos k(x+y)]dy$$

と書くことができる．両辺を x で微分する．積分の中の x 微分を y 微分で書き直すと

$$\frac{df_n(x)}{dx} = -\frac{1}{2\pi}\int_0^{\pi}\sum_{k=1}^{n}\frac{d}{dy}[\cos k(x-y) + \cos k(x+y)]dy$$

であるから容易に積分でき，その結果は

$$\frac{df_n(x)}{dx} = \frac{1}{\pi}\sum_{k=1}^{n}[\cos kx(1-\cos k\pi)]$$

となる．当然のことながら，奇数次の項だけの和になる．そこで改めて $df_{2n-1}(x)/dx$ を求めれば

$$\frac{df_{2n-1}(x)}{dx} = \frac{2}{\pi}\sum_{k=1}^{n}\cos(2k-1)x$$

$$= \frac{\sin 2nx}{\pi \sin x} \qquad (5.18)$$

が得られる．

$df_{2n-1}(x)/dx$ が 0 になる点は

$$x = \frac{m\pi}{2n} \qquad m = \pm 1, \pm 2, \cdots, 2n-1$$

である．ここで $f_{2n-1}(x)$ は極大，あるいは極小値をとる．極大，極小点の数は n が増えるにつれて増加するが，同時に極大値，極小値が順調に正解に近づくとは限らない．図からわかるように，$x = \pi/2n$ の極値は n が 2 倍になるとかえって高くなっているようにすら見える．

$f_{2n-1}(x)$ の値を見積もるには (5.18) 式を積分すればよい．特に，$x = \pi/2n$ における値は

$$\begin{aligned} f_{2n-1}(\pi/2n) &= \frac{1}{2} + \frac{1}{\pi}\int_0^{\pi/2n} \frac{\sin 2ny}{\sin y} dy \\ &= \frac{1}{2} + \frac{1}{2n\pi}\int_0^{\pi} \frac{\sin y}{\sin(y/2n)} dy \end{aligned}$$

である．n が大きいとすれば，積分は

$$\frac{1}{2n\pi}\int_0^{\pi} \frac{\sin y}{\sin(y/2n)} dy \sim \frac{1}{\pi}\int_0^{\pi} \frac{\sin y}{y} dy$$

で近似することができる．最後の積分

$$\mathrm{Si}(x) = \int_0^x \frac{\sin y}{y} dy$$

は積分正弦関数と呼ばれ，$x = \pi$ で最大値

$$\mathrm{Si}(\pi) = 1.8519370$$

をとる．したがって $f_{2n-1}(\pi/2n)$ の最初のピークは

$$f_{2n-1}(\pi/2n) \sim 1.089489 \qquad n \to \infty$$

となって，n を無限大にしても正確な値 1 には収束しない．

テレスコーピング法　初等関数，たとえば三角関数などは，システムのライブラリーに組み込まれているから自分でプログラムを書く必要はないが，次のような関数はライブラリーだけではすまず，自分でプログラムを書く必要がある．

$$\mathrm{Sinc}(x) = \frac{\sin x}{x} \tag{5.19}$$

この関数はサンプリング定理にも出てくるものであるし，散乱の問題でもよく出てくる関数である．

$\sin x$ をライブラリー関数を用いるときには一点 $x = 0$ を除いては上の数式通りに計算することができる．しかし $x = 0$ 付近では 0/0 に近い計算をすることになるので，ライブラリー関数の $\sin x$ が $x = 0$ 付近で $\sin x$ を精度よく計算するとしても，$\sin x/x$ を精度よく計算してくれるとは限らない．そのために $x = 0$ を含む狭い区間では別の方法を考えなければならない．

そこで $\sin x$ をテーラー展開した近似式

$$\mathrm{Sinc}(x) \doteq 1 - \frac{x^2}{3!} + \frac{x^4}{5!} \tag{5.20}$$

を考えてみる．テーラー展開の性質として，近似の誤差は打ち切った最初の項，この例では $x^6/7!$ 程度であるから，x の範囲をたとえば $x^2 < 1/2$ とすれば上の近似式の誤差は

$$\frac{1}{7!}\frac{1}{2^3} = 2.5 \times 10^{-5}$$

である．テーラー展開の誤差は展開の中心では 0 であるが，中心を離れるにつれて冪乗の形で増加する．上の見積もりはその最大値を表している．

テーラー展開では近似区間の端で誤差が最大になるので，ある意味では効率が悪い．この欠点を簡便に解決する方法としてランチョス (Lanczos, 発音はよくわからない) のテレスコーピング法がある．近似の範囲を $x^2 \le 1/2$ として

$$x^2 = \frac{1}{2}\sin\theta$$

と変数変換すると，先の近似式で捨てた項は

$$\frac{x^6}{7!} = \frac{1}{7!2^3}\sin^3\theta = \frac{1}{7!2^3}\left(\frac{3}{4}\sin\theta - \frac{1}{4}\sin 3\theta\right)$$

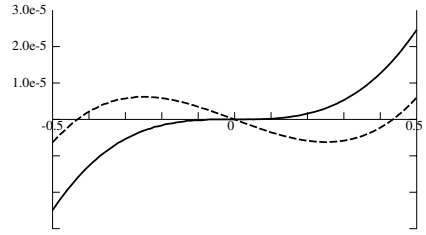
になる．$\sin\theta$，$\sin 3\theta$ の絶対値は 1 以下であるから第一項だけをとることにすれば

$$\frac{1}{7!2^3}\frac{3}{4}\sin\theta = \frac{1}{7!2^3}\frac{3}{2}x^2$$

を先の近似式に加えれば近似式

$$\mathrm{Sinc}(x) \doteq 1 - 0.16670387x^2 + \frac{x^4}{120} \tag{5.21}$$

が得られる．二つの近似式 (5.20)，(5.21) 式の誤差を下図に示してある．横軸は x^2 である．負まで示してあるのは x が虚数のときも考えたかったからである．実線がテーラー展開 (5.20) 式で，誤差は展開の中心 $x = 0$ を離れるにつれて急激に増加している．これに対して同じ三次式の破線の方は誤差の最大値は (5.20) 式の誤差の半分以下であり，近似区間で誤差は一様である．



チェビシェフ展開 テレスコーピング法で誤差が一様化されたのは，$\sin n\theta$ の振幅が一定であるところからきている．これに対して x^n は上限がないために近似誤差が大きくなるのである．

そこでこの方法を一般化する．$|x| \le 1$ で定義された関数 $f(x)$ を近似するために，まず

$$x = \cos\theta \tag{5.22}$$

で変数変換する．先の例とは異なり，今度は cos を用いている．$f(\cos\theta)$ は θ の偶関数であるから，これを (5.16) 式によってフーリエ級数で展開すれば $\cos k\theta$ の項だけが現れて

$$f(\cos\theta) \sim \frac{1}{2}c_0 + \sum_{k=1}^{\infty} c_k \cos k\theta \tag{5.23}$$
$$c_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\cos\theta)\cos k\theta d\theta$$

となる．$\cos k\theta$ の振幅は 1 であるから，$|c_k|$ の減衰の様子をみれば所望の精度に応じてどこで打ち切ればよいかがわかる．この展開を n 次で打ち切ったときの近似関数

$$f_n(x) = \frac{1}{2}c_0 + \sum_{k=1}^{n} c_k \cos k\theta \tag{5.24}$$

の誤差はおよそ $n+1$ 次の項で決まるから，誤差は区間内で一定の振幅 $|c_{n+1}|$ で振動しており，区間の端でとくに大きくなるようなことはない．したがってチェビシェフ展開を有限項で打ち切った (5.24) 式はミニマックス近似に非常に近い．

なお，上式をある x について計算するには §8 に導いてあるゲルツェルの方法が便利である．すなわち

$$\begin{aligned} &V_n = V_{n+1} = 0 \\ &V_{k-1} = c_k + 2xV_k - V_{k-1} \\ &\quad k = n, n-1, \cdots, 1 \\ &f_n(x) = \frac{1}{2}c_0 + xV_0 - V_1 \end{aligned} \tag{5.25}$$

一旦 c_k が求められると，(5.24) 式を x の多項式の形に直すこともできる．$\cos n\theta$ に倍角の公式を用いると $\cos n\theta$ が x の n 次多項式になる．すなわち

$$T_n(x) = \cos(n\cos^{-1}x) \tag{5.26}$$

は x の n 次多項式になり，これをチェビシェフの多項式という．最初の数項は

$$T_0(x) = 1 \quad T_1(x) = x \quad T_2(x) = 2x^2 - 1$$
$$T_3(x) = 4x^3 - 3x \quad T_4(x) = 8x^4 - 8x^2 + 1$$

である．漸化式

$$T_{n+1}(x) - 2xT_n(x) + T_{n-1}(x) = 0 \tag{5.27}$$

が成り立つことは cos の公式から簡単にわかる．この公式を用いれば (5.24) 式を x の多項式に変換するのは容易である．

テレスコーピング法の一般化 チェビシェフ展開 (5.23) 式の問題点は係数 c_k の積分が困難であるという点である．そこで先に述べた方法を拡張する．近似範囲を $|x| \le 1$ とし，関数 $f(x)$ のテーラー展開

$$f(x) = \sum_{n=0}^{\infty} a_n x^n$$

の係数 a_n がわかっているものとする．x の冪乗はチェビシェフ多項式を用いて

$$x^n = \frac{1}{2^{n-1}}\sum_{k=0}^{[n/2]} \binom{n}{k} T_{n-2k}(x) \tag{5.28}$$

と表すことができる．ここに $[n/2]$ は $n/2$ の整数部分であり，n が偶数のときには最後の項 $T_0(x)$ の係数に 1/2 を掛けなければならない (岩波全書，数学公式 ，p.89 の公式にはこの特例が欠けている)．この関係を用いて先のテーラー展開をすべてチェビシェフ多項式で置きかえると，展開

$$f(x) = \sum_{k=0}^{\infty} c_k T_k(x)$$

の係数 c_k を求めることができる．$|c_k|$ の大きさを見て途中で打ち切ることができるのは先と同様である．打ち切った近似式を (5.27) 式によって再び冪級数に変換することも簡単であるが，この過程で桁落ちや丸め誤差が生じる恐れがあるので，なるべくチェビシェフ多項式のまま用いる方が望ましい．

Sinc 関数の近似式 Sinc 関数 (5.19) 式はもともとたちの良い関数であるから，低次のテーラー展開で十分な近似が得られるが，テレスコーピング法でどれだけ精度が上がるか計算してみた例を以下に示す．これは (5.19) 式の $|x| \leq \pi$ における近似式を求めるためにスケーリングしたものであるが，(5.20) 式にみられるように x が虚数のときにも成り立つものである．すなわち

$$\mathrm{Sinc}(\pi\sqrt{x}) = \sum_{n=0}^{\infty} a_n x^n$$
$$a_n = \frac{(-1)^n \pi^{2n}}{(2n+1)!}$$

の展開係数 a_n と，これから求めたチェビシェフ展開の係数 c_n は次の表のようになる．

n	a_n	c_n
0	1.000000e+0	1.415723e+0
1	−1.644934e+0	−1.789468e+0
2	8.117425e−1	4.190149e−1
3	−1.907519e−1	−4.842341e−2
4	2.614786e−2	3.296368e−3
5	−2.346082e−3	−1.473941d−4
6	1.484289e−4	4.654237e−6
7	−6.975877e−6	−1.092549e−7
8	2.531219e−7	1.980867e−9
9	−7.304716e−9	−2.85340e−11

この表から，同じ次数ならチェビシェフ展開の方が誤差が一桁以上小さいことがわかる．

> Sinc 関数には特別な思い入れがあるので，あまり適切ではないが例題として用いた．三十年以上前に，表面波の分散曲線を計算するプログラムを開発しているとき，三角関数をライブラリー関数を用いると時間がかかりすぎるので (いまはそんなことはないが)，近似多項式を直接プログラムの中に書き込むことを考えた．そのために Sinc 関数とその微分の最良近似多項式が必要になり，自分で作る羽目になったというわけである．

選点直交多項式 チェビシェフ展開の係数 c_k を求めるもう一つの方法を述べる．$\cos n\theta$ は $0 \leq \theta \leq \pi$ に n 個の零点

$$\theta_k = (k - 1/2)\frac{\pi}{n} \qquad k = 1, 2, \cdots, n \tag{5.29}$$

をもつ．θ_k に関して次の関係が成り立つ．

$$\sum_{k=1}^{n} \cos m\theta_k = \sum_{k=1}^{n} \sin m\theta_k = 0 \qquad m \neq 0 \tag{5.30}$$

この関係は実際に三角関数の和を計算すれば得られるが，θ_k が複素平面上の単位円上に等間隔に並んだ点の偏角であり，上式がすべての点の実数部，虚数部の和であることからも明らかである．

これに対応して $T_n(x)$ は $-1 \leq x \leq 1$ に n 個の零点

$$T_n(x_k) = 0 \qquad x_k = \cos\theta_k \tag{5.31}$$

をもつ．この零点に関して (5.30) 式から次の関係が成り立つ．

$$\sum_{k=1}^{n} T_i(x_k) T_j(x_k) = \begin{cases} 0 & i \neq j \\ n/2 & i = j \neq 0 \\ n & i = j = 0 \end{cases} \tag{5.32}$$
$$i, j < n$$

この関係は $T_i(x)$ と $T_j(x)$ が離散的な点 x_k 上で直交していることを示している．

いま $-1 \leq x \leq 1$ で定義された関数 $f(x)$ に対して

$$c_k = \frac{2}{n} \sum_{j=1}^{n} f(x_j) T_k(x_j) \tag{5.33}$$

を求め

$$f_n(x) = \frac{1}{2} c_0 + \sum_{k=1}^{n-1} c_k T_k(x) \tag{5.34}$$

とすると $f_n(x)$ は

$$f_n(x_k) = f(x_k)$$

を満たす補間関数である．これは (5.33) 式を (5.34) 式に代入し，直交関係 (5.32) 式を用いれば導くことができる．

ミニマックス近似 区間 $[a,\ b]$ で関数 $f(x)$ を n 次の多項式

$$p_n(x) = a_0 + a_1 x + \cdots + a_n x^n$$

で近似するとき，誤差の最大値

$$\varepsilon = \max_{a \leq x \leq b} |p_n(x) - f(x)| \tag{5.35}$$

を最小にするような係数 $(a_0, a_1, \cdots, a_n)$ が一義的に決まる．このような近似式をミニマックス近似，

あるいは最良近似という．ミニマックス近似は，区間内で $n+2$ 個の誤差 (絶対値) の極大値をもち，その値はすべて等しく，隣同士の符号が反対になっている．

先に導いた二次式の近似式 (5.21) の誤差 (前図の破線) は，端点を含めて 4 個の極大をもち，その大きさはほとんど等しい．その意味でこれはミニマックス近似に非常に近い．

n 次までのチェビシェフ多項式を用いて関数を近似したとき，誤差はおよそ $T_{n+1}(x)$ に比例する．$T_{n+1}(x)$ の振幅は x によらず一定であるから，チェビシェフ展開はミニマックス近似に非常に近い．このことを用いるとミニマックス近似を比較的簡単に求めることができる．

区間 $|x| \le 1$ で関数 $f(x)$ を n 次多項式で近似するとき，誤差を $T_{n+1}(x)$ に比例するようにすることができれば，この近似多項式はチェビシェフ展開に近くなる．ところがこの誤差項は $T_{n+1}(x)$ の零点，(5.29) 式の n を $n+1$ にして得られる $n+1$ 個の x_j で 0 になる．したがってここでの近似式の誤差は 0 である．これは多項式の係数を a_i とすれば

$$a_0 + a_1 x_j + a_2 x_j^2 + \cdots + a_n x^n = f(x_j) \quad (5.36)$$
$$j = 1, 2, \cdots, n+1$$

を意味している．上式は $n+1$ 個の未知数 a_i に関する連立方程式であるから解くことができる．係数が求められると誤差

$$e(x) = p_n(x) - f(x)$$

を計算することができ，$e(x)$ の極大，極小の位置 x_k と，誤差 $|e(x_k)|$ を計算することができる．ミニマックスであれば区間の両端を含めて x_k は $n+2$ 個あるはずであり，$|e(x_k)|$ はすべて等しいはずであるが，実際にはそうはならない．そこで $|e(x_k)|$ の平均値を ε として，今度は

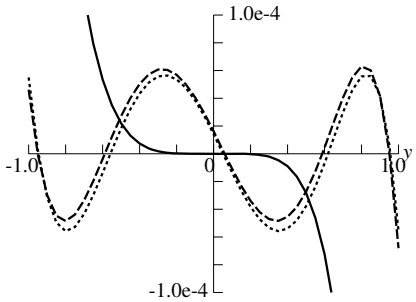
$$p_n(x_k) = f(x_k) \pm \varepsilon \quad (5.37)$$
$$k = 1, 2, \cdots, n+2$$

から a_i を改めて解きなおす．$\pm$ は $e(x_k)$ の符号と同じに選ぶ．上式は未知数よりも式の数の方が多いが，区間の両端のうちのどちらかを省いて解を求めればよい．こうして求まった新しい近似式の誤差曲線を計算し，上の手続きを繰り返せばミニマックス近似が求められる．

例をあげる．$\tan x/x$ を

$$x = \frac{\pi}{4}\sqrt{y} \qquad -1 \le y \le 1$$

と変数変換して y の 4 次式で $\tan x/x$ を近似する．y が負のときには tan は tanh になる．$T_5(y)$ の零点の座標を用いて (5.36) 式から決めた係数の誤差曲線が下図の破線である．曲線はわずかに非対称で，$y = 1$ の誤差が最も大きい．しかしこれを改良しても誤差が大幅に改良されるとは思えない．しかしあえて反復を一回行ってみると，誤差曲線は点線のようになる．たしかにミニマックスに近くはなっているが，改良はごく僅かである．



$\tan x/x$ にはテーラー展開

$$\frac{\tan x}{x} = 1 + \frac{1}{3}x^2 + \frac{2}{15}x^4 + \frac{17}{315}x^6 + \frac{62}{2835}x^8 + \cdots \quad (5.38)$$

がある．x^8 までの展開を用いて誤差曲線を描いたものが同じ図に実線で示してある．この展開は $|y| \sim 0.3$ までは非常に精度が高いが，それ以後は急激に誤差が大きくなる．同じ y^4 までの近似式でも，テーラー展開とミニマックス近似ではこれだけの違いがある．

重みつき直交多項式　内積 (5.8) 式のかわりに，重み関数 $w(x) \ge 0$ を用いて内積を定義することもできる．$p_k(x)$ を x の k 次の多項式とし，最高次の係数を $\mu_k \ne 0$ とする．下に示す例ではすべて $\mu_0 = 1$ である．$p_k(x)$ は次の直交関係を満足する．

$$(p_j, p_k) = \int_a^b w(x) p_j(x) p_k(x) dx = \lambda_k \delta_{jk} \quad (5.39)$$

よく用いられる直交多項式として，次のようなものがある．

名称	記号	$[a, b]$	$w(x)$	λ_n	α_k	β_k	γ_k
ルジャンドル	$P_n(x)$	$[-1, 1]$	1	$2/(2n+1)$	$2-1/k$	0	$1-1/k$
チェビシェフ	$T_n(x)$	$[-1, 1]$	$1/\sqrt{1-x^2}$	$\pi/2$	2	0	1
ラゲール	$L_n(x)$	$[0, \infty)$	$\exp(-x)$	1	$-1/k$	$2-1/k$	$1-1/k$
エルミート	$H_n(x)$	$(-\infty, \infty)$	$\exp(-x^2)$	$\sqrt{\pi}2^n n!$	1	0	$k-1$

直交多項式 $p_k(x)$ に対して次のような漸化式が形式的に成り立つ．

$$
\begin{aligned}
&p_0(x) = \mu_0 \qquad p_{-1}(x) = 0 \\
&p_k(x) = \mu_k x^k + \cdots \\
&p_k(x) = (\alpha_k x - \beta_k) p_{k-1}(x) - \gamma_k p_{k-2}(x) \\
&\alpha_k = \frac{\mu_k}{\mu_{k-1}} \qquad \beta_k = \frac{\alpha_k}{\lambda_{k-1}}(x p_{k-1},\, p_{k-1}) \quad (5.40) \\
&\gamma_k = \frac{\alpha_k \lambda_{k-1}}{\alpha_{k-1}\lambda_{k-2}}
\end{aligned}
$$

形式的という意味は，上式から α_k や β_k などが決まるわけではないからである．これらの値は上の表に示してある．上の関係は $p_k(x)$ が $p_{k-1}(x)$ と $p_{k-2}(x)$ に直交するという性質から導かれる．

また恒等式

$$
\begin{aligned}
\sum_{k=0}^{n-1} \frac{p_k(x)p_k(y)}{\lambda_k} &= \frac{\mu_{n-1}}{\mu_n \lambda_{n-1}} \\
&\times \frac{p_n(x)p_{n-1}(y) - p_{n-1}(x)p_n(y)}{x-y} \qquad (5.41)
\end{aligned}
$$

が成り立つ．これを**クリストッフェル・ダルブーの恒等式**という．これは非常に複雑な式に見えるが，分母を払った左辺を x だけの関数と考えると，これは x の n 次の多項式にほかならない．したがってこれは $p_n(x), p_{n-1}(x), \cdots$ で展開できるはずである．そこで左辺と $p_k(x)$ との内積を漸化式 (5.40) を利用して計算すれば右辺が得られる．

直交多項式による補間公式　これだけの準備が調うと $p_k(x)$ を用いた補間公式が導かれる．$p_n(x)$ の零点を $x_1, x_2, \cdots, x_n$ とすると，(5.41) 式から

$$
\begin{aligned}
\frac{1}{w_j} &\equiv \sum_{k=0}^{n-1} \frac{[p_k(x_j)]^2}{\lambda_k} \\
&= \frac{\mu_{n-1}}{\mu_n \lambda_{n-1}} p_{n-1}(x_j) p_n'(x_j) \qquad (5.42) \\
\sum_{k=0}^{n-1} \frac{p_k(x_j)p_k(x_l)}{\lambda_k} &= 0 \qquad j \neq l
\end{aligned}
$$

が成り立つ．第一式は (5.41) 式で $x = x_j$，$y \to x_j$ の極限であり，第二式は $y = x_l$ と置いたものである．これも一種の直交関係である．ただし，チェビシェフ多項式の直交関係 (5.32) と違って，ここでは多項式の次数についての和になっている．

この式を念頭に置いて

$$
c_k = \frac{1}{\lambda_k} \sum_{j=1}^{n} w_j p_k(x_j) f(x_j) \qquad (5.43)
$$

とすると

$$
f_n(x) = \sum_{k=0}^{n-1} c_k p_k(x) \qquad (5.44)
$$

が補間公式になる．

これが補間公式であることを示すには，上式に (5.43) 式を代入して和の順序を変えると

$$
f_n(x) = \sum_{j=1}^{n} \left[w_j \sum_{k=0}^{n-1} \frac{p_k(x_j)p_k(x)}{\lambda_k} \right] f(x_j)
$$

となる．ここで $x = x_l$ と置けば (5.42) 式によって $j = l$ の項だけが残って $f_n(x_l) = f(x_l)$ が成り立つ．したがって (5.44) 式は与えられた点を通る補間公式になっている．

すぐ上の式の角括弧の中を (5.41) 式を用いて書き直せば

$$
f_n(x) = \sum_{j=1}^{n} \frac{p_n(x)}{(x-x_j)p_n'(x_j)} f(x_j) \qquad (5.45)
$$

が得られる．これはラグランジェの公式 (5.2) と同じ形である．x_j は $p_n(x)$ の零点であるから，これが成り立つのは当然である．

選点直交多項式という術語は,「数学公式　」(岩波全書) ではもっと狭い意味に定義しているが,ここでは補間公式が導かれるという意味に用いている．

連分数 関数 $f(x)$ が連分数

$$f(x) = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \ddots}}} \tag{5.46}$$

で定義される場合がある．a_i, b_i は x の関数である．連分数をこのように書くと場所をとりすぎるので，次のように省略することが多い．

$$f(x) = b_0 + \frac{a_1}{b_1+}\ \frac{a_2}{b_2+}\ \frac{a_3}{b_3+}\ \frac{a_4}{b_4+}\ \cdots \tag{5.47}$$

初等関数の連分数展開としては

$$\tan x = \frac{x}{1-}\ \frac{x^2}{3-}\ \frac{x^2}{5-}\cdots$$
$$e^x = 1 + \frac{x}{1-}\ \frac{x}{2+}\ \frac{x}{3-}\ \frac{x}{2+}\cdots\frac{x}{2n+1-}\ \frac{x}{2+}\cdots$$
$$\log(1+x) = \frac{x}{1+}\ \frac{x}{2+}\ \frac{x}{3+}\ \frac{2x}{2+}\ \frac{2x}{5+}\cdots$$
$$\cdots\frac{nx}{2+}\ \frac{nx}{2n+1+}\cdots$$
$$\tan^{-1} x = \frac{x}{1+}\ \frac{x^2}{3+}\ \frac{4x^2}{5+}\cdots\frac{n^2x^2}{2n+1+}\cdots$$

などがある．また，ベッセル関数，ノイマン関数などの円筒関数は漸化式

$$Z_{\nu-1}(x) + Z_{\nu+1}(x) = \frac{2\nu}{x} Z_\nu(x)$$

を満足しているので，その比は次のような連分数表示ができる．

$$\frac{Z_{\nu-1}(x)}{Z_\nu(x)} = \frac{2\nu}{x} - \frac{1}{Z_\nu(x)/Z_{\nu+1}(x)}$$
$$= \frac{2\nu}{x} - \frac{1}{2(\nu+1)/x-}\ \frac{1}{2(\nu+2)/x-}$$
$$\frac{1}{2(\nu+3)/x-}\cdots$$
$$\frac{Z_{\nu+1}(x)}{Z_\nu(x)} = \frac{x}{2(\nu+1)-}\ \frac{x^2}{2(\nu+2)-}$$
$$\frac{x^2}{2(\nu+3)-}\cdots$$

ルジャンドル関数など，同様な漸化式を満たす特殊関数に対しても，連分数表示を導くことができる．

連分数を計算するのに (5.47) 式の右端から左に向かって計算するのが最も常識的である．第 n 項の分母を

$$\Delta_n = b_n + \frac{a_{n+1}}{b_{n+1} + \cdots}$$

と定義すれば，第二項の分母は Δ_{n+1} にほかならないから，漸化式

$$\Delta_n = b_n + \frac{a_{n+1}}{\Delta_{n+1}} \tag{5.48}$$

が得られる．十分大きな N に対して初期値を

$$\Delta_{N+1} = b_{N+1}$$

として，上の漸化式を $n = N, N-1, \cdots, 0$ の順に計算すれば Δ_0 が求める $f(x)$ である．

この方法は単純ではあるが，どこで打ち切ればよいかは N を変えてみて精度を確認しなければならないので試行錯誤が必要になる．しかし (5.47) 式を左から計算するにはウォリスの方法という古典的なアルゴリズムがある．これは次のようにまとめられる．

$$\begin{aligned} &A_{-1} = 1 \quad B_{-1} = 0 \quad A_0 = b_0 \quad B_0 = 1 \\ &A_j = b_j A_{j-1} + a_j A_{j-2} \\ &B_j = b_j B_{j-1} + a_j B_{j-2} \quad j = 1, 2, \cdots \\ &f_n = \frac{A_n}{B_n} \end{aligned} \tag{5.49}$$

f_n は (5.47) 式を a_n, b_n まで計算したものになっている．この式が正しいことは $j = 1, 2$ くらいまでは手計算で確かめることができる．一般的には少し手間はかかるが数学的帰納法を用いて証明することができる．反復は $|f_j - f_{j-1}|$ が十分小さくなった止めればよい．

(5.49) 式は A_j, B_j に関して線型であるから，取り扱いが簡単である．しかし場合によってはこれらの絶対値が非常に大きくなってしまう場合がある．その場合にはスケーリングを行う必要がある．

この方法の変形として

$$C_j = \frac{A_j}{A_{j-1}} \quad D_j = \frac{B_{j-1}}{B_j}$$

と置くと

$$f_j = f_{j-1} C_j D_j$$

である．C_j D_j は次の漸化式を満たしている．

$$\begin{aligned} &f_0 = C_0 = b_0 \quad D_0 = 0 \\ &C_j = b_j + \frac{a_j}{C_{j-1}} \\ &D_j = \frac{1}{b_j + a_j D_{j-1}} \quad j = 1, 2, \cdots \end{aligned} \tag{5.50}$$

$|C_jD_j-1|$ が十分小さくなったら反復を止めればよい．この方法では C_j や D_j の分母が 0 になってしまうおそれがあるが，そのときには計算機内で許される最小の数値で置きかえれば，次のステップでは正しい値に補正される．

成層構造中を伝わる弾性波動を計算するのによく用いられている方法にトムソン・ハスケル法がある．これは形の上では線型の方程式 (5.49) 式に相当している．しかし同じ問題を反射係数を用いて (5.50) 式の形で解くことができる．また，電磁探査で必要になる成層大地の反射係数は通常 (5.50) 式の形で計算されているが，これは逆に (5.49) 式のような線型問題として解くこともできる．

有理関数近似　連分数を有限項で打ち切ると多くの場合多項式の比，すなわち有理関数が得られる．たとえば $\tan x$ を 3 項で打ち切ると

$$\frac{\tan x}{x} \doteq \frac{1}{1-}\,\frac{x^2}{3-}\,\frac{x^2}{5} = \frac{15-x^2}{15-6x^2}$$

という簡単な近似式が得られる．この近似式は単純なわりには驚くほど精度が高い．下表に $\tan x/x$ の値と，近似式の誤差を示す．

x	$\tan x/x$	有理式	テーラー
0.10	1.003346	0.640e-9	0.009e-9
0.20	1.013550	4.204e-8	0.923e-9
0.30	1.031121	5.001e-7	5.432e-8
0.40	1.056983	2.990e-6	9.938e-7
0.50	1.092605	1.239e-5	9.631e-6
0.60	1.140228	4.110e-5	6.275e-5
0.70	1.203269	1.182e-4	3.124e-4
0.80	1.287048	3.098e-4	1.285e-3
0.90	1.400176	7.675e-4	4.601e-3
1.00	1.557408	1.852e-3	1.490e-2

ここでテーラーとあるのは，$\tan x$ のテーラー展開 (5.38) 式の誤差である．$x>0.5$ になると有理式の誤差の方がテーラー展開のそれよりも小さい．

パラメーターの数が同じ多項式近似と有理関数近似を比べると，有理関数近似の方が誤差が 1 桁，場合によっては数桁も小さくなることがある．また，有理関数近似に対してもミニマックスの定理が成り立つ．すなわち，以下のように分子が m 次多項式，分母が n 次多項式の場合，区間内の誤差の極大は $m+n+2$ 個あり，絶対値が等しく符号は交替している．

$f(x)$ を，分子分母がそれぞれ m 次，n 次の多項式である有理関数

$$f(x) \sim \frac{A_m(x)}{B_n(x)} = \frac{\sum_{i=0}^{m} a_i x^i}{1+\sum_{j=1}^{n} b_j x^j} \tag{5.51}$$

で近似することを考える．a_i や b_j は連分数展開のときの a_i や b_j とはまったく関係ない．a_i や b_j は比だけが問題であるから，ここでは $b_0=1$ とする．これらの係数を求める最も単純な方法は，$m+n+1$ 個の点 x_i を**適当に**選んで上式が等号になるようにすることである．すなわち連立一次方程式

$$\begin{aligned} &a_0+a_1x_i+a_2x_i^2+\cdots+a_mx_i^m \\ &\quad = f(x_i)(1+b_1x_i+b_2x_i^2+\cdots+b_nx_i^n) \\ &\qquad i=1,2,\cdots,m+n+1 \end{aligned} \tag{5.52}$$

から係数 $a_0\sim a_m, b_1\sim b_n$ を決める．しかし x_i を適当に選ぶのは難しく，選び方によってはラグランジュ補間のように点と点の間で補間値が暴れてしまうおそれもある．それよりも，分点を $m+n+1$ 個よりも多くとって，上式を最小二乗法的に解く方が安全である．このときの分点 x_i としては十分高次のチェビシェフ多項式の零点を用いるとうまくいくことが多い．

一旦係数が求められると，誤差曲線

$$e(x) = \frac{A_m(x)}{B_n(x)} - f(x)$$

を描くことができる．もしこの近似式がミニマックス近似なら，両端を含めて $m+n+2$ 個の極大極小点 x_j をもち

$$\begin{aligned} &A_m(x_j) = [f(x_j)\pm\varepsilon]\,B_n(x_j) \\ &\qquad j=1,2,\cdots,m+n+2 \end{aligned} \tag{5.53}$$

を満足しているはずである．ε は誤差の絶対値の最大値である．$\pm$ の符号は誤差の符号が交替していることを表している．そこでいま求めた誤差曲線 $e(x)$ から誤差の極大極小点の位置 x_j と ε を推定し，改めて上式を解いて係数 a_i b_j を決めなおす．新たに決められた近似式から $e(x)$ を計算して再び同じ手続きをくりかえせば，ミニマックス近似に近い近似式を

求めることができる．なお，(5.53) 式では $m+n+1$ 個の未知数 a_i b_j に対して式が $m+n+2$ 個あるが，実は ε も未知数であるあるから数は合っている．

(5.52) 式を最小二乗法的に解いて近似有理式を求めるという方法は，測定値を有理式近似するのに役に立つかもしれない．

パデ展開　反復を行わないでも有理式近似を求める方法にパデ展開がある．

多項式 $A_m(x)$, $B_n(x)$ を求めるために $f(x)$ のテーラー展開

$$f(x) = \sum_{k=0}^{\infty} c_k x^k$$

が知られていると仮定する．これを (5.51) 式に代入して分母を払うと

$$B_n(x) \sum_{k=0}^{\infty} c_k x^k \sim A_m(x)$$

となる．両辺の x^i の係数が等しくなるように a_i, b_j を決めることができれば，テーラー展開に対応した有理式近似が求まることになる．

上式の左辺は

$$\sum_{j=0}^{n} \sum_{k=0}^{\infty} b_j c_k x^{j+k} = \sum_{i=0}^{\infty} \left(\sum_{j=0}^{i} b_j c_{i-j} \right) x^i$$

と書きかえることができる．右辺は x^m までしかないので

$$\sum_{j=0}^{i} b_j c_{i-j} = a_i \qquad i = 0, 1, \cdots, m \tag{5.54}$$

が成り立たなければならない．右辺には x^m よりも高次の項がないが，左辺には生じてしまうのでその係数が 0 にならなければならない．

$$\sum_{j=0}^{i} b_j c_{i-j} = 0 \tag{5.55}$$
$$i = m+1, m+2, \cdots, m+n$$

$i = m+n$ で止めているのは，b_j は $b_0 = 1$ は既知であるが，未知数は $b_1 \sim b_n$ の n 個しかないから n 個以上の条件をつけるわけにはいかないからである．なお，この式の和の上限が i になっているが，b_j は b_n までしかないから実際の上限は $\min(i, n)$ である．

c_k は既知であるから (5.55) 式は $b_1 \sim b_n$ に関する連立一次方程式である．これを解いて (5.50) 式の左辺に代入すれば a_i が求められる．これがパデ展開である．

先の $\tan x/x$ のテーラー展開をもとに $m = n = 2$ のパデ展開を求めた結果

$$A_2(x) = 1 - 0.11111233x^2 + 1.0582954\text{e-}3x^4$$
$$B_2(x) = 1 - 0.44444567x^2 + 1.5873516\text{e-}2x^4$$

が得られた．このパデ展開の誤差は $0 \le x \le 1$ では単精度の丸め誤差の限界に近く，10^{-7} 以下である．パデ展開のもとになった x^8 までのテーラー展開の誤差は，前の表にあるとおり $x = 1$ で 1.5×10^{-2} であるから，このパデ展開は**もとになった関数よりも高い精度**が得られている．また，先にあげた $\tan x$ の収束半径は $\pi/2$ であるが，8 次までの近似式は $x > 1$ ではまったく信用できない．これに対して上で求めたパデ展開は $\tan x$ の極 $x = \pi/2$ とほとんど同じところに極をもち，しかも極を越えた $x = 2$ でも 10^{-4} の精度をもっている．このように，パデ展開はもとになった級数には見えないが，その関数が本来もっている性質を引き出してしまうという特徴がある．これはありがたいことではあるが，反面意図しない結果が生じるおそれもある．実は上で求められたパデ展開は，$\tan x$ の連分数表示を漸化式 (5.49) 式を用いて $j = 5$ まで計算した

$$A_5 \sim 1 - \frac{1}{9}x^2 + \frac{1}{945}x^4$$
$$B_5 \sim 1 - \frac{4}{9}x^2 + \frac{1}{63}x^4$$

を再現している．

パデ展開で求めた多項式 $A_m(x)$ と $B_n(x)$ が互いに素であるという保証はない．極端な場合には $A_m(x)$ が $B_n(x)$ で割り切れてしまって，$A_m(x)/B_n(x)$ が単なる多項式になってしまう場合さえある．これはテーラー展開の係数 c_k の性質によって決まり，次数の低下が起こらないための条件も求められているが，ここでは省略する．

パデ展開はもとになったテーラー展開よりはるかに良い近似を与えるし，収束半径の心配もない．初等関数についてパデ展開が解析的に求められているので，代表的なものを下に示す．

$$e^x \sim \frac{1 + x/2 + x^2/10 + x^3/120}{1 - x/2 + x^2/10 - x^3/120}$$

$$\sqrt{x} \sim \frac{1+10x+5x^2}{5+10x+x^2}$$
$$\sqrt[3]{x} \sim \frac{5+35x+14x^2}{14+35x+5x^2}$$
$$\frac{\tan^{-1}x}{x} \sim \frac{945+735x^2+202x^4}{945+1050x^2+225x^4}$$

すぐにわかるように，$\sqrt{x}$ や $\sqrt[3]{x}$ の近似式は $x=0$ 付近では精度がでない．たとえば $\sqrt{x}$ の近似式では最大誤差 ε は

$$0.8 < x < 1.2 \quad \varepsilon = 8\times10^{-7}$$

であるが，範囲を広げると

$$0.5 < x < 2.0 \quad \varepsilon = 3\times10^{-4}$$

となる．

チェビシェフ・パデ展開　テーラー展開に比べてチェビシェフ展開の方が優れていることは既に述べた．そこで，$-1 \le x \le 1$ で定義された関数の，テーラー展開のかわりにチェビシェフ展開

$$f(x) = \sum_{k=0}^{\infty} c_k T_k(x)$$

がわかっているとする．こんどは (5.51) 式の $A_m(x)$, $B_n(x)$ としては多項式ではなく

$$A_m(x) = \sum_{i=0}^{m} a_i T_i(x) \qquad B_n(x) = \sum_{j=0}^{n} b_j T_j(x)$$

を仮定する．テーラー展開のときには $B_n(x)f(x)$ が簡単に計算できたが，こんどはそうはいかない．

まず，cos の加法定理を利用すると

$$2T_j(x)T_k(x) = T_{j+k}(x) + T_{|j-k|}(x)$$

が成り立つことは容易にわかる．この関係を用いると

$$\begin{aligned} T_j(x)\sum_{k=0}^{\infty} c_k T_k(x) &= \frac{1}{2}\sum_{k=0}^{\infty} c_k [T_{j+k}(x) + T_{|j-k|}(x)] \\ &= \sum_{k=0}^{\infty} c_k^{(j)} T_k(x) \end{aligned}$$

の展開係数 $c_k^{(j)}$ を求めることができる．明らかに $c_k^{(0)} = c_k$ である．したがって

$$B_n(x)f(x) = \sum_{j=0}^{n}\sum_{k=0}^{\infty} b_j c_k^{(j)} T_k(x) \sim \sum_{i=0}^{m} a_i T_i(x)$$

でなければならない．これより，前と同じ考え方で

$$\sum_{j=0}^{n} b_j c_k^{(j)} = 0 \quad k = m+1, m+2, \cdots, m+n$$
$$\sum_{j=0}^{n} b_j c_k^{(j)} = a_k \quad k = 0, 1, 2, \cdots, m \qquad (5.56)$$

が導かれる．第一式を b_j について解き，これを第二式に代入すれば a_i が求められる．

参考文献

一松　信，1963：近似式，竹内書店．

6 行列の変換と固有値問題

行列の固有値問題に入る前に，行列の変換について述べる．既に §4 において，ガウスの消去法が係数行列 $\boldsymbol{A}$ を LU 分解して解くことに相当していることを知った．本節ではまず行列を QR 分解する方法を二つ述べる．なお，以下では行列の要素は実数とする．

グラム・シュミット法（QR 分解） ベクトルを直交化する方法として，グラム・シュミット法がよく知られている．この方法を行列の列ベクトルに対して適用する．

$n \times m$ $(n \geq m)$ の行列 $\boldsymbol{A}$ を n 次元の列ベクトル $\boldsymbol{a}_j$ の集まりと考えて

$$\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_m] \tag{6.1}$$

と書く．$\boldsymbol{A}$ は正方行列ではなくてもいいことに注意する．$\boldsymbol{a}_j$ は直交していないのでこれらを正規直交化する．まず

$$\boldsymbol{q}_1 = \frac{\boldsymbol{a}_1}{r_{11}} \qquad r_{11} = \|\boldsymbol{a}_1\|$$

と置く．ノルムは L_2 ノルムである．次の列 $\boldsymbol{a}_2$ に含まれている $\boldsymbol{q}_1$ 成分の大きさは $\boldsymbol{q}_1$ と $\boldsymbol{a}_2$ の内積で表されるから，$\boldsymbol{a}_2$ から $\boldsymbol{q}_1$ 成分を差し引いて

$$\boldsymbol{u}_2 = \boldsymbol{a}_2 - \boldsymbol{q}_1 r_{12} \qquad r_{12} = \boldsymbol{q}_1^T \boldsymbol{a}_2$$

を作る．ここで ^{T} は転置行列を意味している．$\boldsymbol{u}_2$ が $\boldsymbol{q}_1$ に直交することは明らかである．そこで $\boldsymbol{u}_2$ の長さを 1 に正規化した

$$\boldsymbol{q}_2 = \frac{\boldsymbol{u}_2}{r_{22}} \qquad r_{22} = \|\boldsymbol{u}_2\|$$

が次のベクトル $\boldsymbol{q}_2$ である．

これを一般化すれば，k 番目のベクトル $\boldsymbol{q}_k$ は

$$\boldsymbol{u}_k = \boldsymbol{a}_k - \sum_{j=1}^{k-1} \boldsymbol{q}_j r_{jk} \qquad r_{jk} = \boldsymbol{q}_j^T \boldsymbol{a}_k$$

$$\boldsymbol{q}_k = \frac{\boldsymbol{u}_k}{r_{kk}} \qquad r_{kk} = \|\boldsymbol{u}_k\| \tag{6.2}$$

によって計算できる．$\boldsymbol{u}_k$ は $\boldsymbol{a}_k$ からこれまでに決まった直交ベクトル $\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_{k-1}$ の成分をすべて差し引いたものであるから，前に決まったベクトルすべてに直交している．$\boldsymbol{u}_k$ の式を

$$\boldsymbol{a}_k = \boldsymbol{q}_k r_{kk} + \sum_{j=1}^{k-1} \boldsymbol{q}_j r_{jk} = \sum_{j=1}^{k} \boldsymbol{q}_j r_{jk}$$

と書き換えてみれば，これが行列の関係で

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R} \qquad \boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_m] \tag{6.3}$$
$$\boldsymbol{R} = [r_{ij}]$$

と書き表されることがわかる．ここで $\boldsymbol{Q}$ は $n \times m$ 行列，$\boldsymbol{R}$ は $m \times m$ の右上三角行列

$$\boldsymbol{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ & r_{22} & \cdots & r_{2m} \\ & & \ddots & \vdots \\ 0 & & & r_{mm} \end{bmatrix}$$

である．$\boldsymbol{q}_k$ の正規直交性から，$\boldsymbol{Q}$ は

$$\boldsymbol{Q}^T \boldsymbol{Q} = \boldsymbol{I}_m \tag{6.4}$$

を満たしている．$\boldsymbol{I}_m$ は $m \times m$ の単位行列である．ただし $n \neq m$ のときには $\boldsymbol{Q}\boldsymbol{Q}^T$ は単位行列にはならない．(6.4) 式のみが成り立つ行列を，以下では半直交行列と呼ぶことにする．

修正グラム・シュミット法 グラム・シュミット法はわかりやすいが丸め誤差に弱く，ベクトル間の直行性がだんだん悪くなってくる．これはまとめて $\boldsymbol{q}_j$ 成分を引いているからである．そこで一つの $\boldsymbol{q}_k$ が求まるたびに，その成分をすべての $\boldsymbol{a}_j$ から引くことにする．

まず最初の行列を改めて

$$\boldsymbol{A} = [\boldsymbol{a}_1^{(1)}, \boldsymbol{a}_2^{(1)}, \cdots, \boldsymbol{a}_m^{(1)}]$$

と置く．$\boldsymbol{q}_1$ は先と同様に定義する．異なるのは，$\boldsymbol{q}_1$ が求まった段階で，$\boldsymbol{a}_2^{(1)}, \boldsymbol{a}_3^{(1)}, \cdots, \boldsymbol{a}_m^{(1)}$ すべてから $\boldsymbol{q}_1$ 成分を差し引いてこれを $\boldsymbol{a}_2^{(2)}$ などとすることである．すなわち，一般に $\boldsymbol{q}_{k-1}$ までが求まった段階では次の $\boldsymbol{q}_k$ は

$$\boldsymbol{a}_j^{(k)} = \boldsymbol{a}_j^{(k-1)} - \boldsymbol{q}_{k-1} r_{k-1,j}$$
$$r_{k-1,j} = \boldsymbol{q}_{k-1}^T \boldsymbol{a}_j^{(k-1)} \quad j = k, k+1, \cdots, m \tag{6.5}$$
$$\boldsymbol{q}_k = \frac{\boldsymbol{a}_k^{(k)}}{r_{kk}} \qquad r_{kk} = \|\boldsymbol{a}_k^{(k)}\|$$

によって計算される．$\boldsymbol{q}_m$ が求まったときに計算終了になる．こうして求まった $\boldsymbol{Q}$，$\boldsymbol{R}$ はもちろん (6.3), (6.4) 式を満足している．

(6.5) 式には r_{kk} による割り算が現れている．これはガウスの消去法などにおけるピボットにほかならない．これが大きい方が丸め誤差が小さくなるから，$\boldsymbol{a}_j^{(k)}$ $(j \geq k)$ が求まった段階でノルムが最大の $\boldsymbol{a}_j^{(k)}$ と $\boldsymbol{a}_k^{(k)}$ とを入れかえて計算を行なうのがよい．これは列の入れかえであるから，後で結果を利用するときには注意が必要である．なお，この計算法では $\boldsymbol{Q}$ は $\boldsymbol{A}$ の上に上書きして求めることができるが，$\boldsymbol{R}$ には別のメモリーが必要である．

$\boldsymbol{Q}$ の直交性を利用すると連立方程式を簡単に解くことができる．$n = m$ のとき連立方程式 $\boldsymbol{Ax} = \boldsymbol{b}$ に左から $\boldsymbol{Q}^T$ を掛けると

$$\boldsymbol{Q}^T\boldsymbol{QRx} = \boldsymbol{Q}^T\boldsymbol{b} \quad \boldsymbol{Rx} = \boldsymbol{Q}^T\boldsymbol{b}$$

と書き換えることができる．左辺は上三角方程式であるから後退代入法で解くことができる．右辺は r_{jk} と全く同じ形をしている．そこで $\boldsymbol{A}$ の右端に右辺の列ベクトル $\boldsymbol{b}$ を加え，これを $\boldsymbol{a}_{n+1}^{(1)}$ として反復 (6.5) 式を繰り返す．最後に求められた $n \times (n+1)$ 行列 $\boldsymbol{R}$ の最後の列 $r_{i,n+1}$ が上式の右辺である．これを右辺に用いて上三角方程式を解けばよい．なお，n 次元ベクトルは n 個の直交ベクトルで完全に表されるから，最後に求められたベクトル $\boldsymbol{a}_{n+1}^{(n+1)}$ は 0 ベクトルになっているはずである．もちろんこの方法はガウスの消去法よりは効率が悪い．

以下では単にグラム・シュミット法と呼ぶときにも，実際の計算には修正グラム・シュミット法を用いることを前提にする．

ハウスホールダー変換（QR 分解） ノルムが $\sqrt{2}$ の任意のベクトル $\boldsymbol{w}$ を用いて

$$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{w}\boldsymbol{w}^T \quad \|\boldsymbol{w}\|^2 = 2 \tag{6.6}$$

を作ると，$\boldsymbol{P}$ は対称な直交行列

$$\boldsymbol{P}^T = \boldsymbol{P} \quad \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I} \tag{6.7}$$

である．ただし $\boldsymbol{I}$ は単位行列である．この行列を**基本直交変換**という．この行列を任意のベクトル $\boldsymbol{a}$ に掛けたとき，ベクトルのノルムは保存される．すなわち

$$\|\boldsymbol{Pa}\|^2 = \boldsymbol{a}^T\boldsymbol{P}^T\boldsymbol{Pa} = \|\boldsymbol{a}\|^2 \tag{6.8}$$

が成り立つ．$\boldsymbol{P}$ として (6.6) 式を用いた変換をハウスホールダー変換というが，ここではこの変換を用いて行列の QR 分解を行う．

あるベクトル $\boldsymbol{a}$ が与えられたとき，$\boldsymbol{w}$ をうまく選ぶと

$$\boldsymbol{Pa} = \boldsymbol{a} - \boldsymbol{w}(\boldsymbol{w}^T\boldsymbol{a}) = [s,\, 0,\, \cdots,\, 0]^T \tag{6.9}$$
$$s^2 = \|\boldsymbol{a}\|^2$$

のように $\boldsymbol{Pa}$ の第一成分以外はすべて 0 にすることができる．上式に左から $\boldsymbol{a}^T$ を掛けると

$$\|\boldsymbol{a}\|^2 - (\boldsymbol{w}^T\boldsymbol{a})^2 = sa_1$$

となる．a_1 は $\boldsymbol{a}$ の第一成分である．これから $\boldsymbol{w}^T\boldsymbol{a}$ が決まり，(6.9) 式の両辺を比較することによって $\boldsymbol{w}$ の成分が決まる．$\boldsymbol{w}$ の具体的な形は

$$\boldsymbol{w}^T\boldsymbol{a} = \sqrt{s(s-a_1)} \quad s = \|\boldsymbol{a}\|\text{sign}\,(-a_1)$$
$$\boldsymbol{w} = \frac{1}{\boldsymbol{w}^T\boldsymbol{a}}[a_1 - s, a_2, \cdots, a_n]^T \tag{6.10}$$

で与えられる．s の符号は $s - a_1$ の計算で桁落ちが起こらないように選んである．実際の計算では上式を用いるよりも

$$\boldsymbol{P} = \boldsymbol{I} + \frac{\boldsymbol{u}\boldsymbol{u}^T}{s(a_1 - s)} \tag{6.11}$$
$$\boldsymbol{u}^T = [a_1 - s, a_2, \cdots, a_n]$$

を用いた方が便利である．任意のベクトル $\boldsymbol{b}$ に $\boldsymbol{P}$ を掛けたとき，結果は

$$\boldsymbol{Pb} = \boldsymbol{b} + \beta\boldsymbol{u} \quad \beta = \frac{\boldsymbol{u}^T\boldsymbol{b}}{s(a_1 - s)}$$

と，簡単に計算することができる．

$n \times m$ $(n \geq m)$ 行列 $\boldsymbol{A}$ が与えられたとき，上の $\boldsymbol{a}$ として $\boldsymbol{A}$ の一列目のベクトル $\boldsymbol{a}_1$ を用いて $\boldsymbol{P}$ を作り，これを $\boldsymbol{P}_1$ とする．積 $\boldsymbol{P}_1\boldsymbol{A}$ の 1 列目は第 2 成分以下がすべて 0 になる．次に，$\boldsymbol{P}_1\boldsymbol{A}$ の二列目の第二成分以下の $n-1$ 成分のベクトルを $\boldsymbol{a}$ と考えて $\boldsymbol{P}_2$ を作る．これを $\boldsymbol{P}_1\boldsymbol{A}$ に掛けたときに一行目，一列目が変化しないためには，$\boldsymbol{P}_2$ の一行目，一列目は $(1,1)$ 成分が 1，それ以外は 0 でなければならない．一般に $\boldsymbol{P}_k$ は $(k,\, k)$ 要素より右下の行列が対象になる．すなわち $\boldsymbol{P}_k$ は

$$\boldsymbol{P}_k = \left[\begin{array}{c|c} \boldsymbol{I}_{k-1} & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{I}_{n-k+1} - \boldsymbol{w}\boldsymbol{w}^T \end{array}\right]$$

の形をしている．$\boldsymbol{I}_k$ は $k \times k$ の単位行列である．

このようにして左から $\boldsymbol{P}_k$ を掛けていくと，対角要素より下に 0 が導入されていき，行列 $\boldsymbol{A}$ が $m(m \leq n)$ 列のとき

$$\boldsymbol{P}_m \boldsymbol{P}_{m-1} \cdots \boldsymbol{P}_1 \boldsymbol{A} = \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{bmatrix} \tag{6.12}$$

の形になる．$\boldsymbol{R}$ は $m \times m$ の上三角行列，$\boldsymbol{0}$ は $(n-m) \times m$ の零行列である．なお，$m = n$ のときには $\boldsymbol{P}_m$ は必要ない．

直交変換の積を

$$\boldsymbol{Q} = \boldsymbol{P}_1 \boldsymbol{P}_2 \cdots \boldsymbol{P}_m \tag{6.13}$$

とする．$\boldsymbol{Q}$ は (6.3) 式の $\boldsymbol{Q}$ と違って $n \times n$ の正方行列である．この $\boldsymbol{Q}$ を用いると (6.12) 式は

$$\boldsymbol{Q}^T \boldsymbol{A} = \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{bmatrix}$$

となる．$\boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I}$ であるから，この式から

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

が成り立つ．$n > m$ のとき，$\boldsymbol{Q}$ の最初の m 列をとれば上式は (6.3) 式と同じである．ただし，(6.12) 式のような変換を行って $\boldsymbol{R}$ を求めるだけなら，$\boldsymbol{Q}$ をわざわざ計算する必要はない．また，$\boldsymbol{R}$ は $\boldsymbol{A}$ に上書きすることができる．したがって計算は修正グラム・シュミット法よりも効率がよい．

連立方程式 $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ の場合 $(n = m)$，$\boldsymbol{A}$ に $\boldsymbol{P}_k$ を掛けていくと同時に右辺 $\boldsymbol{b}$ にも左から $\boldsymbol{P}_k$ を掛けていくと，もとの方程式は

$$\boldsymbol{P}_{n-1} \boldsymbol{P}_{n-2} \cdots \boldsymbol{P}_1 \boldsymbol{A}\boldsymbol{x} = \boldsymbol{R}\boldsymbol{x} = \boldsymbol{P}_{n-1} \boldsymbol{P}_{n-2} \cdots \boldsymbol{P}_1 \boldsymbol{b}$$

に変換されるから，後退代入法によって容易に解くことができる．このときには $\boldsymbol{Q}$ を求める必要はまったくない．

$\boldsymbol{QL}$ 分解　前二項では行列 $\boldsymbol{A}$ を直交行列 $\boldsymbol{Q}$ と右上三角行列 $\boldsymbol{R}$ の積 $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ に分解したが，反対に $\boldsymbol{Q}$ と左下三角行列

$$\boldsymbol{L} = \begin{bmatrix} l_{11} & & & 0 \\ l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{m1} & l_{m2} & \cdots & l_{mm} \end{bmatrix}$$

の積 $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{L}$ に分解することもできる．もちろんこの $\boldsymbol{Q}$ は前の $\boldsymbol{Q}$ とは異なる．

計算は前と逆順に行なう．わかりやすいようにグラム・シュミット法で説明する．はじめに

$$\boldsymbol{q}_m = \frac{\boldsymbol{a}_m}{l_{mm}} \qquad l_{mm} = \|\boldsymbol{a}_m\|$$

とする．次に $\boldsymbol{a}_{m-1}$ から $\boldsymbol{q}_m$ 成分を差し引いて

$$\begin{aligned} \boldsymbol{u}_{m-1} &= \boldsymbol{a}_{m-1} - l_{m,m-1}\boldsymbol{q}_m \\ l_{m,m-1} &= \boldsymbol{q}_m^T \boldsymbol{a}_{m-1} \\ \boldsymbol{q}_{m-1} &= \frac{\boldsymbol{u}_{m-1}}{l_{m-1,l-1}} \qquad l_{m-1,m-1} = \|\boldsymbol{u}_{m-1}\| \end{aligned}$$

である．一般項は

$$\begin{aligned} \boldsymbol{u}_k &= \boldsymbol{a}_k - \sum_{j=k+1}^{m} \boldsymbol{q}_j l_{jk} \qquad l_{jk} = \boldsymbol{q}_j^T \boldsymbol{a}_k \\ \boldsymbol{q}_k &= \frac{\boldsymbol{u}_k}{r_{kk}} \qquad l_{kk} = \|\boldsymbol{u}_k\| \end{aligned} \tag{6.14}$$

である．(6.14) 式が分解 $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{L}$ を表している．これを修正グラム・シュミット法の形式に直すのは簡単である．

ハウスホールダー変換でも QL 分解を行うことができる．このときにははじめに m 列の第 1 から第 $n-1$ 成分が 0 になるような変換を行う．後ろの列から順次右上に 0 を導入すれば (6.12) 式に相当する式は

$$\boldsymbol{P}_m \boldsymbol{P}_{m-1} \cdots \boldsymbol{P}_1 \boldsymbol{A} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{L} \end{bmatrix}$$

になる．

固有値問題と相似変換　以下では固有値問題を解くことになるが，その際に基本となるのは相似変換である．固有値問題であるから，考えるのは $n \times n$ の正方行列である．

正方行列 $\boldsymbol{A}$ に対して

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x} \tag{6.15}$$

を満足する数 λ が行列 $\boldsymbol{A}$ の固有値，$\boldsymbol{x}$ が固有値 λ に属する固有ベクトルである．固有値は行列式

$$\det(\lambda \boldsymbol{I} - \boldsymbol{A}) = 0$$

の根としても求められる．

適当な正則行列 $\boldsymbol{P}$ を用いて相似変換

$$\boldsymbol{B} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} \tag{6.16}$$

を行うと，行列 $\boldsymbol{A}$ と $\boldsymbol{B}$ の固有値は等しい．なぜなら (6.15) 式に左から $\boldsymbol{P}^{-1}$ を掛けると

$$\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}\boldsymbol{P}^{-1}\boldsymbol{x} = \lambda\boldsymbol{P}^{-1}\boldsymbol{x}$$

よって

$$\boldsymbol{B}\boldsymbol{y} = \lambda\boldsymbol{y} \quad \boldsymbol{y} = \boldsymbol{P}^{-1}\boldsymbol{x}$$

となるからである．そこで $\boldsymbol{B}$ が三角行列になるような変換 $\boldsymbol{P}$ が求まれば，$\boldsymbol{A}$ の固有値が $\boldsymbol{B}$ の対角要素から求められることになる．また $\boldsymbol{A}$ の固有値ベクトルは $\boldsymbol{B}$ の固有ベクトル $\boldsymbol{y}$ から

$$\boldsymbol{x} = \boldsymbol{P}\boldsymbol{y}$$

によって求められる．

三重対角化（対称行列） ハウスホールダー変換

$$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{w}\boldsymbol{w}^T \quad \|\boldsymbol{w}\|^2 = 2$$

の $\boldsymbol{w}$ として今度は第一成分が 0 の

$$\boldsymbol{w} = [0, *, *, \cdots, *]^T$$

の形のものを用いる．$*$ は任意の数を表している．このとき $\boldsymbol{P}$ の一行，一列は (1, 1) 成分が 1 であるほかは 0 になる．したがってこの $\boldsymbol{P}$ を左から掛けても一行目は変化せず，右から掛けても一列目は変化しない．この性質を用いて対称行列 $\boldsymbol{A}$ に相似変換を施した結果 $\boldsymbol{B} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}$ を

$$\boldsymbol{B} = \begin{bmatrix} * & * & 0 & \cdots & 0 \\ * & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \cdots & * \end{bmatrix}$$

の形にすることができる．

$\boldsymbol{A}$ の第一列からなるベクトルを $\boldsymbol{a}_1$ とする．ハウスホールダー変換 $\boldsymbol{P}_1$ を用いて上のような形になるためには $\boldsymbol{P}_1\boldsymbol{a}_1$ が

$$\boldsymbol{P}_1\boldsymbol{a}_1 = \boldsymbol{a}_1 - \boldsymbol{w}(\boldsymbol{w}^T\boldsymbol{a}_1) = [a_{11}, s, 0, \cdots, 0]^T$$

の形にならなければならない．上式に左から $\boldsymbol{a}_1^T$ を掛けた式と，直交変換ではノルムが変化しないという式を書けば

$$\|\boldsymbol{a}_1\|^2 - (\boldsymbol{w}^T\boldsymbol{a}_1)^2 = a_{11}^2 + sa_{21}$$
$$\|\boldsymbol{a}_1\|^2 = a_{11}^2 + s^2$$

が成り立つ．a_{11} と a_{21} は $\boldsymbol{a}_1$ の第一，二成分である．これらの式から

$$s = \mathrm{sign}(-a_{21})\sqrt{\sum_{i=2}^{n} a_{i1}^2}$$
$$\boldsymbol{w}^T\boldsymbol{a}_1 = \sqrt{s(s - a_{21})} \tag{6.17}$$

が得られる．よって $\boldsymbol{w}$ は

$$\boldsymbol{w} = \frac{1}{\boldsymbol{w}^T\boldsymbol{a}_1}[0,\, a_{21} - s,\, a_{31}, \cdots,\, a_{n1}]^T \tag{6.18}$$

とすればよい．$\boldsymbol{w}$ の計算で桁落ちが起きないように s の符号は a_{21} の符号と反対に選ぶ．また相似変換を行うときに $\boldsymbol{P}_1$ の成分を求めずに，(6.11) 式の形，すなわち

$$\boldsymbol{P}_1 = \boldsymbol{I} + \frac{\boldsymbol{u}\boldsymbol{u}^T}{s(a_{21} - s)}$$
$$\boldsymbol{u}^T = [0, a_{21} - s, a_{31}, \cdots, a_{n1}]$$

を用いて計算を行うのは QR 分解のときと同じである．

このようにして決まった $\boldsymbol{P}_1$ を用いて相似変換 $\boldsymbol{B} = \boldsymbol{P}_1\boldsymbol{A}\boldsymbol{P}_1$ を行うと $\boldsymbol{B}$ の一行，一列は

$$b_{11} = a_{11} \quad b_{12} = b_{21} = s$$

以外はすべて 0 になる．次に一行，一列を無視して二行，二列以下の正方行列に対して同じことを行えば二行，二列の部分に 0 が導入される．このようにして $\boldsymbol{A}$ が $n \times n$ のときには $n - 2$ 回の相似変換によって $\boldsymbol{A}$ は三重対角行列

$$\boldsymbol{B} = \boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P}$$

$$= \begin{bmatrix} \alpha_1 & \beta_1 & 0 & & & \\ \beta_1 & \alpha_2 & \beta_2 & 0 & & \\ 0 & \beta_2 & \alpha_2 & \beta_3 & \ddots & \\ & 0 & \ddots & \ddots & \ddots & 0 \\ & & \ddots & \ddots & \alpha_{n-1} & \beta_{n-1} \\ 0 & & & 0 & \beta_{n-2} & \alpha_n \end{bmatrix}$$

$$\boldsymbol{P} = \boldsymbol{P}_1\boldsymbol{P}_2\cdots\boldsymbol{P}_{n-2} \tag{6.19}$$

に変換される．

ヘッセンベルグ行列　対称行列を三重対角化する方法を非対称行列に適用したらどうなるか．$\boldsymbol{P}_1\boldsymbol{A}$ で 1 列目に 0 が導入されることは同じであるが，右から $\boldsymbol{P}_1$ を掛けたときには非対称性のために 1 行目に 0 が導入されない．したがって最後まで相似変換を続けた結果は三重対角行列 (6.19) 式ではなく

$$\boldsymbol{B} = \begin{bmatrix} * & * & \cdots & \cdots & * \\ * & * & \cdots & \cdots & * \\ 0 & * & \cdots & \cdots & * \\ & \ddots & \ddots & \ddots & \vdots \\ 0 & & 0 & * & * \end{bmatrix} \tag{6.20}$$

の形，すなわち対角要素より一段下の成分までが非零になる．このような形の行列をヘッセンベルグ行列という．

後で述べる非対称行列の固有値を求める QR 法などでは，もとの行列を一旦ヘッセンベルグ型に変換してから行った方が計算が効率的になる．

二重対角化　$\boldsymbol{A}$ の 1 列目の第二成分以下を 0 にするハウスホールダー変換 $\boldsymbol{P}_1$(6.10) 式を用いて

$$\boldsymbol{B} = \boldsymbol{P}_1\boldsymbol{A}$$

を計算する．次に $\boldsymbol{B}$ の一行目の行ベクトル $\boldsymbol{b}_1^T$ を用いて

$$\boldsymbol{b}_1^T\boldsymbol{Q}_1 = [b_{11}, s, 0, \cdots, 0]$$

が成り立つようなハウスホールダー変換 $\boldsymbol{Q}_1$ を求める．これは (6.17)，(6.18) 式から求めることができる．この行列を $\boldsymbol{B}$ の右から掛けても一列目は変化しないので，結果は

$$\boldsymbol{P}_1\boldsymbol{A}\boldsymbol{Q}_1 = \begin{bmatrix} * & * & 0 & \cdots & 0 \\ 0 & * & \cdots & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & \cdots & \cdots & * \end{bmatrix}$$

の形になる．このような操作を続ければ $\boldsymbol{A}$ は

$$\boldsymbol{P}^T\boldsymbol{A}\boldsymbol{Q} = \begin{bmatrix} * & * & 0 & \cdots & 0 \\ 0 & * & * & \cdots & 0 \\ \vdots & 0 & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & * \\ 0 & \cdots & \cdots & 0 & * \end{bmatrix} \tag{6.21}$$

$$\boldsymbol{P} = \boldsymbol{P}_1\boldsymbol{P}_2\cdots\boldsymbol{P}_{n-1}$$

$$\boldsymbol{Q} = \boldsymbol{Q}_1\boldsymbol{Q}_2\cdots\boldsymbol{Q}_{n-1}$$

の形，すなわち対角成分とその上の成分以外は 0 の二重対角の形に変換される．ただしこの変換は相似変換ではないから，この行列から固有値を求めることはできない．この変換は特異値分解のときに利用される．

ヤコビ法（対称行列）　対称行列の固有値を求めるには，二次元の座標回転に相当する相似変換を繰り返して $\boldsymbol{A}$ を対角化する方法が，考え方としては最も簡単である．いま $\boldsymbol{P}$ として

$$\boldsymbol{P} = \begin{bmatrix} 1 & 0 & \cdots & & & & & 0 \\ 0 & 1 & \cdots & & & & & \\ \vdots & \vdots & \ddots & & & & & \\ & & & c & \cdots & s & & \\ & & & \vdots & \ddots & \vdots & & \\ & & & -s & \cdots & c & & \\ & & & & & & \ddots & 0 \\ 0 & & & & & & 0 & 1 \end{bmatrix} \tag{6.22}$$

と置く．ここに

$$p_{kk} = p_{ll} = c = \cos\varphi$$

$$p_{kl} = -p_{lk} = s = \sin\varphi$$

で，ほかの要素は単位行列と同じである．この行列が直交行列

$$\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$$

であること，またこの行列による相似変換

$$\boldsymbol{B} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P}$$

によって行列の対称性が保存されることもわかる．

いま

$$\boldsymbol{B}' = \boldsymbol{P}^T\boldsymbol{A} \quad \boldsymbol{B} = \boldsymbol{B}'\boldsymbol{P}$$

と置くと，$\boldsymbol{B}'$ は k, l 行だけが変化し，$\boldsymbol{B}$ は k, l 列だけが変化する．$\boldsymbol{B}'$，$\boldsymbol{B}$ の成分は次のように表される．

$$\begin{cases} b'_{kj} = a_{kj}\cos\varphi - a_{lj}\sin\varphi \\ b'_{lj} = a_{kj}\sin\varphi + a_{lj}\cos\varphi \end{cases} \quad 1 \le j \le n$$
$$\begin{cases} b_{ik} = b'_{ik}\cos\varphi - b'_{il}\sin\varphi \\ b_{il} = b'_{ik}\sin\varphi + b'_{il}\cos\varphi \end{cases} \quad 1 \le i \le n \tag{6.23}$$

b'_{ik}，b'_{il} は実は，先の性質により，$i = k, l$ 以外は $\boldsymbol{A}$ の成分と同じである．

$\boldsymbol{B}$ の (k, l) 成分をもとの成分で表すと

$$\begin{aligned} b_{kl} = b_{lk} &= (a_{kk} - a_{ll})\sin\varphi\cos\varphi \\ &+ a_{kl}(\cos^2\varphi - \sin^2\varphi) \end{aligned} \tag{6.24}$$

であるから，これが 0 になるように φ を選べば (k, l)，(l, k) 成分が消去できる．次に $\boldsymbol{B}$ の別の非対角要素を選んでこれを 0 にするという操作を続ければ，最終的に $\boldsymbol{A}$ を対角行列 $\boldsymbol{D}$ に変換することができるであろう．ただし，一旦 0 にした非対角要素も，次の変換で 0 でなくなってしまうから，この方法は有限回の反復で解が求まる方法ではない．

このような方法で本当に $\boldsymbol{A}$ が対角行列に変換できるかという疑問が生じる．これは簡単に証明できる．一般に直交変換 $\boldsymbol{B} = \boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P}$ では行列の二乗和が保存される．すなわち

$$\sum_{i,j} b_{ij}^2 = \sum_{i,j} a_{ij}^2$$

が成り立つ．一方，ヤコビ法の変換 (6.23) 式では

$$b_{kk}^2 + 2b_{kl}^2 + b_{ll}^2 = a_{kk}^2 + 2a_{kl}^2 + a_{ll}^2$$

が成り立っているが，b_{kl} が 0 になるように φ を選ぶのがヤコビ法であるから

$$b_{kk}^2 + b_{ll}^2 > a_{kk}^2 + a_{ll}^2 \qquad a_{kl} \neq 0$$

が成り立つ．すなわち，一回の変換で対角成分が増加し，その分だけ非対角成分が減少するから，この反復は収束する．

回転角 φ は (6.24) 式の $b_{kl} = 0$ から

$$\tan 2\varphi = \frac{2a_{kl}}{a_{ll} - a_{kk}}$$

を満たさなければならないが，必要なのは φ ではなく，$\cos\varphi$，$\sin\varphi$ であるから，これらを次のようにして直接求めた方がよい．$b_{kl} = 0$ の式から

$$t^2 + \frac{a_{ll} - a_{kk}}{a_{kl}}t - 1 = 0 \qquad t = \tan\varphi$$

よって

$$t = \frac{a_{kk} - a_{ll}}{2a_{kl}} \pm \sqrt{\left(\frac{a_{kk} - a_{ll}}{2a_{kl}}\right)^2 + 1} \tag{6.25}$$

でなければならない．二つの根のうち絶対値の大きな方を選ぶものとする．そうすると

$$c = \frac{1}{\sqrt{1 + t^2}} \qquad s = ct$$

によって $\boldsymbol{P}$ の要素が決まる．

数値計算のために (6.23) 式，(6.25) 式を書きかえる．非対角成分 a_{kl} は 0 に近づけていくのであるから (6.25) 式に含まれる

$$\tau = \frac{a_{ll} - a_{kk}}{2a_{kl}}$$

は $\pm\infty$ に近づいていく．そのときにも t として絶対値の大きな方を選ぶと，s の計算で $0 \times \infty$ が現れる．これを避けるためには逆に絶対値の小さい方の根を選ぶ．たとえば $\tau > 0$ のときには根と係数の関係から

$$t = \frac{1}{\tau + \sqrt{\tau^2 + 1}} \qquad \tau > 0$$

とする．$\tau < 0$ のときも含めると

$$t = \frac{\mathrm{sign}(-\tau)}{|\tau|(1 + \sqrt{1 + 1/\tau^2})}$$

とすればよい．sign は符号関数である．

次に (6.23) 式の対角成分は，たとえば

$$b_{kk} = c^2 a_{kk} + s^2 a_{ll} - 2csa_{kl}$$

となる．ところで c や s は

$$a_{ll} - a_{kk} = \frac{c^2 - s^2}{cs}a_{kl}$$

になるように決めたものである．上式を用いて a_{ll} を消去し，同様に b_{ll} からは a_{kk} を消去すると

$$b_{kk} = a_{kk} - ta_{kl} \qquad b_{ll} = a_{ll} + ta_{kl} \tag{6.26}$$

が得られる．このような書きかえをしたのは，もとの量＋補正量という形にして丸め誤差を少なくした

かったからである．同様に (6.23) 式の非対角項も

$$\begin{aligned} b_{kj} &= a_{kj} - s(a_{lj} + \gamma a_{kj}) \\ b_{lj} &= a_{lj} + s(a_{kj} - \gamma a_{lj}) \\ \gamma &= \frac{s}{1+c} \qquad j \neq k, l \end{aligned} \tag{6.27}$$

と書きかえられる．

次に 0 にすべき要素 a_{kl} をどのように選ぶかという問題がある．非対角要素の中で絶対値が最大の要素を (k, l) に選ぶのが合理的なように思えるが，これを見つけるのには $n^2/2$ 回の比較が必要になる．毎回このような比較をするよりは，非対角要素を端から順に，すなわち $(k,l) = (1,2),(1,3)\cdots(1,n),(2,3),(2,4)\cdots$ のように機械的に選んだ方が手っ取り早い．もちろん，既に小さくなった要素を 0 にするのは無駄であるから，閾値よりも小さな要素は飛ばすというような工夫も必要である．

(6.23) 式の形の直交変換 $\boldsymbol{P}_1, \boldsymbol{P}_2, \cdots$ によって行列 $\boldsymbol{A}$ が対角化されたとする．すべての積を改めて

$$\boldsymbol{P} = \boldsymbol{P}_1 \boldsymbol{P}_2 \cdots$$

と書くと

$$\boldsymbol{P}^T \boldsymbol{A} \boldsymbol{P} = \boldsymbol{D} \qquad \boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{D}$$

が成り立つ．$\boldsymbol{D}$ は固有値 λ_j を対角要素とする対角行列である．$\boldsymbol{P}$ の j 列目のベクトルを $\boldsymbol{p}_j$ とすれば，最後の式は

$$\boldsymbol{A}\boldsymbol{p}_j = \boldsymbol{p}_j \lambda_j$$

である．したがって変換行列の積を計算しておけば，その列ベクトルが固有ベクトルである．

三重対角行列の固有値　先にハウスホールダー変換を用いた相似変換によって，実対称行列を三重対角化する方法を述べた．このように三重対角された行列 (6.19) 式から行列式

$$p_n(\lambda) \equiv \det(\lambda \boldsymbol{I} - \boldsymbol{B})$$

を作ると，これは λ の n 次多項式になる．この多項式の零点がもとの行列 $\boldsymbol{A}$ の固有値にほかならない．

そこでこの行列式の左上の $k \times k$ の主行列式の値を $p_k(\lambda)$ とする．たとえば

$$\begin{aligned} p_1(\lambda) &= \lambda - \alpha_1 \\ p_2(\lambda) &= \begin{vmatrix} \lambda - \alpha_1 & -\beta_1 \\ -\beta_1 & \lambda - \alpha_2 \end{vmatrix} \end{aligned}$$

等である．一般に

$$p_k(\lambda) = \begin{vmatrix} \lambda - \alpha_1 & -\beta_1 & 0 & & & 0 \\ -\beta_1 & \lambda - \alpha_2 & -\beta_2 & 0 & & \\ 0 & -\beta_2 & \lambda - \alpha_3 & -\beta_3 & \ddots & \\ & \ddots & \ddots & \ddots & \ddots & 0 \\ & & \ddots & \ddots & \lambda - \alpha_{k-1} & -\beta_{k-1} \\ 0 & & & 0 & -\beta_{k-1} & \lambda - \alpha_k \end{vmatrix} \tag{6.28}$$

である．これを最後の行あるいは列について展開すれば，漸化式

$$p_k(\lambda) = (\lambda - \alpha_k) p_{k-1}(\lambda) - \beta_{k-1}^2 p_{k-2}(\lambda) \tag{6.29}$$

が得られる．これが $k = 2$ に対しても成立するように

$$p_0(\lambda) = 1$$

と定義しておく．固有値を求めるために $p_k(\lambda)$ の微分が必要になるが，これは漸化式

$$\begin{aligned} p_k'(\lambda) = (\lambda - \alpha_k) p_{k-1}'(\lambda) - \beta_{k-1}^2 p_{k-2}'(\lambda) \\ + p_{k-1}(\lambda) \end{aligned} \tag{6.30}$$

によって計算できる．

$n \times n$ 実対称行列 $\boldsymbol{A}$ の固有値は上のようにして求めた n 次多項式 $p_n(\lambda)$ の零点である．実対称行列の固有値はすべて実数であるから，固有値の存在範囲がわかればニュートン法などで零点を求めるの

は容易である．ただし §3 でも注意しておいたように，多項式 $p_n(\lambda)$ を実際に求めてしまうとその係数の誤差が大きくなってしまう．しかし漸化式 (6.29), (6.30) 式を用いて $p_n(\lambda)$ の関数値や微係数の値を計算すればそのような心配はない．

スツルムの定理　ある λ を固定したとき，数列

$$p_n(\lambda),\, p_{n-1}(\lambda),\, \cdots,\, p_0(\lambda)$$

を左から順に見ていったときに符号が変化した回数を $N(\lambda)$ と定義すると，$N(\lambda)$ は λ よりも大きい固有値の数に等しい．これがスツルムの定理である．いま

$$a < \lambda < b \qquad p_n(a) \neq 0 \neq p_n(b)$$

のとき，この区間 $(a,\, b)$ に含まれる $p_n(\lambda)$ の零点の数が

$$N(a) - N(b) \tag{6.31}$$

で与えられることはスツルムの定理からただちに導かれる．

この定理を用いると，固有値がただ一つある区間を二分法を用いて絞り込むことは容易である．その際に，固有値の存在範囲が知られていなければならないが，行列 (6.19) の行ごとの要素の絶対値の和の最大値を，すなわち

$$r = \max_{1 \leq i \leq n} \left(|\beta_{i-1}| + |\alpha_i| + |\beta_i| \right)$$

とすると，すべての固有値は

$$-r < \lambda < r$$

の範囲にある (ゲルシェゴリンの定理)．これを最初の区間 $(a,\, b)$ とし，二分法で固有値を絞り込んだ後は，ニュートン法などで固有値の精度を上げることができる．

冪乗法　正方行列 $\boldsymbol{A}$(対称行列でなくてもよい) に対して，適切な初期ベクトル $\boldsymbol{x}_0$ から出発して

$$\boldsymbol{x}_j = \boldsymbol{A}\boldsymbol{x}_{j-1} \tag{6.32}$$

という反復を繰り返すと，$j \to \infty$ では $\boldsymbol{x}_j$ は $\boldsymbol{A}$ の絶対値最大の固有値 λ_1 に属する固有ベクトルに収束し

$$\lim_{j \to \infty} \frac{\boldsymbol{x}_j^T \boldsymbol{x}_j}{\boldsymbol{x}_j^T \boldsymbol{x}_{j-1}} = \lambda_1 \tag{6.33}$$

が成り立つ．これによって絶対値最大の固有値と固有ベクトルが，行列とベクトルの積だけで計算できる．

$n \times n$ の行列 $\boldsymbol{A}$ の固有値がすべて異なり

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$$

とし，これらに属する固有ベクトルを $\boldsymbol{u}_k$，すなわち

$$\boldsymbol{A}\boldsymbol{u}_k = \lambda_k \boldsymbol{u}_k \tag{6.34}$$

とする．固有ベクトルは互いに一次独立であるから，初期ベクトルは

$$\boldsymbol{x}_0 = c_1\boldsymbol{u}_1 + c_2\boldsymbol{u}_2 + \cdots + c_n\boldsymbol{u}_n$$

の形に展開できる．$c_1 \neq 0$ とすれば反復 (6.32) 式によって

$$\begin{aligned} \boldsymbol{x}_j = \boldsymbol{A}^j\boldsymbol{x}_0 &= c_1\lambda_1^j\boldsymbol{u}_1 + c_2\lambda_2^j\boldsymbol{u}_2 + \cdots + c_n\lambda_n^j\boldsymbol{u}_n \\ &= c_1\lambda_1^j \left[\boldsymbol{u}_1 + \frac{c_2}{c_1}\left(\frac{\lambda_2}{\lambda_1}\right)^j \boldsymbol{u}_2 + \right. \\ &\qquad \left. \cdots + \frac{c_n}{c_1}\left(\frac{\lambda_n}{\lambda_1}\right)^j \boldsymbol{u}_n \right] \end{aligned}$$

になる．この関係から

$$\lim_{j \to \infty} \frac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|} = \frac{\boldsymbol{u}_1}{\|\boldsymbol{u}_1\|}$$

と (6.33) 式が導かれる．

ここで $c_1 \neq 0$ という条件は重要である．初期ベクトル $\boldsymbol{x}_0$ に $\boldsymbol{u}_1$ 成分が含まれていなければ，$\boldsymbol{x}_j$ は次に大きい固有値 λ_2 に属する $\boldsymbol{u}_2$ に収束するはずであるが，実際には丸め誤差のために $\boldsymbol{u}_1$ 成分が生じ，徐々に成長してくる．しかしそのためには時間が必要であるから，収束は遅くなる．また，当然のことながら，λ_1 が複素数のとき，$\boldsymbol{x}_0$ を実数に選べば反復は収束しない．

実は冪乗法によって固有値や固有ベクトルを求めることはほとんどないが，次に述べる逆反復法の基礎になっている．

逆反復法 行列 $\boldsymbol{A}$ のある固有値 λ_k の近似値を μ とし

$$\boldsymbol{B} = (\mu\boldsymbol{I} - \boldsymbol{A})^{-1} \tag{6.35}$$

と置く．λ_k に属する $\boldsymbol{A}$ の固有ベクトルを $\boldsymbol{u}_k$ とすると

$$\boldsymbol{B}^{-1}\boldsymbol{u}_k = (\mu\boldsymbol{I} - \boldsymbol{A})\boldsymbol{u}_k = (\mu - \lambda_k)\boldsymbol{u}_k$$

であるから

$$\boldsymbol{B}\boldsymbol{u}_k = \frac{1}{\mu - \lambda_k}\boldsymbol{u}_k$$

が成り立つ．すなわち $\boldsymbol{B}$ の固有値の一つは $(\mu - \lambda_k)^{-1}$ である．もし近似値 μ が真値 λ_k に非常に近く，ほかのすべての固有値 λ_i に対して

$$|\mu - \lambda_k| < |\mu - \lambda_i| \quad i \neq k$$

が成り立つとすれば，$\boldsymbol{B}$ の絶対値最大の固有値は $(\mu - \lambda_k)^{-1}$ である．したがって適切な初期ベクトル $\boldsymbol{x}_0$ を用いて行列 $\boldsymbol{B}$ に対して冪乗法を行えば

$$\lim_{j\to\infty} \frac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|} = \frac{\boldsymbol{u}_k}{\|\boldsymbol{u}_k\|}$$
$$\lim_{j\to\infty} \frac{\boldsymbol{x}_j^T\boldsymbol{x}_j}{\boldsymbol{x}_j^T\boldsymbol{x}_{j-1}} = \frac{1}{\mu - \lambda_k} \tag{6.36}$$

が得られる．実際の計算では $(\mu\boldsymbol{I} - \boldsymbol{A})^{-1}$ を掛けるのではなく，連立方程式

$$(\mu\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x}_j = \boldsymbol{x}_{j-1} \tag{6.37}$$

を解くことによって $\boldsymbol{x}_{j-1}$ から $\boldsymbol{x}_j$ を求める．また，(6.36) 式のように $j \to \infty$ まで計算するのではなく，(6.37) 式を二，三回解いた後

$$\lambda_k \doteq \mu - \frac{\boldsymbol{x}_j^T\boldsymbol{x}_{j-1}}{\boldsymbol{x}_j^T\boldsymbol{x}_j}$$

によって λ_k の近似値を計算し，これを新しい μ として (6.37) 式の反復を繰り返すという方法をとった方が効率的である．このとき，同じ μ について (6.37) 式を解くときには，最初にガウスの消去法で LU 分解をしておけば，二回目以降の計算はその結果を利用して効率的に行うことができる．

なお，μ が真の固有値に近くなると $\mu\boldsymbol{I} - \boldsymbol{A}$ は特異行列に近くなるので，(6.37) 式をそのままで反復するとオーバーフローする恐れがある．そこで一回の反復ごとに右辺のベクトルを

$$\|\boldsymbol{x}_{j-1}\| = 1$$

のように正規化しておいた方がよい．$\mu\boldsymbol{I} - \boldsymbol{A}$ が本当に特異行列になったときには，そのときの μ が固有値 λ_k である．

逆反復法は固有値と固有ベクトルの近似値がわかっているときに，解の精度を上げるためによく用いられる．固有ベクトルの近似値がわかっていないときでも，乱数を用いて初期ベクトルを生成してやればほとんどの場合うまくいく．

$\boldsymbol{QR}$ 法 行列 $\boldsymbol{A}$ を QR 分解したとき，$\boldsymbol{A}$ を $\boldsymbol{Q}$ によって相似変換すると

$$\boldsymbol{Q}^{-1}\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{Q}^{-1}(\boldsymbol{Q}\boldsymbol{R})\boldsymbol{Q} = \boldsymbol{R}\boldsymbol{Q}$$

となるから，$\boldsymbol{A}$ と $\boldsymbol{RQ}$ の固有値は等しい．そこで $\boldsymbol{A}$ を QR 分解し

$$\boldsymbol{A}_1 = \boldsymbol{A} = \boldsymbol{Q}_1\boldsymbol{R}_1 \quad \boldsymbol{A}_2 = \boldsymbol{R}_1\boldsymbol{Q}_1$$

を作り，さらに $\boldsymbol{A}_2$ を QR 分解する．一般に分解と積を

$$\boldsymbol{A}_k = \boldsymbol{Q}_k\boldsymbol{R}_k \quad \boldsymbol{A}_{k+1} = \boldsymbol{R}_k\boldsymbol{Q}_k \tag{6.38}$$

で反復する．かなり一般的な条件の下で $\boldsymbol{A}_k$ の対角成分に $\boldsymbol{A}$ の固有値が上から絶対値の大きな順に並ぶか (対称行列のとき)，対角線上の 2×2 の小行列の固有値がもとの行列の固有値になる．

$\boldsymbol{RQ}$ は $\boldsymbol{A}$ の $\boldsymbol{Q}$ による直交変換であるため，$\boldsymbol{A}$ の対称性，三重対角性，ヘッセンベルグ性は保存される．したがって $\boldsymbol{A}$ が実対称行列のときには三重対角化し，非対称のときにはヘッセンベルグ化してから QR 法を行えば計算が非常に効率化される．

原点移動（対称行列） QR 法の収束の速さは冪乗法でもみられたように，固有値の絶対値の比で決まるので，単純な QR 法では収束が遅すぎることがある．収束を加速するためには原点移動を行う．

$\boldsymbol{A}$ のある固有値 λ の近似値を v とするとき，$\boldsymbol{A} - v\boldsymbol{I}$ の固有値は $\lambda - v$ であるから，$\boldsymbol{A} - v\boldsymbol{I}$ に対して QR 法を行えば右下の対角成分は速く $\lambda - v$ に収束するであろう．$\lambda - v$ の絶対値が小さければ小さいほど収束は速くなる．

そこで次のような反復を行う．

$$\begin{aligned}\boldsymbol{A}_k - v_k\boldsymbol{I} &= \boldsymbol{Q}_k\boldsymbol{R}_k \\ \boldsymbol{A}_{k+1} &= \boldsymbol{R}_k\boldsymbol{Q}_k + v_k\boldsymbol{I}\end{aligned} \tag{6.39}$$

$\boldsymbol{A}_k$ を QR 分解するのではなく，$\boldsymbol{A}_k - v_k\boldsymbol{I}$ を QR 分解している．第一式から

$$\boldsymbol{R}_k\boldsymbol{Q}_k = \boldsymbol{Q}_k^T\boldsymbol{A}_k\boldsymbol{Q}_k - v_k\boldsymbol{I}$$

であるから，これを第二式に代入すれば

$$\boldsymbol{A}_{k+1} = \boldsymbol{Q}_k^T\boldsymbol{A}_k\boldsymbol{Q}_k$$

となる．すなわち，$\boldsymbol{A}_{k+1}$ は $\boldsymbol{A}_k$ の相似変換になっている．したがって $k \to \infty$ における $\boldsymbol{A}_k$ の振舞いは単純な QR 法と同様であるが，v_k の選び方によって収束が速くなる．

v_k は任意であるが，対称行列のときには $\boldsymbol{A}_k$ の第 (n,n) 成分が絶対値最小の固有値に近づくことから，この成分を v_k に選ぶのが効果的であろう．このようにして，第 n 行で (n,n) 成分以外が十分に小さくなれば，これを絶対値最小の固有値 λ_n とみなすことができる．λ_n が求まったときには第 n 行，n 列を除いた $(n-1)\times(n-1)$ 行列に対して同じ手続きを繰り返せばよい．このように固有値が一つ求められるたびに行列の次数を下げて計算を進めることができる．ただし非対称行列のときには固有値が複素数になるから，v_k を複素数にして複素数計算をしなければならないのでこの方法は適当ではない．

ヘッセンベルグ行列に関する補題　非対称行列の固有値を実数計算だけで求める方法を事項に述べるが，その際に必要になる補助定理を示しておく．

ある与えられた正則行列 $\boldsymbol{B}$ が

$$\boldsymbol{B}\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{H} \tag{6.40}$$

の形に書くことができたとする．ここに $\boldsymbol{Q}$ は直交行列，$\boldsymbol{H}$ はヘッセンベルグ行列である．このとき $\boldsymbol{Q}$ の第一列目が与えられれば $\boldsymbol{Q}$ の残りの列，および $\boldsymbol{H}$ を求めることができる．もし $\boldsymbol{H}$ の対角下の要素をすべて正とするならば，$\boldsymbol{Q}$ と $\boldsymbol{H}$ は一義的に決まる．

これを証明するには，与えられた条件で実際に $\boldsymbol{Q}$ と $\boldsymbol{H}$ を作って見せればよい．そこでまず $\boldsymbol{Q}$ の各列を $\boldsymbol{q}_j$ として

$$\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_n]$$

と書くと，(6.40) 式は

$$\begin{aligned}&\boldsymbol{B}\,[\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots] \\ &\quad = [\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots]\begin{bmatrix} h_{11} & h_{12} & \cdots \\ h_{21} & h_{22} & \cdots \\ 0 & h_{32} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}\end{aligned} \tag{6.41}$$

の形に書かれることになる．この式の両辺の一列目は

$$\boldsymbol{B}\boldsymbol{q}_1 = \boldsymbol{q}_1 h_{11} + \boldsymbol{q}_2 h_{21}$$

であるが，$\boldsymbol{q}_j$ が正規直交系であることから

$$\boldsymbol{q}_1^T\boldsymbol{B}\boldsymbol{q}_1 = h_{11}$$

が成り立つ．仮定により $\boldsymbol{q}_1$ が与えられているから，上式より h_{11} が決まる．h_{11} が決まると

$$\boldsymbol{q}_2 h_{21} = \boldsymbol{B}\boldsymbol{q}_1 - \boldsymbol{q}_1 h_{11}$$

の右辺は既知になる．両辺の二乗から

$$h_{21}^2 = \|\boldsymbol{B}\boldsymbol{q}_1 - \boldsymbol{q}_1 h_{11}\|^2$$

が得られるので，正の根を選ぶことにし

$$h_{21} = \|\boldsymbol{B}\boldsymbol{q}_1 - \boldsymbol{q}_1 h_{11}\|$$

とすれば h_{21} が決まり

$$\boldsymbol{q}_2 = \frac{1}{h_{21}}(\boldsymbol{B}\boldsymbol{q}_1 - \boldsymbol{q}_1 h_{11})$$

から $\boldsymbol{q}_2$ が決まる．

(6.41) 式の二列目は

$$\boldsymbol{B}\boldsymbol{q}_2 = \boldsymbol{q}_1 h_{12} + \boldsymbol{q}_2 h_{22} + \boldsymbol{q}_3 h_{32}$$

であるから，既に決まっている $\boldsymbol{q}_1, \boldsymbol{q}_2$ を左から掛けることにより

$$h_{12} = \boldsymbol{q}_1^T\boldsymbol{B}\boldsymbol{q}_2 \qquad h_{22} = \boldsymbol{q}_2^T\boldsymbol{B}\boldsymbol{q}_2$$

より h_{12}, h_{22} が決まり，これらを用いて

$$\begin{aligned}h_{32} &= \|\boldsymbol{B}\boldsymbol{q}_2 - \boldsymbol{q}_1 h_{12} - \boldsymbol{q}_2 h_{22}\| \\ \boldsymbol{q}_3 &= \frac{1}{h_{32}}(\boldsymbol{B}\boldsymbol{q}_2 - \boldsymbol{q}_1 h_{12} - \boldsymbol{q}_2 h_{22})\end{aligned}$$

から h_{32} と $\boldsymbol{q}_3$ が決まる．以下同様にしていけば $\boldsymbol{Q}$ のすべての列と，$\boldsymbol{H}$ のすべての要素が決まることは明らかであろう．

二段 QR 法（フランシスの方法） 要素が実数でも，非対称行列の固有値は一般には複素数であるが，必ず複素共役の組になっている．このことを利用すると実数計算だけで複素固有値を求めることができる．

まず，行列 $\boldsymbol{A}$ をヘッセンベルグ型に変換し，これを改めて $\boldsymbol{A}$ とする．これに対して次のように二段階の QR 法を行う．

$$\begin{aligned}\boldsymbol{A}_k - v_k\boldsymbol{I} &= \boldsymbol{Q}_k^T\boldsymbol{R}_k \\ \boldsymbol{A}_{k+1} &= \boldsymbol{R}_k\boldsymbol{Q}_k + v_k\boldsymbol{I}\end{aligned} \tag{6.42}$$

$$\begin{aligned}\boldsymbol{A}_{k+1} - v_{k+1}\boldsymbol{I} &= \boldsymbol{Q}_{k+1}^T\boldsymbol{R}_{k+1} \\ \boldsymbol{A}_{k+2} &= \boldsymbol{R}_{k+1}\boldsymbol{Q}_{k+1} + v_{k+1}\boldsymbol{I}\end{aligned} \tag{6.43}$$

QR 分解ではなく Q^TR 分解になっているのは後での計算の便宜上である．はじめにいくつかの関係式を導いておく．(6.42)，(6.43) 式からそれぞれ前に導いたと同じ関係

$$\begin{aligned}\boldsymbol{A}_{k+1} &= \boldsymbol{Q}_k\boldsymbol{A}_k\boldsymbol{Q}_k^T \\ \boldsymbol{A}_{k+2} &= \boldsymbol{Q}_{k+1}\boldsymbol{A}_{k+1}\boldsymbol{Q}_{k+1}^T\end{aligned}$$

が成り立っていることに注意する．$\boldsymbol{A}_{k+1}$ の式を (6.43) 式の第一式に代入すると

$$\boldsymbol{A}_k - v_{k+1}\boldsymbol{I} = \boldsymbol{Q}_k^T\boldsymbol{Q}_{k+1}^T\boldsymbol{R}_{k+1}\boldsymbol{R}_k$$

が得られる．この式と (6.42) 式の第一式から

$$\begin{aligned}\boldsymbol{M} &\equiv (\boldsymbol{A}_k - v_{k+1}\boldsymbol{I})(\boldsymbol{A}_k - v_k\boldsymbol{I}) \\ &= \boldsymbol{Q}^T\boldsymbol{R} \\ \boldsymbol{Q} &= \boldsymbol{Q}_{k+1}\boldsymbol{Q}_k \qquad \boldsymbol{R} = \boldsymbol{R}_{k+1}\boldsymbol{R}_k\end{aligned} \tag{6.44}$$

が得られる．この $\boldsymbol{Q}$ を用いると

$$\boldsymbol{A}_{k+2} = \boldsymbol{Q}\boldsymbol{A}_k\boldsymbol{Q}^T$$

が成り立つ．すなわち $\boldsymbol{A}_{k+2}$ は $\boldsymbol{A}_k$ の相似変換である．

上の関係を

$$\boldsymbol{A}_k\boldsymbol{Q}^T = \boldsymbol{Q}^T\boldsymbol{A}_{k+2}$$

と書きかえる．先の補助定理によれば $\boldsymbol{A}_{k+2}$ がヘッセンベルグ行列になるためには $\boldsymbol{Q}^T$ の第一列，すなわち $\boldsymbol{Q}$ の第一行がわかればよい．ところで (6.44) 式によれば，$\boldsymbol{Q}$ は $\boldsymbol{M}$ を上三角行列に変換する直交行列である．これは (6.9)，(6.10) 式のタイプのハウスホールダー変換 $\boldsymbol{P}_k$ を用いて

$$\boldsymbol{Q} = \boldsymbol{P}_{n-1}\boldsymbol{P}_{n-2}\cdots\boldsymbol{P}_1$$

と表すことができる．$\boldsymbol{Q}$ の第一行は，ハウスホールダー変換の性質により $\boldsymbol{P}_1$ の第一行に等しい．そこで $\boldsymbol{P}_1$ を求める．

$\boldsymbol{P}_1$ は $\boldsymbol{M}$ の一列目に 0 を導入する変換である．$\boldsymbol{A}_k$ がヘッセンベルグ行列であることから，$\boldsymbol{M}$ の第一列は三成分しかなく，$\boldsymbol{A}_k$ の成分を a_{ij} とすればそれらは

$$\begin{aligned}m_{11} = p_1 &= a_{11}^2 - (v_k + v_{k+1})a_{11} + a_{12}a_{21} \\ &\quad + v_kv_{k+1} \\ m_{21} = q_1 &= a_{21}(a_{11} + a_{22} - v_k - v_{k+1}) \\ m_{31} = r_1 &= a_{21}a_{32}\end{aligned} \tag{6.45}$$

である．これを用いてハウスホールダー変換 (6.10) 式を計算すれば

$$\begin{aligned}s &= \mathrm{sign}(-p_1)\sqrt{p_1^2 + q_1^2 + r_1^2} \\ \boldsymbol{w} &= \frac{1}{\sqrt{s(s-p_1)}}[p_1 - s, q_1, r_1, 0, \cdots, 0]^T\end{aligned} \tag{6.46}$$

が得られる．$\boldsymbol{w}$ は最初の三成分以外は 0 である．これで $\boldsymbol{P}_1$ が求まった．

補助定理の証明で示したように，$\boldsymbol{A}_{k+1}$ を求めるには実は $\boldsymbol{P}_1$ の第一行だけで十分であるが，$\boldsymbol{A}_{k+1}$ は $\boldsymbol{A}_k$ の相似変換であるから，$\boldsymbol{P}_1$ を用いて $\boldsymbol{A}_k$ を相似変換してみる．まず $\boldsymbol{P}_1$ を左から掛けると

$$\boldsymbol{P}_1\boldsymbol{A}_k = \begin{bmatrix} * & * & * & \cdots \\ * & * & * & \cdots \\ * & * & * & \cdots \\ 0 & 0 & * & \cdots \\ 0 & 0 & 0 & * \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

となって一列目の第二成分以下が 0 にならない．これは当然で，$\boldsymbol{P}_1$ は $\boldsymbol{M}$ の第一列に 0 を導入するように作られているからである．さらに右から $\boldsymbol{P}_1$ を

掛けると

$$\boldsymbol{P}_1\boldsymbol{A}_k\boldsymbol{P}_1 = \begin{bmatrix} * & * & * & \cdots \\ p_2 & * & * & \cdots \\ q_2 & * & * & \cdots \\ r_2 & * & * & \cdots \\ 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

の形になる．そこで次の相似変換では一列目の q_2, r_2 の部分を 0 にするハウスホールダー変換 $\boldsymbol{P}_2$ を用いる．これは

$$\boldsymbol{P}_2 = \left[\begin{array}{c|c} 1 & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{I}_{n-1} - \boldsymbol{w}\boldsymbol{w}^T \end{array}\right]$$

の形をしている．$\boldsymbol{w}\boldsymbol{w}^T$ は (6.46) 式の p_1, q_1, r_1 を p_2, q_2, r_2 で置き換えたものである．このように前の列の非ヘッセンベルグ成分を 0 にするような変換を $n-1$ 回行えばヘッセンベルグ行列 $\boldsymbol{A}_{k+2}$ が求められる．最終的な変換直交行列は

$$\boldsymbol{Q} = \boldsymbol{P}_{n-1}\boldsymbol{P}_{n-2}\cdots\boldsymbol{P}_1$$

と表されるが，この $\boldsymbol{Q}$ は必要ない．それどころか QR 分解そのものも必要ない．この計算は非常に速い．$\boldsymbol{P}_k$ を掛けるときには，ある列あるいは行に対して，三つの成分しか相手にしないからである．

シフト量 v_k, v_{k+1} としては $\boldsymbol{A}_k$ の右下の 2×2 行列の固有値を用いる．これらは固有値方程式

$$\begin{vmatrix} \lambda - a_{n-1,n-1} & -a_{n-1,n} \\ -a_{n,n-1} & \lambda - a_{n,n} \end{vmatrix} = 0$$

の根である．したがって (6.45) 式に現れる項は

$$v_k + v_{k+1} = a_{n-1,n-1} + a_{n,n}$$
$$v_k v_{k+1} = a_{n-1,n-1}a_{n,n} - a_{n-1,n}a_{n,n-1}$$

である．$\boldsymbol{A}$ が実数行列のときにはこれらは実数になるから，計算はすべて実数で行うことができる．

反復を繰り返しているうちに $\boldsymbol{A}_k$ の $a_{n,n-1}$ が十分小さくなったら $a_{n,n}$ が固有値である．このときには $\boldsymbol{A}_k$ の最後の行と列を除いた $(n-1)\times(n-1)$ 行列について反復を続行する．そうでなくても $a_{n-1,n-2}$ が十分小さくなったときには右下の 2×2 行列から二つの固有値が求められる．このときには二行二列を取り除くことができる．

行列のサイズが大きいときには $\boldsymbol{A}_k$ を分割することも計算量を減らすのに有効である．いま l 行の要素 $a_{l,l-1}$ が 0 とみなせるときには，図を描いてみればわかるように，l 行から n 行までと，l 行より上の $l-1$ 行はそれぞれ独立なヘッセンベルグ行列になっている．そこで下の $n-l+1$ 行について相似変換を行ってすべての固有値を求め，その後で残りの部分の変換を行えばよい．しかし分割された下の部分がさらに分割が可能になることもある．そこで簡単には次のような方法をとればよい．

一番下の n 行から上に向かって捜して，l 行の $a_{l,l-1}$ が十分小さいことがわかったとする．もし $l=n$ であれば a_{nn} が固有値である．このときには n 行，n 列を消去することができる．もし $l=n-1$ ならば右下の 2×2 小行列から二個の固有値が求められ，最後の二行，二列を消去することができる．以上のどちらでもないときには，l 行から n 行までを一つのヘッセンベルグ行列と考えて，この項の最初で述べた相似変換を最後の行まで行う．すなわち，(6.45) 式の (1,1) 成分を (l,l) 成分に読みかえればよい．たとえば

$$a_{11} \to a_{ll} \qquad a_{12} \to a_{l,l+1} \qquad a_{32} \to a_{l+2,l+1}$$

などである．これを初期値として最後の行まで相似変換を行った後，再び最後の行から上に向かって $a_{l,l-1}$ が 0 になるような行 l を捜す，という反復を繰り返す．この方法は，単に計算を効率化するだけでなく，行列全体の相似変換を繰り返したときに特定の非対角要素だけを小さくするような無限ループに陥る危険性を防ぐためにも有効である．

二段 QR 法では実行列の固有値が実数計算だけで求められる．しかし固有ベクトルまで求めようとすると，どうしても複素数計算を行わなければならないし，計算も非常に複雑になる．そこで非対称行列の場合には固有値だけは実数計算で求め，固有ベクトルが必要な場合には逆反復法で求めるのが効率的である．

特異値分解　$n\times m$ $(n \geq m)$ 行列 $\boldsymbol{A}$ は

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T \tag{6.47}$$

の形に分解することができる．ここに $\boldsymbol{U}$ は $n \times m$ の直交行列，$\boldsymbol{V}$ は $m \times m$ の直交行列，$\boldsymbol{\Lambda}$ は $m \times m$ の対角行列である．$\boldsymbol{U}$ と $\boldsymbol{V}$ は直交関係

$$\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_m \quad \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_m$$

を満足する．$\boldsymbol{V}$ は正方行列であるから

$$\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}_m$$

を満たすが，$\boldsymbol{U}$ は正方行列でないので $\boldsymbol{U}\boldsymbol{U}^T$ は単位行列にはならない．すなわち，$\boldsymbol{U}$ は半対角行列である．

分解 (6.47) 式を行列 $\boldsymbol{A}$ の特異値分解という．対角行列 $\boldsymbol{\Lambda}$ の対角要素を行列 $\boldsymbol{A}$ の特異値という．特異値は正，または 0 である．0 でない特異値の数 p を行列 $\boldsymbol{A}$ のランク (階数) という．$n \geq m$ のときは $p \leq m$ である．$p = m$ のときをフルランクという．$p < m$ のときはランク落ちがあるといい，このとき行列 $\boldsymbol{A}$ は特異行列である．特異値の最大値と最小値の比

$$\kappa(\boldsymbol{A}) = \frac{\max(\lambda_i)}{\min(\lambda_i)} \tag{6.48}$$

がこの行列の条件数である．特異値は 0 になることもあり得る．このときには条件数は無限大になり，$\boldsymbol{A}$ は特異行列である．

この分解がどのような意味をもっているかは最小二乗法の項で詳しく説明する．ここでは特異値分解の計算法だけについて述べる．

前に述べたように，$n \times m (n \geq m)$ の行列 $\boldsymbol{A}$ は列に 0 を導入する (6.9) 式タイプのハウスホールダー変換 $\boldsymbol{P}_k$ と，行に 0 を導入する (6.17) 式タイプのハウスホールダー変換 $\boldsymbol{Q}_k$ を用いて二重対角行列 $\boldsymbol{B}$(6.21) 式に変換することができる．

$$\begin{aligned} \boldsymbol{B} &= \boldsymbol{P}^T\boldsymbol{A}\boldsymbol{Q} \\ \boldsymbol{P} &= \boldsymbol{P}_1\boldsymbol{P}_2\cdots\boldsymbol{P}_m \\ \boldsymbol{Q} &= \boldsymbol{Q}_1\boldsymbol{Q}_2\cdots\boldsymbol{Q}_{m-1} \end{aligned} \tag{6.49}$$

$m = 4$ のときには $\boldsymbol{B}$ は次のような形をしている．

$$\boldsymbol{B} = \begin{bmatrix} w_1 & r_2 & 0 & 0 \\ 0 & w_2 & r_3 & 0 \\ 0 & 0 & w_3 & r_3 \\ 0 & 0 & 0 & w_4 \\ \vdots & \vdots & \vdots & 0 \\ 0 & \cdots & \cdots & 0 \end{bmatrix} \tag{6.50}$$

(6.47) 式は $\boldsymbol{U}$ と $\boldsymbol{V}$ の直交性から

$$\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{V} = \boldsymbol{\Lambda}$$

と書きかえることができるから，$\boldsymbol{B}$ を直交行列 $\boldsymbol{S}$ と $\boldsymbol{T}$ を用いて

$$\boldsymbol{S}^T\boldsymbol{B}\boldsymbol{T} = \boldsymbol{\Lambda}$$

のように対角行列に変換することができれば，分解 (6.47) 式が完成することになる．上式から直交行列 $\boldsymbol{U}$ と $\boldsymbol{V}$ が

$$\boldsymbol{U} = \boldsymbol{P}\boldsymbol{S} \quad \boldsymbol{V} = \boldsymbol{Q}\boldsymbol{T}$$

で与えられることも明らかである．

二重対角行列 $\boldsymbol{B}$ を対角行列に変換するのには，ヤコビ法と同じ考え方を用いる．はじめに一列と二列の間の回転行列

$$\boldsymbol{T}_1 = \left[\begin{array}{cc|c} \cos\varphi_1 & -\sin\varphi_1 & 0 \\ \sin\varphi_1 & \cos\varphi_1 & 0 \\ \hline \multicolumn{2}{c|}{\boldsymbol{0}} & \boldsymbol{I} \end{array}\right]$$

を $\boldsymbol{B}$ に右から掛けると次の形になる．

$$\boldsymbol{B}' = \boldsymbol{B}\boldsymbol{T} = \begin{bmatrix} * & * & 0 & 0 \\ * & * & r_3 & 0 \\ 0 & \cdots & \cdots & 0 \\ \vdots & \cdots & \cdots & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix}$$

はじめ 0 であった (2,1) 成分が 0 でなくなって $\boldsymbol{B}'$ は二重対角行列ではなくなる．そこで $\boldsymbol{T}_1$ と同じ形をした行列で φ_1 を θ_1 に変えた行列 $\boldsymbol{S}_1$ の転置を左から掛けると

$$\boldsymbol{B}'' = \boldsymbol{S}^T\boldsymbol{B}' = \begin{bmatrix} * & * & * & 0 \\ * & * & * & 0 \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

の形になってしまう．そこで二重対角行列に近づけるために，$\boldsymbol{B}''$ の (2,1) 成分が 0 になるように θ_1 を決めることにする．しかしこれでは (1,3) 成分は 0 にならない．そこで $\boldsymbol{T}_1$ を右下に一段ずらした二列と三列を回転する行列 $\boldsymbol{T}_2$ の回転角 φ_2 をうまく選んで (1,3) 成分を 0 にし，新たに生じた (3,2) 成分を $\boldsymbol{S}_2^T$ を左から掛けて 0 にするという操作を続ける

と，下図の数字の順番に非二重対角成分が現れては消えて，最後には再び二重対角行列に戻る．この操作を追い込みという．

$$\begin{bmatrix} * & * & 2 & & \\ 1 & * & * & 4 & \\ & 3 & * & * & 6 \\ & & 5 & * & * \\ & & & 7 & * \end{bmatrix}$$

一般に $\boldsymbol{T}_k$ や $\boldsymbol{S}_k$ は

$$\left[\begin{array}{c|cc|c} \boldsymbol{I}_{k-1} & & & \\ \hline & c & -s & \\ & s & c & \\ \hline & & & \boldsymbol{I}_{m-k-1} \end{array}\right]$$

の形をしており，一回の追い込みは

$$\boldsymbol{S}_{m-1}^T \boldsymbol{S}_{m-2}^T \cdots \boldsymbol{S}_1^T \boldsymbol{B} \boldsymbol{T}_1 \boldsymbol{T}_2 \cdots \boldsymbol{T}_{m-1}$$

の演算で表される．

肝心なのは最初の $\boldsymbol{T}_1$ の回転角 φ_1 をどのように決めるかということである．これを決めなければその後の回転角を決めることができない．詳しい説明は省略するが，フランシスの方法と似たようなやり方をする．すなわち $\boldsymbol{T}_1$ の第一列

$$[\cos\varphi_1,\, \sin\varphi_1,\, 0, \cdots, 0]^T$$

が

$$\boldsymbol{B}^T\boldsymbol{B} - s\boldsymbol{I}$$

の第一列に比例するように選ぶ．ここに s は原点移動量である．先の記号を用いると上式の第一列は

$$[w_1^2 - s,\, w_1 r_2,\, 0, \cdots, 0]^T$$

である．また，移動量 s は $\boldsymbol{B}^T\boldsymbol{B}$ の右下の 2×2 小行列

$$\begin{bmatrix} w_{m-1}^2 + r_{m-1}^2 & w_{m-1} r_m \\ w_{m-1} r_m & w_m^2 + r_m^2 \end{bmatrix}$$

の固有値のうちの小さい方を選ぶ．このようにして一回の追い込みが完全に定義されたことになる．追い込みを反復すると，QR 法と同様に $\boldsymbol{B}$ の非対角成分が減少して対角行列に収束する．

収束を加速するには，QR 法のときと同様に行列を分割するとよい．二重対角行列 $\boldsymbol{B}$ を右下から $r_m, w_{m-1}, r_{m-1}, w_{m-2}, \cdots$ の順に調べていき

$$r_l \neq 0 \quad \text{で} \quad w_{l-1} = 0$$

となるような l を求める．もし w_{l-1} を 0 に保ったまま r_l を 0 にするような直交変換があれば，$\boldsymbol{B}$ は l 行以下とそれより上の二つの二重対角行列に分割することができる．分割ができれば l 行以下に追い込みをかければよい．

r_l を消去するためには回転行列を左から掛ける．はじめに $l-1$ 行と l の回転行列を左から掛けて r_l，すなわち $(l-1, l)$ 成分が 0 になるように回転角を決める．そうすると新たに $(l-1, l+1)$ 成分が生じるので，次に $l-1$ 行と $l+1$ 行の回転行列を掛けて $(l-1, l+1)$ 成分を 0 にすると $(l-1, l+2)$ 成分が生じる．一般に $l-1$ 行と k 行の回転で $(l-1, k)$ 成分を消去するようにすれば，最後には l 行以下を二重対角化することができる．そこで l 行以下に追い込みをかければよい．なお，$l = m$ になったら w_m が特異値であるから，最後の一行一列を除いて再び計算を行えばよい．

簡単な例をあげる．行列

$$\boldsymbol{A} = \begin{bmatrix} 1.260 & 0.840 & 0.630 & 0.504 \\ 0.840 & 0.630 & 0.504 & 0.420 \\ 0.630 & 0.504 & 0.420 & 0.360 \\ 0.504 & 0.420 & 0.360 & 0.315 \\ 0.420 & 0.360 & 0.315 & 0.280 \end{bmatrix}$$

は対称行列の一部である．上に述べた方法で五回の追い込みで収束し，この行列の特異値が

$$\begin{bmatrix} 2.558200\text{e}{+0} \\ 1.799580\text{e}{-1} \\ 6.208599\text{e}{-3} \\ 9.967085\text{e}{-5} \end{bmatrix}$$

と求まった．したがって，この行列の条件数は (6.48) 式により

$$\kappa(\boldsymbol{A}) = 2.567 \times 10^4$$

である．これは十進 6 桁の単精度計算ではかなり悪条件の (ill-conditioned) 行列といわなければならない．

これまで $\boldsymbol{U}$ や $\boldsymbol{V}$ の計算についてはあまり触れなかったが，二重対角化の過程を含めて，行列に左からたとえば $\boldsymbol{S}^T$ を掛けたときには $\boldsymbol{U}$ に右から $\boldsymbol{S}$ を掛け，右から $\boldsymbol{T}$ を掛けたときには $\boldsymbol{V}$ の右から $\boldsymbol{T}$ を掛ければよい．はじめに $\boldsymbol{U}$ や $\boldsymbol{V}$ を単位行列に初期化しておくのはもちろんである．

7 数値積分

関数 $f(x)$ が与えられているとして，定積分

$$I = \int_a^b f(x)dx \tag{7.1}$$

を計算することを考える．はじめに最も簡単な公式をあげる．ここでは $h = b - a$ である．

中点公式

$$I_C = hf\Big(\frac{a+b}{2}\Big) \tag{7.2}$$
$$I - I_C = \frac{h^2}{4!}[f'(b) - f'(a)] - \frac{7h^4}{2^3 \cdot 6!}[f^{(3)}(b) - f^{(3)}(a)] + \frac{31h^6}{2^3 \cdot 3 \cdot 8!}[f^{(5)}(b) - f^{(5)}(a)] + \cdots$$

誤差の公式はオイラーの和の公式から導かれたものである．誤差には奇数次の微係数しか現れてこない．右辺は必ずしも収束するとは限らない．

台形公式

$$I_T = \frac{h}{2}[f(a) + f(b)] \tag{7.3}$$
$$I - I_T = -\frac{h^2}{2 \cdot 3!}[f'(b) - f'(a)] + \frac{h^4}{6!}[f^{(3)}(b) - f^{(3)}(a)] - \frac{h^6}{3! \cdot 7!}[f^{(5)}(b) - f^{(5)}(a)] + \cdots$$

誤差の項はオイラー・マクローリン展開から導かれる．面白いことに，誤差項は中点公式よりも係数が大きい．

中点公式，台形公式の幾何学的意味は明らかであろう．図を描いてみれば中点公式の方が誤差が少ないことも理解できる．

シンプソンの公式 上の二つの公式を組み合わせて，誤差の h^2 の項を消去したものである．

$$I_S = \frac{h}{6}\Big[f(a) + 4f\Big(\frac{a+b}{2}\Big) + f(b)\Big] = \frac{1}{3}(2I_C + I_T) \tag{7.4}$$
$$I - I_S = -\frac{h^4}{2^2 \cdot 6!}[f^{(3)}(b) - f^{(3)}(a)] + \frac{5h^6}{2 \cdot 3! \cdot 8!}[f^{(5)}(b) - f^{(5)}(a)] + \cdots$$

この公式は三点を通る二次式を積分しても求められる．

ニュートン・コーツ型公式 $f(x)$ を x の多項式で近似して，その多項式を積分することによってさまざまな公式が得られる．以下では座標点が等間隔に並んでいるものとし

$$f_j = f(x_j) \qquad h = x_{j+1} - x_j$$

とする．また以下の公式の誤差項に現れる ξ は積分区間内のある座標を表す．

$$\int_{x_0}^{x_1} f(x)dx = \frac{h}{2}(f_0 + f_1) - \frac{h^3}{12}f''(\xi) \quad \text{台形公式} \tag{7.5}$$
$$\int_{x_0}^{x_2} f(x)dx = \frac{h}{3}(f_0 + 4f_1 + f_2) - \frac{h^5}{90}f^{(4)}(\xi) \quad \text{シンプソンの公式} \tag{7.6}$$
$$\int_{x_0}^{x_3} f(x)dx = \frac{3h}{8}(f_0 + 3f_1 + 3f_2 + f_3) - \frac{3h^5}{80}f^{(4)}(\xi) \quad \text{シンプソンの } \tfrac{3}{8} \text{ 公式} \tag{7.7}$$
$$\int_{x_0}^{x_4} f(x)dx = \frac{2h}{45}(7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4) - \frac{8h^7}{945}f^{(6)}(\xi) \tag{7.8}$$
$$\int_{x_0}^{x_8} f(x)dx = \frac{4h}{14175}(989f_0 + 5888f_1 - 928f_2 + 10496f_3 - 4540f_4 + 10496f_5 - 928f_6 + 5888f_7 + 989f_8) - \frac{2368h^{11}}{467775}f^{(10)}(\xi) \tag{7.9}$$

今度は誤差項に偶数次の微分しか現れてこないが，これはたとえば台形公式の場合，誤差の第一項 $f'(b) - f'(a)$ を平均値の定理を用いて $hf''(\xi)$ で近似しているからである．またシンプソンの公式の誤差項の係数が (7.4) 式と異なるのは，(7.4) 式では $b - a = h$ としているのに対して (7.6) 式では $x_2 - x_0 = 2h$ としているからである．なお最後の公式には負の係数が現れていて桁落ちの心配がないわけではないが，積分区間の数が 2 の冪乗の公式は便利であるのであげておいた．

上の公式群は次数があがるにつれて誤差項の h の次数が高くなっていくので，高次の公式ほど誤差が極端に小さくなるような錯覚をするが，必ずしもそうではない．高次になるほど誤差項の微分の階数も上がるからである．たとえば $f(x)$ が単振動

$$f(x) = e^{i\omega x}$$

のとき，4 区間の公式 (7.8) 式の誤差項は

$$-\frac{h(i\omega h)^6}{118.1}e^{i\omega\xi}$$

である．一方，8 区間の公式 (7.9) 式の誤差項は

$$-\frac{h(i\omega h)^{10}}{197.5}e^{i\omega\xi}$$

である．したがって $\omega h > 1$ ならば (7.9) 式の誤差の方が大きくなる可能性もある．もちろん，周期関数 $f(x)$ を表現するためには一周期あたり最低でも 10 点は必要であろう．このときには h は

$$\omega h < 2\pi/10 < 1$$

を満たすように選ばなければならないから，(7.9) 式の誤差の方が (7.8) 式の誤差より小さくなる．しかし単振動のときには h を選ぶ基準があるが，複雑な関数になると高階微分がどのような振舞をするかがわからないので，高次の公式でかえって誤差が大きくなるおそれもある．

ニュートン・コーツ型の積分公式は，隣の区間を同じ方法で積分して，いわゆる複合公式として用いることができる．たとえば台形公式を x_{2n} まで適用すれば

$$\begin{aligned}\int_{x_0}^{x_{2n}} f(x)dx = h\big[&\frac{1}{2}f_0 + f_1 + f_2 + \cdots \\ &+ f_{2n-1} + \frac{1}{2}f_{2n}\big] \\ &-\frac{h^2}{12}\big[f'(x_{2n}) - f'(x_0)\big] + \cdots\end{aligned} \tag{7.10}$$

という近似公式が得られる．シンプソンの公式からは

$$\begin{aligned}\int_{x_0}^{x_{2n}} f(x)dx = \frac{h}{3}\big[&f_0 + 4f_1 + 2f_2 + \cdots \\ &+ 2f_{2n-2} + 4f_{2n-1} + f_{2n}\big] \\ &-\frac{h^4}{180}\big[f^{(3)}(x_{2n}) - f^{(3)}(x_0)\big] + \cdots\end{aligned} \tag{7.11}$$

が得られる．

ニュートン・コーツ型公式には開いた公式というのもある．これは x_0 から x_n まで積分するときに端点 x_0 と x_n の値を用いないもので，たとえば

$$\int_{x_0}^{x_3} f(x)dx = \frac{3h}{2}(f_1 + f_2) + \frac{h^3}{4}f''(\xi)$$

である．このタイプの公式は精度も悪いし，複合公式として用いる場合に隣の区間との間に隙間が生じるので利用しない方がよい．

周期関数の積分　$f(x)$ が周期関数のときに一周期にわたって積分すると，(7.3)，(7.4) 式などに現れる $f^{(2k-1)}(x)$ が積分の上下限で等しくなり，誤差項が見かけ上 0 になる．極端な例として

$$\int_0^{\pi} \cos^2\theta d\theta = \frac{\pi}{2}$$

を積分するときに $h = \pi/2$ として点 $\theta = 0, \pi/2, \pi$ の値をを用いて台形公式で計算すると

$$I_T = \frac{\pi}{2}\Big(\frac{1}{2} + 0 + \frac{1}{2}\Big) = \frac{\pi}{2}$$

が得られる．これは正確な値である．h を細かくしても結果は同じである．すなわち，周期関数を積分するときには，含まれている周期よりも短い h を用いて台形公式で積分すれば正しい値が得られることになる．

周期関数ではないが，被積分関数の性質によって精度が異常に高くなる場合がある．よくあげられる例は

$$\int_{-1}^{1} \frac{1}{1+x^2}dx$$

である．この被積分関数の三階微分は

$$f'''(x) = \frac{24x(1-x^2)}{(1+x^2)^4}$$

であるから，シンプソンの公式を用いたときには (7.4) 式によって誤差は h^4 ではなく h^6 のオーダーになる．

ロンバーグの補外法　誤差が h の何乗に比例するかがわかっていると，この性質を用いて組織的に精度の高い近似値を求めていくことができる．

積分区間 $[a, b]$ を n 等分して

$$h = \frac{b-a}{n} \qquad x_j = a + jh \quad j = 0, 1, 2, \cdots, n$$

とする．各区間に台形公式を用いて計算した近似値を

$$I_T(h) = h\big[\frac{1}{2}f_0 + f_1 + \cdots \\ \cdots + f_{n-1} + \frac{1}{2}f_n\big] \tag{7.12}$$

とすると，この近似値の誤差は

$$I - I_T(h) = \alpha h^2 + \beta h^4 + \cdots$$

である．α, β は h によらない定数である．そこで分割を二倍にして同じ公式を用いれば

$$I - I_T(h/2) = \frac{1}{4}\alpha h^2 + \frac{1}{16}\beta h^4 + \cdots$$

となる．両式から h^2 の項を消去すれば

$$I = \frac{1}{3}\left[4I_T(h/2) - I_T(h)\right] - \frac{1}{4}\beta h^4 + \cdots$$

が得られる．(7.9) 式と (7.10) 式を用いれば，右辺第一項はシンプソンの公式にほかならないことがわかる．さらに分割を半分にして h^4 の項を消去するとさらに精度の高い近似値が得られる．これをシステマティックに行うと次のような公式になる．

$$S_0^{(k)} = I_T(h/2^k) \quad k = 0,\, 1,\, 2, \cdots \\ S_m^{(k)} = S_{m-1}^{(k+1)} + \frac{S_{m-1}^{(k+1)} - S_{m-1}^{(k)}}{4^m - 1} \\ m = 1,\, 2, \cdots \tag{7.13}$$

$S_m^{(k)}$ は m と k を交互にふやしていく．表で書けば

$$\begin{array}{lll} S_0^{(0)} & S_1^{(0)} & S_2^{(0)} \\ S_0^{(1)} & S_1^{(1)} & \\ S_0^{(2)} & & \end{array}$$

となる．$S_0^{(0)}$ の次に h を半分にした $S_0^{(1)}$ を計算しこれらから $S_1^{(0)}$ を計算する．次に h をさらに半分にした $S_0^{(2)}$ を計算し，$S_0^{(1)}$ と $S_0^{(2)}$ から $S_1^{(1)}$ を，これと前に計算しておいた $S_1^{(0)}$ から $S_2^{(0)}$ を計算する．この表においては右上の値が最も精度が高い．

$S_m^{(k)}$ は二次元の表になっているが，実は一次元の配列で十分である．$S_0^{(1)}$ が求められたとき，これと $S_0^{(0)}$ とから $S_1^{(0)}$ を計算した後には $S_0^{(0)}$ は不要になるので，ここに $S_1^{(0)}$ を上書きすればよい．同様に $S_1^{(1)}$ は $S_0^{(1)}$ に上書きすればよい．上の表の場合には上から $S_2^{(0)}$，$S_1^{(1)}$，$S_0^{(2)}$ だけが残っている．

当然のことながら，h を半分にしたときには新たに増えた分点における $f(x)$ だけを計算すればよい．

h を半分にする手続きはいくらでも続けることができるが，m をあまり大きくまで計算するのは考え物である．ラグランジェの補間公式のところでも述べたように，次数の高い補間公式は暴れる可能性がある．補外法は h，$h/2$，$h/4$ などの値から補間公式を作り，$h = 0$ のときの値を推定するものである．したがって $m = 5$ 程度までで止めておいた方が安全である．

実際は，被積分関数が解析的なら収束は非常に速い．積分

$$\int_0^2 x^4 \log(x + \sqrt{1 + x^2})dx$$

をはじめ全区間を四等分 ($h = 0.5$) した場合の結果は次のようになった．

k	$S_0^{(k)}$	$S_1^{(k)}$	$S_2^{(k)}$	$S_3^{(k)}$
0	9.254511	8.155954	8.153362	8.153365
1	8.430593	8.153524	8.153365	
2	8.222792	8.153375		
3	8.170728			

右上の $S_3^{(0)}$ が最終結果である．最終結果が得られるまでに 32 等分で十分である．

指数変換を用いた積分公式 (7.3) 式の誤差項が 0 になるのは周期関数のときだけではない．積分

$$\int_{-\infty}^{\infty} e^{-x^2}dx$$

の被積分関数 $f(x) = e^{-x^2}$ の微係数は積分の上下限 $x = \pm\infty$ で急速に 0 に収束する．したがってこれを台形公式で計算すれば高精度の値が得られることが期待される．実際，極端な例として上式を $h = 1$ として $x = 0$, ±1, ±2, ±3, ±4 の値を用いて計算すると 1.772637 が得られる．これは真値 $\sqrt{\pi} = 1.772453$ に非常に近い．もし $h = 0.5$ を用いれば有効数字 10 桁以上の値が得られるであろう．

これを参考にして次のような方法を考える．求める積分範囲を変換して

$$I = \int_{-1}^{1} f(x)dx$$

とする．積分変数 x を

$$x = \tanh y \quad dx = \frac{dy}{\cosh^2 y} \tag{7.14}$$

によって y に変換すれば，求める積分は

$$I = \int_{-\infty}^{\infty} f(\tanh y)\frac{dy}{\cosh^2 y}$$

になる．y の被積分関数 $f(\tanh y)/\cosh^2 y$ は $y = \pm\infty$ で $e^{-2|y|}$ で 0 に収束する．収束の速さは先の例よりも遅いが，台形公式を用いたときに誤差が小さくなることは十分期待される．そこで y の刻み幅を h とすれば

$$I_{SE} = h\sum_{k=-\infty}^{\infty} w_k f(x_k) \tag{7.15}$$

$$y_k = kh \quad x_k = \tanh y_k \quad w_k = \frac{1}{\cosh^2 y_k}$$

という積分公式が得られる．和の上下限は $\pm\infty$ となっているが，重み w_k は急速に 0 近づくので有効桁数に応じたところで打ち切ればよい．

分点の間隔 h は，はじめは 1/2 あるいは 1/4 にとり，次々に半分にしていく．h を変えるごとに分点の座標 x_k や重み w_k を計算するのは面倒であるから，はじめに $h = 1/2^M$ で座標や重みを計算した表を作っておき，表をとびとびに引くことにより $h = 1/2, 1/4, 1/8, \cdots$ のときの x_k や w_k を求めるようにプログラムしておくとよい．(7.15) 式で h を w_k に含めないで，和の前に出るように書いてあるのはそのような計算に便利であるからである．

刻み幅 h のときの $I_{SE}(h)$ の誤差はおよそ

$$I - I_{SE}(h) \sim \exp\Big(-\frac{c}{h}\Big) \tag{7.16}$$

で表されることが知られている．c は定数である．したがって刻み幅を半分にしたときの誤差は

$$I - I_{SE}(h/2) \sim (I - I_{SE}(h))^2$$

になる．$I_{SE}(h/2)$ が十分真値 I に近いとすれば右辺の I を $I_{SE}(h/2)$ で置き換えて

$$\sqrt{|I - I_{SE}(h/2)|} \sim |I_{SE}(h/2) - I_{SE}(h)|$$

が成り立つ．このことを利用すれば刻み幅 h を順々に半分にしていき，上式の右辺が所望の誤差以下になったら反復を止めるというやり方ができる．

なお，$f(x) = 1$ のときには積分値は 2 になるはずであるから

$$h\sum_k w_k = 2$$

が成り立たなければならない．この関係は数表のチェックに用いることができる．

二重指数変換を用いた積分公式　さらに効率のよいのが変数変換 (7.14) 式のかわりに

$$x = \tanh\Big(\frac{\pi}{2}\sinh y\Big) \tag{7.17}$$

$$dx = \frac{\frac{\pi}{2}\cosh y}{\cosh^2\Big(\frac{\pi}{2}\sinh y\Big)}dy$$

を用いるものである (森，1975)．この変換では y の被積分関数は $y \to \pm\infty$ では

$$\exp(-c\exp|y|)$$

のオーダーで減衰する．そこでこの変換を二重指数変換 (DE) と呼ぶ．この変換を用いた積分公式は直ちに書き下すことができて

$$I_{DE}(h) = h\sum_{k=-\infty}^{\infty} w_k f(x_k) \tag{7.18}$$

$$y_k = kh \quad x_k = \tanh\Big(\frac{\pi}{2}\sinh y_k\Big)$$

$$w_k = \frac{\frac{\pi}{2}\cosh y_k}{\cosh^2\Big(\frac{\pi}{2}\sinh y_k\Big)}$$

となる．二重指数変換の誤差も (7.16) 式を満たしている．収束の判定も前項と同じにすることができる．

ラプラス変換の計算　$f(t)$ のラプラス変換は

$$F(p) = \int_0^{\infty} f(t)e^{-pt}dt \quad p > 0$$

で定義される．ラプラス変数 p が大きいときには e^{-pt} の減衰が激しいので t の小さいところを細かく計算しないと正確な積分が得られないし，反対に p が小さいときには減衰が緩やかなので大きな t まで考慮する必要がある．

そこで

$$pt = e^y \quad dt = \frac{1}{p}e^y dy \tag{7.19}$$

と変数変換すると，積分は

$$F(p) = \frac{1}{p}\int_{-\infty}^{\infty} f(t)w(y)dy \tag{7.20}$$

となる．ここに

$$w(y) = \exp(y - e^y) \tag{7.21}$$

である．したがって積分公式は

$$F(p) = \frac{h}{p} \sum_{k=-\infty}^{\infty} f(t_k) w_k \tag{7.22}$$
$$y_k = kh \quad pt_k = e^{y_k} \quad w_k = w(y_k)$$

となる．

このやり方は結果的には二重指数変換 (7.16) 式と同じような形になっている．ただし (7.16) 式が $y = 0$ に関して対称であるのに対して，(7.19) 式は対称性が非常に悪い．たとえば重み関数 $w(y)$ が 10^{-10} になるのは

$$y = -20.7 \quad pt = 1.0 \times 10^{-9}$$
$$y = 2.0 \quad pt = 20.7$$

の二点である．これは指数関数 e^{-pt} の減衰によって $pt > 20.7$ からの寄与は無視できるのに対して，原点 $t = 0$ 付近は細かく計算しなければならないことを意味している．

チェビシェフの積分公式　これまでは関数値を等間隔に計算していたが，積分だけを求めるのであれば必ずしも等間隔である必要はない．以下では積分範囲を $[-1, 1]$ にスケールして

$$I = \int_{-1}^{1} f(x) dx$$

を考える．この積分の近似値として

$$I_{Ch} = \frac{2}{n} \sum_{k=1}^{n} f(x_k) \tag{7.23}$$

の形を仮定する．$f(x)$ が n 次までの x の多項式に対して I_{Ch} の誤差が 0 になるという条件から分点の座標 x_k を決めることができる．たとえば $n = 2$ の場合には条件式は

$$x_1 + x_2 = 0 \quad x_1^2 + x_2^2 = \frac{2}{3}$$

であるから，座標は

$$x_2 = -x_1 = \frac{1}{\sqrt{3}}$$

となる．

$n = 8$，$n > 9$ に対しては分点が複素数になるので積分公式は得られない．求められている分点を下に示す．

n	$\pm x_k$	n	$\pm x_k$
2	0.57735 02692	3	0
			0.7071067812
4	0.18759 24741		5 0
	0.79465 44723		0.37454 14096
			0.83249 74870
6	0.26663 54015	7	0
	0.42251 86538		0.32391 18105
	0.86624 68181		0.52965 67753
			0.88386 17008
9	0		
	0.16790 61842		
	0.52876 17831		
	0.60101 86554		
	0.91158 93077		

チェビシェフの積分公式の分点

ガウスの積分公式　チェビシェフの積分公式は重みがすべて 1 であるという意味で簡単であるが，重みを変えることによってさらに精度の高い積分公式が得られる．積分の近似値として

$$I_G = \sum_{k=1}^{n} w_k f(x_k) \tag{7.24}$$

とすれば $2n$ 個のパラメーター x_k, w_k を自由に選ぶことができる．これらのパラメーターを $f(x)$ が $2n - 1$ 次多項式のときまで正確になるように選ぶようにすれば，これらを一義的に決めることができる．たとえば $n = 3$ のときにはこれらのパラメーターを決める式は

$$\begin{aligned}
w_1 + w_2 + w_3 &= 2 \\
w_1 x_1 + w_2 x_2 + w_3 x_3 &= 0 \\
w_1 x_1^2 + w_2 x_2^2 + w_3 x_3^2 &= \frac{2}{3} \\
w_1 x_1^3 + w_2 x_2^3 + w_3 x_3^3 &= 0 \\
w_1 x_1^4 + w_2 x_2^4 + w_3 x_3^4 &= \frac{2}{5} \\
w_1 x_1^5 + w_2 x_2^5 + w_3 x_3^5 &= 0
\end{aligned}$$

となる．非線型の非常に複雑な式のように見えるが，対称性を用いれば簡単に解けて

$$x_2 = 0 \quad x_1 = -x_1 \quad x_1^2 = \frac{3}{5}$$
$$w_1 = w_3 = \frac{5}{9} \qquad w_2 = \frac{8}{9}$$

となる．一般の次数のときにはこのような方法で分点の座標や重みを求めることは非常に面倒になる．しかしルジャンドルの多項式を用いた内挿公式を積分することによって求めることができる．

直交多項式を用いた積分公式　重み関数を $w(x)$ とする区間 $[a,b]$ における直交多項式 $p_j(x)$ は既に §5 で導いてある．$p_n(x)$ の零点を $x_k\,(k=1,2,\cdots,n)$ とするとこれらの点を通る $n-1$ 次の内挿公式は

$$f_{n-1}(x) = \sum_{j=0}^{n-1} \frac{1}{\lambda_j} \left[\sum_{k=1}^{n} w_k p_j(x_k) f(x_k) \right] p_j(x)$$

であった．$f(x)$ の重み付き積分の近似値として $f_{n-1}(x)$ の積分をとれば

$$I_G = \int_a^b w(x) f_{n-1}(x) dx$$
$$= \sum_{j=0}^{n-1} \frac{1}{\lambda_j} \left[\sum_{k=1}^{n} w_k p_j(x_k) f(x_k) \right] \int_a^b w(x) p_j(x) dx$$

である．最後の積分は

$$\int_a^b w(x) p_j(x) dx$$
$$= \frac{1}{\mu_0} \int_a^b w(x) p_0(x) p_j(x) dx = \frac{\lambda_0}{\mu_0} \delta_{0j}$$

であるから，j についての和は $j=0$ だけでよい．したがって積分の近似公式

$$\int_a^b w(x) f(x) dx \doteq I_G = \sum_{k=1}^{n} w_k f(x_k) \quad (7.25)$$

が得られた．

$w(x)=1$ のときには上式の左辺は通常の積分であり，$p_n(x)$ はルジャンドルの多項式 $P_n(x)$ になる．このときの分点と重みは

$$P_n(x_k) = 0 \qquad k = 1,2,\cdots,n$$
$$w_k = \frac{2(1-x_k^2)}{[nP_{n-1}(x_k)]^2} \quad (7.26)$$

と表される．この公式は $n-1$ 次の近似式を積分して得られたものであるが，実際には $2n-1$ 次多項式まで正確な値を与える．下に低次の分点と重みを示す．

n	$\pm x_k$	w_k
2	0.57735 02692	1
3	0	0.88888 88889
	0.77459 66692	0.55555 55556
4	0.33998 10435	0.65214 51549
	0.86113 63116	0.34785 48451
5	0	0.56888 88889
	0.53846 93101	0.47862 86705
	0.90617 98459	0.23692 68851

ガウスの積分公式の分点と重み

ルジャンドル多項式以外の直交多項式を用いると，重みの付いた積分に対する積分公式を導くことができる．たとえばラゲールの多項式を用いれば

$$\int_0^\infty f(x) e^{-x} dx \doteq \sum_k w_k f(x_k)$$

という形の積分公式が導かれる．しかし上で述べたガウスやチェビシェフの公式は複合公式として，つまり小区間ごとに適用できるのに対して，重みが 1 でない直交多項式を用いた場合には全区間に一つの公式を用いなければならないので，精度を高めるには次数を上げるほかない．

特異点の処理　話をわかりやすくするために，積分

$$I = \int_0 \frac{g(x)}{\sqrt{x}} dx$$

を考える．$g(x)$ は $x=0$ で有限な関数であるとする．積分の上限はいまのところ問題ではない．この積分は可積分である．なぜなら

$$x = y^2$$

と変数変換すれば

$$I = \int_0 2g(y^2) dy$$

となって積分できるからである．したがってこのような場合には $x=0$ 付近は特別に積分しなければならない．

似たような例で

$$I = \int_0 g(x)\sqrt{x}dx$$

がある．これはもちろん可積分であるが，$x = 0$ 付近で $\sqrt{x}$ が急激に増えるので，全区間を同じ間隔の分点を用いて積分しようとすると $x = 0$ 付近で足を引っ張られて分点の数が多くなってしまう．この場合にも $x = 0$ 付近だけは特別に計算しなければならない．この場合にも同じ変数変換を行えば

$$I = \int_0 2g(y^2)y^2dy$$

となる．一般に

$$I = \int_0 g(x)x^{\pm\alpha}dx \quad 0 < \alpha < 1$$

のときの対処法は明らかであろう．

いまは $x = 0$ に特異点がある場合を例としたが，積分区間内に特異点がある場合には特異点が積分の上下限になるように区間を分ける必要がある．したがって二重指数変換 (7.16) 式を用いる場合にも特異点があるとすれば $y = \pm\infty$ にあるとしてよい．この場合には特別な変数変換をする必要はない．なぜならこの公式では x が端点になることはないからである．ただし丸め誤差のために $x = \pm 1$ になってしまうことがあるかもしれない．これを避けるためには次のようにすればよい．

いま $x = 1$ に特異点があったとする．このとき (7.16) 式により

$$1 - x = \frac{\exp\left(-\frac{\pi}{2}\sinh y\right)}{\cosh\left(\frac{\pi}{2}\sinh y\right)}$$

が成り立つから，(7.17) 式の $f(x)$ を計算するときに引数として x ではなく $1 - x$ を与えるようにしておけば $x = 1$ 付近の $f(x)$ が正確に計算できることになる．$x = -1$ に特異点があるときも同様である．

適応的積分法　積分を行うとき積分区間全体にわたって一つの公式，たとえばガウスの公式を適用するわけではなく，いくつかの部分区間に分けて適用する．しかし積分区間を等間隔に分けるのは効率が悪い．被積分関数の変化が激しいところでは狭い区間を用い，変化が緩やかなところでは区間の幅を広くとるのが効率的であることはいうまでもない．

ある区間の積分をするとしよう．刻み幅 h で積分した値 $I(h)$ が正しいかどうかをチェックするために，刻み幅を半分の $h/2$ にしたときの値 $I(h/2)$ との差から誤差を評価して先に進むかどうかを判定する．刻み幅を半分にしたときの積分公式をどのように選ぶかが一つの問題である．台形公式を用いれば積分区間全体の刻み幅を半分にすることは容易である．これはロンバーグ法にほかならない．またニュートン・コーツ型の公式を順に (7.5)，(7.6)，(7.8)，(7.9) 式と用いていくこともできる．しかしどちらの方法でも区間全体の刻み幅を小さくしていくのは効率が悪い．必要なところは細かく刻み，必要のないところでは粗く刻む方が効率的である．

まず全区間における積分の近似値を刻み幅 h で計算しておく．この値を $I(h)$ とする．たとえば 4 区間の近似公式 (7.8) 式で計算したとすれば，この誤差は

$$I - I(h) = -\frac{8h^7}{945}f^{(6)}(\xi)$$

で表される．次に区間を半分にして左側と右側に刻み幅 $h/2$ の同じ近似公式を用いて計算した全区間の近似値を $I(h/2)$ とする．この近似値の誤差は

$$I - I(h/2) = -\frac{8h^7}{2^7 \cdot 945}[f^{(6)}(\xi_L) + f^{(6)}(\xi_R)]$$

と表される．ξ_L，ξ_R は半分に分割した左側と右側の区間内のある点である．全区間が狭い場合にはこれらは $f^{(6)}(\xi)$ と等しいと考えてよいだろう．そこで上式から $f^{(6)}(\xi)$ を消去すれば，誤差評価式

$$I - I(h/2) = \frac{1}{63}[I(h/2) - I(h)] \tag{7.27}$$

が導かれる．右辺があらかじめ定められた誤差の上限 ε よりも小さければ，すなわち

$$\frac{1}{63}|I(h/2) - I(h)| < \varepsilon \tag{7.28}$$

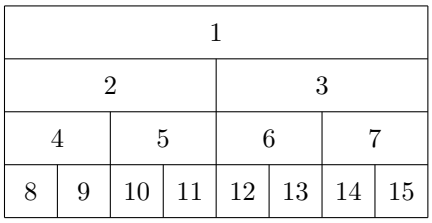
が成り立てば，$I(h/2)$ をその区間の積分の近似値として採用する．このとき $I(h/2)$ を全区間の積分の近似値として採用するよりも，(7.23) 式より

$$I \doteq I(h/2) + \frac{1}{63}[I(h/2) - I(h)] \tag{7.29}$$

を採用した方が精度が高い．

もしこの条件が満たされなかったとき，左半分の区間をさらに半分にして $I(h/2)$ と $I(h/4)$ の比較を行う．ただし区間が半分になっているから，誤差の

上限としてはεではなく$\varepsilon/2$を用いなければならない．この条件が満たされたなら右半分についても同じような計算を行う．満たされなかったら左半分をさらに二分する．



小区間の二分木表現

このように分割を繰り返していくと，全区間はいろいろな幅の小区間に分かれる．これを統一的に処理するために二分木の考え方を使う．図の一番上のレベル0の区間1が全区間である．2番目のレベル1の区間2と3の値を使って全区間1の積分を評価する．これが基準を満たしていなければ小区間4と5を用いて誤差を評価するという手順を繰り返す．たとえば区間8で基準が満たされたとすると，次には区間9の処理を行わなければならない．区間9で基準が満たされたときには次には区間10の処理ではなく，一つ上のレベルの区間5の処理を行わなければならない．区間5が基準を満たさなければ一つ下のレベルの区間10の処理を行う．このように二分木を行ったりきたりしながら全区間の積分が求められることになる．

参考文献

森 正武，1975：数値解析と複素関数論，筑摩書房．

8 差分方程式と特殊関数の計算

ベッセル関数などの特殊関数と差分方程式の間になんの関係があるかと思うかもしれないが，数値計算上は両者には密接な関係がある．ここで取り上げる特殊関数は主に円筒関数であるが，ここで述べる議論は漸化式を満たすほかの関数についてもあてはまる．

差分方程式 単振子の運動を記述する二階の斉次常微分方程式

$$\frac{d^2y(t)}{dt^2} + \omega^2 y(t) = 0 \tag{8.1}$$

を $t = n\Delta t$ と離散化して

$$y_n = y(n\Delta t)$$

と書き，二階微分を中心差分

$$\frac{d^2y(t)}{dt^2} \doteq \frac{y_{n+1} - 2y_n + y_{n-1}}{\Delta t^2}$$

で近似すると，もとの微分方程式は斉次差分方程式

$$y_{n+1} - [2 - (\omega\Delta t)^2]y_n + y_{n-1} = 0 \tag{8.2}$$

で近似される．

これを少し一般化して

$$y_{n+1} + \alpha y_n + \beta y_{n-1} = 0 \tag{8.3}$$

という形の差分方程式を考える．α，β は n によらない定数である．y_0 と y_{-1} が与えられたとすれば，上式を $n = 0, 1, \cdots$ の順に計算すれば任意の y_n が求められるはずである．

α，β が n によらないときには，微分方程式からの類推で y_n が

$$y_n \sim z^{-n}$$

の形に書けると仮定する．z^n ではなく z^{-n} としたのは後の式との整合性のためである．これを (8.3) 式に代入すると

$$z^{-2} + \alpha z^{-1} + \beta = 0$$

が成り立たなければならない．したがって上式の根を

$$z_1^{-1},\, z_2^{-1} = \frac{1}{2}\left[-\alpha \pm \sqrt{\alpha^2 - 4\beta}\right]$$

とすれば，(8.3) 式の一般解は

$$y_n = Az_1^{-n} + Bz_2^{-n} \tag{8.4}$$

と書くことができる．

A，B は微分方程式における積分定数に相当するものであるが，差分方程式の場合には微分方程式の場合とは違って，$n = 0$ だけの初期条件から決めることはできない．差分方程式の場合，y_0 と y_{-1} が与えられているとすれば

$$y_0 = A + B \quad y_{-1} = Az_1 + Bz_2 \tag{8.5}$$

から A と B を決めることができる．

$|z_1| < |z_2|$ とする．B が 0 でなくても，n が増えていけば必ず z_1 成分が卓越するようになってしまう．これは常微分方程式の場合でも同様である．仮に $A = 0$ であったとしても，数値計算の場合には丸め誤差のために必ず z_1 成分が現れるから，やはり z_1 成分が卓越してしまう．逆に非常に大きな N で y_N と y_{N+1} が与えられたとして A と B を決めて，差分方程式 (8.3) を逆順で，すなわち $n = N, N-1, \cdots$, に対して

$$y_{n-1} = -\frac{1}{\beta}(\alpha y_n + y_{n+1})$$

を計算する場合には，先とは逆に n が減るについれはいずれは z_2 成分が卓越してしまう．

以上をまとめると，差分方程式 (8.3) 式の特性根が $|z_1| < |z_2|$ を満たしているとき，(8.3) 式を順方向に計算すれば最終的には z_1 成分が生き残り，反対に逆方向に計算すれば z_2 成分が生き残る．$|z_1| = |z_2|$ のときにはどちら向きに計算しても両成分ともに生き残る．

絶対値の小さい方の特性根が $|z_1| < 1$ であるとすると，(8.3) 式を順方向に計算すると $n \to \infty$ で y_n は発散してしまう．したがって (8.3) 式が $n \to +\infty$ で安定であるためには

$$|z_1| > 1 \quad |z_2| > 1 \tag{8.6}$$

が成り立たなければならない．いいかえれば，(8.3) 式が順方向に安定であるためには，特性根がすべて複素 z 平面の単位円の外になければならない．これに関する詳しい議論は z 変換の項で行う．

なお，$\omega^2 > 0$ のとき差分方程式 (8.2) の特性根は二つとも複素数で，絶対値は 1 である．したがっ

て (8.2) 式はどちら向きに計算しても安定である．$|\omega\Delta t| \ll 1$ のときには特性根は

$$z_1,\, z_2 \doteq 1 \pm i\omega\Delta t \doteq e^{\pm i\omega\Delta t}$$

であるから，(8.2) 式の解はもとの微分方程式 (8.1) の解

$$y(t) = Ae^{-i\omega t} + Be^{i\omega t}$$

を近似している．

三角関数の漸化式　(8.3) 式で

$$\alpha = -2\cos\theta \qquad \beta = 1$$

のときには，特性根は

$$z_1^{-1},\, z_2^{-1} = \cos\theta \pm i\sin\theta = e^{\pm i\theta}$$

となる．両方とも絶対値は 1 であるから，(8.3) 式をどちらの方向に進めても計算は安定に進められる．

初期条件として

$$y_0 = 0$$

が与えられたとする．このとき

$$A + B = 0$$

であるから

$$y_{-1} = A(e^{-i\theta} - e^{i\theta}) = -2iA\sin\theta$$

となる．方程式は線型であるから

$$A = \frac{1}{2i}$$

と選ぶことにすれば，(8.3) 式の解は

$$y_n = \sin n\theta$$

となる．同様に

$$y_0 = 1$$

の解を求めれば

$$y_n = \cos n\theta$$

が得られる．これら二つの解はともに (8.3) 式を満たしている．これらは三角関数の漸化式

$$\begin{aligned}\sin(n+1)\theta + \sin(n-1)\theta &= 2\cos\theta\sin n\theta \\ \cos(n+1)\theta + \cos(n-1)\theta &= 2\cos\theta\cos n\theta \end{aligned} \quad (8.7)$$

に対応している．この漸化式を用いれば，一度 $\sin\theta$, $\cos\theta$ を計算しておけば $\cos n\theta$，$\sin n\theta$ がそれぞれ一回の積和で求められるので，フーリエ級数を計算するときに便利である．

円筒関数の計算　ベッセルの微分方程式

$$\frac{d^2Z_n(x)}{dx^2} + \frac{1}{x}\frac{dZ_n(x)}{dx} + \left(1 - \frac{n^2}{x^2}\right)Z_n(x) = 0 \quad (8.8)$$

を満足する解 $Z_n(x)$ を円筒関数という．ここでは n は 0 または正の整数，また $x \geq 0$ とする．円筒関数にはいろいろあるが，よく知られているのはベッセル関数である．これは冪級数

$$J_n(x) = \left(\frac{x}{2}\right)^n \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(k+n)!}\left(\frac{x}{2}\right)^{2k} \quad (8.9)$$

で表すこともできる．この式は任意の x に対して成り立つから，$J_n(x)$ の値を求めるのに用いることができる．しかし x が大きくなると収束が遅くなり，また交替級数であるために桁落ちが激しく，正確な値を得るためには多倍長計算が必要になる．一例として $J_0(x)$ を $x = 10$ に対して単精度で計算した結果を示す．

k	a_k	S_k
0	1.000000e+0	1.000000e+0
1	−2.500000e+1	−2.400000e+1
2	1.562500e+2	1.322500e+2
3	−4.340278e+2	−3.017778e+2
4	6.781684e+2	3.763906e+2
5	−6.781684e+2	−3.017778e+2
6	4.709503e+2	1.691725e+2
7	−2.402808e+2	−7.110825e+1
8	9.385967e+1	2.275143e+1
9	−2.896903e+1	−6.217607e+0
10	7.242259e+0	1.024651e+0

a_k は (8.9) 式の各項を，S_k は k 項までの部分和を表している．部分和は $k = 4$ で最大値 376.4 をとり，その後絶対値は減少していく．$J_0(10)$ の正確な値は -0.2459 であるから，このことは最終的には 3 桁の桁落ちが生じていることを意味している．$J_0(x)$ の場合，$x = 5$ くらいまでは冪級数展開で計算するこ

とができる．$J_n(x)$ には漸近展開もあるが，これも収束性に問題がある．

円筒関数 $Z_n(x)$ は漸化式

$$Z_{n+1}(x) - \frac{2n}{x} Z_n(x) + Z_{n-1}(x) = 0 \quad (8.10)$$

を満たしている．したがって $Z_0(x)$ と $Z_1(x)$ がわかれば上式から $Z_2(x), Z_2(x), \cdots$ が計算できることになる．しかしこの方法が必ずしもうまくいくとは限らない．それは $Z_n(x)$ としてベッセル関数 $J_n(x)$ を用いて計算してみればわかる．たとえば $J_0(1)$ と $J_1(1)$ を与えて上式から計算してみると

n	$J_n(1)$
0	0.765197686
1	0.440050585
2	0.114903484
3	0.019563351
4	0.002476622
5	0.000249625

となる．$n = 5$ で既に 3 桁の桁落ちが生じている．

(8.10) 式は (8.3) 式の係数 α に相当するものが n の関数になっているので先の議論はそのまま成り立たつわけではないが，形式的に特性根を求めれば

$$z_1^{-1}, z_2^{-1} = \frac{n}{x} \pm \sqrt{\frac{n^2}{x^2} - 1}$$

であるから，n が小さい間は特性根の絶対値は 1 で，したがって (8.10) 式を n の増える方向に計算しても問題は生じない．しかし計算を続けていけばいつかは n が x を越え，特性根は実数になる．根と係数の関係から一つの根の絶対値は必ず 1 よりも小さい．したがって漸化式 (8.10) を順方向に計算していくと n が大きくなると

$$Z_n(x) \sim z_1^{-n} \quad |z_1| < 1$$

になる．

ところでベッセル関数 $J_n(x)$ は

$$J_0(x) + 2\sum_{n=1}^{\infty} J_{2n}(x) = 1 \quad (8.11)$$

$$J_0^2(x) + 2\sum_{n=1}^{\infty} J_n^2(x) = 1 \quad (8.12)$$

という関係を満たしている．これらの式はいずれも $n \to \infty$ で $J_n(x)$ が 0 に収束することを示している．これは先の順方向の漸化式の解とは矛盾している．前の議論に従えば，このような関数を表すには漸化式を逆方向に計算すればよいことになる．そこで次のような計算法が考えられる．

求めたい $J_n(x)$ に対して $N \gg n$ となる N を選び

$$y_N = \varepsilon \quad y_{N+1} = 0 \quad (8.13)$$

とする．ε は十分に小さな正の数である．これを初期値として漸化式 (8.10) を逆順に計算する．すなわち

$$y_{k-1} = \frac{2k}{x} y_k - y_{k+1} \quad (8.14)$$
$$k = N, N-1, \cdots, 1$$

同時に和

$$S = y_0 + 2\sum_{k=1}^{N} y_{2k}$$

を計算しておく．y_n は $J_n(x)$ に比例しているはずである．そこで (8.11) 式が成り立つように S でスケーリングすれば，$J_k(x)$ の近似値は

$$J_k(x) \doteq \frac{y_k}{S} \quad (8.15)$$

で与えられる．

ε は十分に小さな正数であるが，n や x によっては (8.14) 式の途中でオーバーフローが起こる可能性がある．そこで y_k の絶対値を監視して，大きくなりすぎたら y_k や S をスケーリングする．

それでは (8.10) 式を順方向に計算したときにはなにが得られるのであろうか．実はこれによって得られるのはノイマン関数 $N_n(x)$ である．しかしノイマン関数を漸化式 (8.10) によって計算するためには $N_0(x)$ と $N_1(x)$ を何らかの方法で求めておく必要がある．

$J_0(x)$ や $J_1(x)$ が (8.13, 14) 式から計算できることを用いると，ウロンスキアン

$$J_n(x)N_{n+1}(x) - J_{n+1}(x)N_n(x) = -\frac{2}{\pi x} \quad (8.16)$$

を利用することが考えられる．$J_0(x)$，$J_1(x)$ を上の方法で計算したとすると，なんらかの方法で $N_0(x)$ が計算できれば上式から $N_1(x)$ が求められるから，漸化式 (8.10) を順方向に計算することができる．な

お $x < 5$ 程度なら $N_0(x)$ は

$$\frac{\pi}{2}N_0(x) = \left(\ln\frac{x}{2} + \gamma\right) J_0(x) + \frac{1}{(1!)^2}\left(\frac{x}{2}\right)^2 - (1+1/2)\frac{1}{(2!)^2}\left(\frac{x}{2}\right)^4 + (1+1/2+1/3)\frac{1}{(3!)^2}\left(\frac{x}{2}\right)^6 - \cdots$$

から計算することができる．ここに γ はオイラーの定数で，$\gamma = 0.5772156\cdots$ である．

変形ベッセル関数　変形ベッセル関数は常微分方程式

$$\frac{d^2Z_n(x)}{dx^2} + \frac{1}{x}\frac{dZ_n(x)}{dx} - \left(1 + \frac{n^2}{x^2}\right) Z_n(x) = 0 \tag{8.17}$$

の解として定義される．二つある基本解のうちのひとつ，第一種変形ベッセル関数 $I_n(x)$ は

$$I_n(x) = i^{-n}J_n(ix) \tag{8.18}$$

と表すことができる．したがって $I_n(x)$ の冪級数は (8.9) 式から容易に求めることができる．

$I_n(x)$ の満たす漸化式は

$$I_{n+1}(x) + \frac{2n}{x}I_n(x) - I_{n-1}(x) = 0 \tag{8.19}$$

である．この漸化式の特性根は

$$z_1^{-1},\, z_2^{-1} = -\frac{n}{x} \pm \sqrt{\frac{n^2}{x^2} + 1}$$

であるから常に二根とも実数で，一方は必ず 1 よりも小さい．したがって $I_n(x)$ を求めるには漸化式 (8.19) 式を逆方向に計算しなければならない．スケールファクターは

$$I_0(x) + 2\sum_{n=1}^{\infty} I_n(x) = e^x \tag{8.20}$$

から決めることができる．

微分方程式 (8.17) のもう一つの基本解，第二種変形ベッセル関数 $K_n(x)$ は漸化式

$$K_{n+1}(x) - \frac{2n}{x}K_n(x) - K_{n-1}(x) = 0 \tag{8.21}$$

を満たしている．この漸化式の特性根は $I_n(x)$ のそれの符号を反対にしたものである．したがって上の漸化式を正方向に計算すると n とともに振幅が増加する解が得られる．これは $K_n(x)$ のもつ性質と同じである．したがって $K_0(x)$ と $K_1(x)$ がわかれば上の漸化式を順方向に計算すれば $K_n(x)$ が得られる．ノイマン関数のときと同様に，ウロンスキアン

$$I_n(x)K_{n+1}(x) + I_{n+1}(x)K_n(x) = \frac{1}{x} \tag{8.22}$$

が成り立つので，$K_0(x)$ さえ求められれば漸化式を用いて $K_n(x)$ が求められる．x が小さいときには $K_0(x)$ は $N_0(x)$ と同じような式で計算することができる．

球ベッセル関数　球ベッセル関数 $j_n(x)$ は x の多項式と三角関数の有限項の和で表すことができる．たとえば

$$\begin{aligned} j_0(x) &= x^{-1}\sin x \\ j_1(x) &= x^{-2}(\sin x - x\cos x) \\ j_2 &= x^{-3}[(3-x^2)\sin x - 3x\cos x] \end{aligned} \tag{8.23}$$

などであるが，このような公式を用いて計算すると桁落ちのために正確な値が得られない．

球ベッセル関数 $j_n(x)$ は漸化式

$$j_{n-1}(x) + j_{n+1}(x) = \frac{2n+1}{x}j_n(x) \tag{8.24}$$

を満たしているので，(8.23) 式を初期値としてこの漸化式を n の増える方向に計算することも考えられるが，これはベッセル関数 $J_n(x)$ のときと同様で，ここでも桁落ちが生じてしまう．

そこで $J_n(x)$ のときと同様に，漸化式 (8.24) を n の減る方向に方向に計算する．スケールファクターは今度の場合は簡単で，$j_0(x)$ が (8.23) 式になるように決めればよい．ただし $j_0(x)$ は $x = k\pi$ で 0 になってしまうので，この付近では $j_1(x)$ を用いてスケールファクターを決めればよい．

球ノイマン関数 $n_n(x)$ は $j_n(x)$ とまったく同じ漸化式 (8.24) を満足している．ノイマン関数 $N_n(x)$ と同様にこれを初期条件

$$\begin{aligned} n_0(x) &= -x^{-1}\cos x \\ n_1(x) &= -x^{-2}(\cos x + x\sin x) \\ n_2(x) &= -x^{-3}[(3-x^2)\cos x + 3x\sin x] \end{aligned} \tag{8.25}$$

を用いて n の増える方向に計算していけばよい．

ルジャンドル多項式 ルジャンドル多項式はルジャンドル関数の特別な場合として定義することもできるが，ここでは次のように漸化式によって定義することにする．

$$\begin{aligned} &P_0(x)=1 \quad P_1(x)=x \quad |x|\le 1\\ &(n+1)P_{n+1}(x)-(2n+1)xP_n(x)\\ &\qquad +nP_{n-1}(x)=0 \end{aligned} \tag{8.26}$$

この漸化式は

$$\begin{aligned} P_{n+1}(x)-\left(2-\frac{1}{n+1}\right)xP_n(x)\\ +\left(1-\frac{1}{n+1}\right)P_{n-1}(x)=0 \end{aligned}$$

と書き換えてみればわかるように，$n\to\infty$ では $x=\cos\theta$ とすれば三角関数の漸化式 (8.7) と同じ形になるので，(8.23) 式はルジャンドル多項式を計算するために順方向に用いても問題ない．

ゲルツェルの方法 フーリエ級数を計算するには与えられた数列 c_n に対して

$$S=\sum_{n=0}^{N}c_ne^{in\theta} \tag{8.27}$$

という形の級数を計算する必要がある．いま

$$z=e^{i\theta}$$

と置くと S は z の N 次式であるから，多項式の計算法の定石にしたがって

$$S=c_0+z(c_1+z(c_2+\cdots+z(c_N)\cdots)$$

と書いて，外から n 番目の括弧の中を Z_n と書くことにすると

$$S=Z_0=c_0+zZ_1 \quad Z_1=c_1+zZ_2\cdots$$

一般に

$$Z_k=c_k+zZ_{k+1}$$

が成り立つ．一番最後の項は

$$Z_N=c_N$$

であるから，次のような計算方式が得られる．

$$\begin{aligned} &Z_{N+1}=0\\ &Z_k=c_k+zZ_{k+1} \quad k=N,N-1,\cdots,0 \end{aligned} \tag{8.28}$$

最後に得られた Z_0 が求める S である．c_n が複素数のときにはここまでである．

c_n が実数 $c_k=a_k$ のときにはさらに簡単になる．漸化式 (8.28) から

$$Z_{k+1}=z^{-1}(Z_k-a_k) \tag{a}$$

$$Z_{k-1}=zZ_k+a_{k-1} \tag{b}$$

したがって

$$Z_{k-1}+Z_{k+1}=(z+z^{-1})Z_k+a_{k-1}-z^{-1}a_k \tag{c}$$

が成り立つ．ここで Z_k を実数部と虚数部に分けて

$$Z_k=U_k+V_k\sin\theta \tag{8.29}$$

と置くと，(c) の虚数部から

$$\begin{aligned} &V_N=0 \quad V_{N+1}=0\\ &V_{k-1}+V_{k+1}=2\cos\theta V_k+a_k\\ &\quad k=N,N-1,\cdots,1 \end{aligned} \tag{8.30}$$

という漸化式が得られる．$V_N=0$ は a_N が実数であるためである．$V_0\sin\theta$ が S の虚数部である．実数部 U_0 を求めるためには (a) 式で $k=0$ としたときの虚数部から

$$U_0=-V_1+V_0\cos\theta+a_0 \tag{8.31}$$

として求められる．あるいは (8.30) 式を $k=0$ まで計算して

$$U_0=V_{-1}-V_0\cos\theta$$

としてもよい．

以上をまとめると，a_k が実数のときには漸化式 (8.30) から V_0，V_1 を計算するだけで

$$\sum_{n=0}^{N}a_k\cos n\theta \qquad \sum_{n=0}^{N}a_n\sin n\theta$$

の両方が計算できることになる．

クレンショーの漸化式 漸化式

$$p_{n+1}(x) = \alpha_n p_n(x) + \beta_{n-1} p_{n-1}(x) \qquad (8.32)$$

を満たすある関数 $p_n(x)$ の積和

$$S = \sum_{n=0}^{N} a_n p_n(x) \qquad (8.33)$$

を求めることを考える．常識的な方法は (8.32) 式から $p_{n+1}(x)$ を計算して係数 a_{n+1} を掛けて加えるという方法である．これはフーリエ級数でいえば，$\sin n\theta$ あるいは $\cos n\theta$ を漸化式 (8.7) で計算して係数 a_n を加えて足すという方法に相当している．しかし前項のゲルツェルの方法では三角関数の漸化式と積和が一つの漸化式にまとめられている．これは上の積和の計算のヒントになる．

改めてフーリエ級数

$$S_n = \sum_{k=0}^{n} a_k V_k \qquad V_n = \sin n\theta \qquad \text{(d)}$$

を考える．ここで V_n は漸化式

$$V_{n-1} + V_{n+1} = 2\cos\theta V_n \qquad \text{(e)}$$

を満たしている．部分和 S_n まで計算したときに次の部分和を求めるときには上式から V_{n+1} を求め

$$S_{n+1} = S_n + a_{n+1} V_{n+1} \qquad \text{(f)}$$

とするのが常識的なやりかたである．これを情報の流れとして図式化したのが Fig. 8.1 である．この図で節点はそこに現れるデータを表し，矢印はデータの流れる方向を示している．データが箱を通るときそこに書いてある数値が掛けられ，節点に合流するときには合流するすべてのデータが加えられる．α は $2\cos\theta$，β は -1 である．

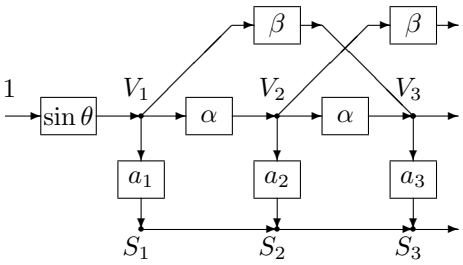


Fig. 8.1

この図の左端から数値 1 を入れると V_1, V_2 などが計算され，a_n が掛けられて一番下の線上の部分和に加えられていく．この図の右側は省略してあるが，下の線の右端から S_N が出てくることは明らかであろう．

そこで同じ流れ図をゲルツェルの方法 (8.30) 式について書いてみると，実は Fig. 8.1 とまったく同じ形になる．二つの図で異なるのはゲルツェルの方法では上の図の V_1, V_2, V_3 のところがそれぞれ $V_0, V_1 V_2$ と番号が一つずれていることと，もっと重要なことは，矢印の向きがすべて反対になっていることである．すなわち，ゲルツェルのアルゴリズムは Fig. 8.1 図の下の線の右端から数値 1 を入れると左端から $V_0 \sin\theta$ が得られるという構造になっている．

右側から 1 を入れたときの出力が左側から 1 を入れたときの出力と同じになることは，ある a_k に注目したときにこれにどのような数が掛けられるかを考えてみればわかる．もちろん両方の値が同じになることは数式的に証明することができる．

式 (d) の $\sin n\theta$ を $\cos n\theta$ に変えたときの流れ図は Fig. 8.1 の V_1 より右側ではまったく同じである．したがって右側から計算を行なう場合には V_1 のところまでは sin のときとまったく同じで，左端に cos に対応した一段を加えることによって (8.31) 式が導かれる．

以上をまとめれば，順方向の計算式を流れ図に書き，それを逆方向にたどることにより，漸化式と積和を一つの公式にまとめることができる．

漸化式 (8.32) と積和 (8.33) のときの流れ図は Fig. 8.2 になる．今度は省略なしに右端まで書いてある．q_n は右に辿るときには $p_n(x)$ であるが，左に辿るときには p_n と混同を避けるために記号を変えてある．sin, cos のときと違って左端が複雑であるが，これを逆方向にたどった式はつぎの漸化式になる．

$$\begin{aligned} &q_{N+1} = q_{N+2} = 0 \\ &q_n = a_n + \alpha_n q_{n+1} + \beta_n q_{n+2} \\ &\qquad n = N, N-1, \cdots, 1 \end{aligned} \qquad (8.34)$$

q_1 まで求められた後，図の左端の計算を行なって

$$\begin{aligned} &q_0 = a_0 + \beta_0 q_2 \\ &S = q_0 p_0(x) + q_1 p_1(x) \end{aligned} \qquad (8.35)$$

が得られる．ゲルツェルの方法はこの特別な場合である．この計算法はルジャンドル関数などについても適用できる．

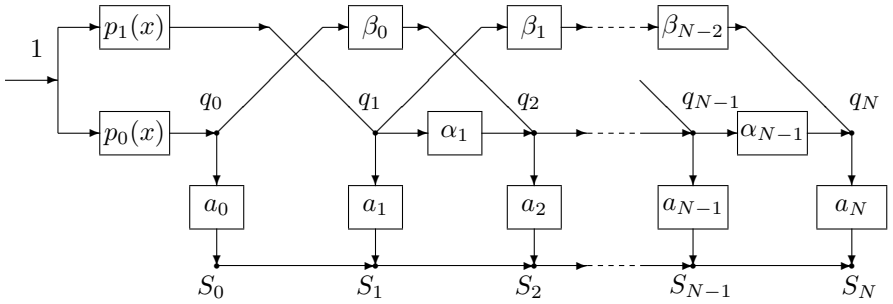


Fig. 8.2

非斉次差分方程式　これまでは斉次の差分方程式だけを考えてきたが，次に方程式 (8.3) の右辺に外力項が加わった非斉次の方程式を

$$y_n + \alpha y_{n-1} + \beta y_{n-2} = x_n \tag{8.36}$$

を考える．便宜上，左辺の y の添字をずらしてある．外力が $n = 0$ から働き始めたとすれば

$$x_n = 0 \quad y_n = 0 \quad n < 0$$

と仮定してよい．上式 (8.36) の解を閉じた形に表すことは不可能ではないが，あまり意味はない．ここでは外力が有界のとき，解が $n \to \infty$ で有界になる場合だけを考える．

y_n を解く前に，外力がインパルスのときの解 h_n を求めておく．これは差分方程式

$$h_n + \alpha h_{n-1} + \beta h_{n-2} = \delta_n = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases} \tag{8.37}$$

の解である．δ_n は離散的なデルタ関数である．h_n は系のインパルス応答と呼ばれる．h_n は実は上式を解かなくても，この式の右辺を 0 にした斉次方程式を，初期条件

$$h_n = \begin{cases} 1 & n = 0 \\ 0 & n < 0 \end{cases}$$

で解くことによって求められる．この解が $n \to \infty$ で有界であるためには特性根が (8.6) 式の条件を満たしていなければならない．この条件が満たされれば h_n は順方向の漸化式

$$h_n = -\alpha h_{n-1} - \beta h_{n-2} \quad n = 1, 2, \cdots$$

で求めることができる．この解が

$$\sum_{n=0}^{\infty} |h_n| < \infty \tag{8.38}$$

を満足することは一般解 (8.4) 式と条件 (8.6) 式から簡単に証明することができる．

次に外力 x_n が

$$x_n = x_0 \delta_n$$

のときの解を求める．これは振幅が x_0 のインパルスが働いたときの解であるから，明らかに

$$y_n = x_0 h_n$$

で表される．外力が時刻 $n = k$ に働く振幅 x_k のインパルスであるときの解は，時刻 k だけ遅れて応答が現れるから

$$y_n = x_k h_{n-k}$$

と表される．したがって一般に時刻 $0, 1, 2, \cdots$ に振幅 $x_0, x_1, x_2 \cdots$ のパルスが次々に働いたときの解は，それぞれのパルスによる解を重ね合わせて

$$y_n = \sum_{k=0}^{n} x_k h_{n-k} = \sum_{k=0}^{n} h_k x_{n-k} \tag{8.39}$$

である．右辺は x_n と h_n の離散的畳み込みである．この解が $n \to \infty$ で有界であることは

$$|y_n| \leq \sum_{k=0}^{n} |h_k||x_{n-k}| \leq \max |x_k| \sum_{k=0}^{\infty} |h_k|$$

と (8.38) 式から明らかである．

実は，実際の計算ではこのような持って回ったやり方はしない．条件 (8.6) 式が満足されているとき，(8.36) 式の解は順方向の漸化式

$$y_n = x_n - \alpha y_{n-1} - \beta y_{n-2} \tag{8.40}$$
$$n = 0, 1, 2, \cdots$$

で簡単に求めることができる．この解が (8.39) 式に等しくなることは数学的帰納法で証明することができる．

(8.39)，(8.40) 式はディジタルフィルターの計算式にほかならない．ディジタルフィルターについては別の節で詳しく説明する．

9 離散的フーリエ変換

離散的フーリエ変換 時間間隔 Δt で読み取られた時系列 x_j が

$$\sum_{j=-\infty}^{\infty} |x_j|^2 < \infty \tag{9.1}$$

を満足していると仮定する．二乗和は時系列の全エネルギーに相当する量であるから，この条件を満たす時系列を，全エネルギー有限なデータと呼ぶことにする．x_j が有限の長さであればもちろんこの条件は満たされている．しかし微動のように時間的に減衰しないで無限に続くデータはこの条件を満たしていない．このようなデータについては別に扱う (§14)．

時系列 x_j に対して

$$X_D(\omega) = \sum_{j=-\infty}^{\infty} x_j e^{ij\omega} \tag{9.2}$$

を x_j の離散的フーリエ変換と呼ぶ．ω は無次元の角周波数であり，実周波数を f，x_j のサンプル間隔を Δt とすると

$$\omega = 2\pi f \Delta t \tag{9.3}$$

で定義される．明らかに $X_D(\omega)$ は周期関数

$$X_D(\omega + 2\pi) = X_D(\omega) \tag{9.4}$$

で，したがって角周波数は

$$|\omega| \le \pi \tag{9.5}$$

の範囲だけを考えれば十分である．実周波数では

$$|f| \le f_N = \frac{1}{2\Delta t} \tag{9.6}$$

の範囲に限られる．この最大の周波数をナイキスト周波数という．分母が $2\Delta t$ であるから，離散的データで表現できる最短の周期は Δt ではなく $2\Delta t$ である．

x_j が (9.1) 式ではなく

$$\sum_{-\infty}^{\infty} |x_j| < \infty$$

を満たしていれば (9.2) 式の和が収束することは明らかである．しかし上式が満たされていなくても，(9.1) 式が満たされていれば

$$\lim_{N\to\infty} \int_{-\pi}^{\pi} \Big| \sum_{j=-N}^{N} x_j e^{ij\omega} - X_D(\omega) \Big|^2 d\omega = 0$$

を満足する $X_D(\omega)$ が存在する．このような収束を，二乗積分の意味での収束と呼ぶが，ここではこのような意味も含めて (9.2) 式のように表すことにする．

離散的フーリエ変換の逆変換は

$$x_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} X_D(\omega) e^{-ij\omega} d\omega \tag{9.7}$$

である．この式は (9.2) 式を代入して和と積分の順序を入れかえれば簡単に証明できる．上式は x_j が単振動 $e^{-ij\omega}$ の重ね合わせとして表されることを意味している．(9.3) 式を用いれば上式は

$$x_j = \int_{-f_N}^{f_N} \Delta t X_D(\omega) e^{-ij\omega} df$$

と書きかえることができるから，$\Delta t X_D(\omega)$ が単位周波数当たりの重みである．これをスペクトル密度，あるいは単にスペクトルと呼ぶ．

(9.2)，(9.7) 式は §5 で考えたフーリエ級数 (5.16) 式と反対の関係になっている．そこでは連続変数 x の関数 $f(x)$ から離散的な係数 (a_k, b_k) を定義していた．

エイリアシング ディジタルデータ x_j はアナログデータ $x(t)$ を時間間隔 Δt おきにサンプルすることによって得られる．Δt は $x(t)$ に含まれている周期よりも短く選ばなければならないのは当然である．もし粗い周期でサンプリングを行えば下の図のようなことが起きる．この現象をエイリアシング (aliasing) という．

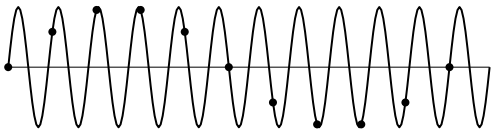


Fig. 9.1 短周期の振動を粗い間隔でサンプルすると (黒丸)，長周期に見える．

$x(t)$ のスペクトルは通常のフーリエ変換

$$X(\sigma) = \int_{-\infty}^{\infty} x(t) e^{i\sigma t} dt \tag{9.8}$$

で定義される．σ は実角周波数で

$$\sigma = 2\pi f$$

である．この逆変換は

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\sigma)e^{-i\sigma t}d\sigma \tag{9.9}$$

で表される．

問題は離散的なスペクトル $X_D(\omega)$ がもともとのスペクトル $X(\omega)$ に一致するかどうかということである．それを調べるには

$$x_j = x(j\Delta t)$$

を (9.2) 式の右辺に代入した

$$\sum_{j=-\infty}^{\infty} x(j\Delta t)e^{ij\omega}$$

を計算してみればよい．右辺に $x(t)$ のフーリエ逆変換 (9.9) 式を代入しても計算することができるが，ここではポアソンの和の公式を用いる．

いま，$f(t)$ のフーリエ変換を $F(\sigma)$ とするとき，$\alpha\beta = 2\pi$ を満たす任意の α, β に対して

$$\sum_{m=-\infty}^{\infty} f(m\alpha) = \sqrt{\frac{\beta}{2\pi\alpha}}\sum_{n=-\infty}^{\infty} F(n\beta) \tag{9.10}$$

が成り立つ．これがポアソンの和の公式である．$f(t)$ として

$$f(t) = x(t)e^{i\omega t/\Delta t}$$

を選ぶと $\sum_j f(j\Delta t)$ が先の和に等しくなる．この $f(t)$ のフーリエ変換は

$$F(\sigma) = X(\sigma + \omega/\Delta t)$$

である．そこで $\alpha = \Delta t$ と選ぶと，ポアソンの和の公式により

$$\begin{aligned} X_D(\omega) &= \sum_{m=-\infty}^{\infty} x(m\Delta t)e^{im\omega} \\ &= \frac{1}{\Delta t}\sum_{n=-\infty}^{\infty} X\left(\frac{\omega + 2\pi n}{\Delta t}\right) \end{aligned}$$

が成り立つ．したがって，もとのデータのスペクトル $X(\sigma)$ と離散的なスペクトル $X_D(\omega)$ の間には

$$\Delta t X_D(\omega) = \sum_{n=-\infty}^{\infty} X\left(\frac{\omega + 2\pi n}{\Delta t}\right) \tag{9.11}$$

の関係があることがわかった．

$X_D(\omega)$ の引数 ω は無次元の角周波数であり，$X(\sigma)$ の引数 σ は次元をもった角周波数であるので，上式ははわかり難くなっている．そこで X の引数も無次元にして，上の関係を図で示したのが Fig. 9.2 である．上段がアナログのスペクトル $X(\omega)$，下段がディジタルのスペクトル $X_D(\omega)$ を表している．無次元の角周波数で考えたとき，アナログのスペクトル $X(\omega)$ の $\omega > \pi$ の部分は $-\pi < \omega$ の部分に移動しており，$X(\omega)$ の $\omega < -\pi$ の部分が $-\pi < \omega$ の部分に移動して，それらが本来あった $X(\omega)$ に加えあわされている．したがって一般には実周波数 f におけるディジタルスペクトルは同じ周波数のアナログスペクトルには等しくない．$X_D(\omega)$ は周期関数になるが，この図では $|\omega| > \pi$ の部分は省略してある．

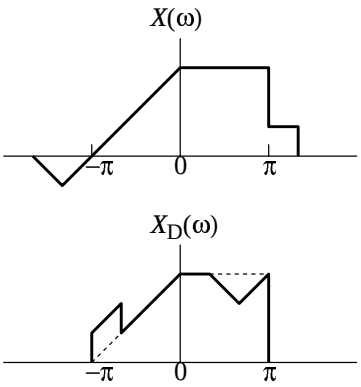


Fig. 9.2

実周波数 f で考えると，周波数 f におけるディジタルスペクトルはアナログスペクトルの f, $f \pm 1/\Delta t$, $f \pm 2/\Delta t, \cdots, f \pm n/\Delta t, \cdots$ におけるスペクトルの和になっている．これは逆に言えば高周波成分 $f \pm 2nf_N$ が低周波成分 f に見えてしまうことになる．これが Fig. 9.1 である．具体的な例をあげる．$f = 13$ Hz のデータを $\Delta t = 0.1$ s で読みとると，$f_N = 5$ Hz であるから，実際に見える周波数は $f = 13 - 2\times 5 = 3$ Hz になる．もちろん 3 Hz の成分がもともと含まれていればこれらのスペクトルの和になる．

式からも図からもわかるように，もし $X(\sigma)$ がナイキスト周波数以上で 0 であるならば，$X_D(\omega)$ と $X(\sigma)$ は等しくなる．すなわち

$$\Delta t X_D(\omega) = X(\omega/\Delta t) \tag{9.12}$$

$$|\sigma| > \frac{\pi}{\Delta t} \quad \text{で} \quad X(\sigma) = 0 \text{ のとき}$$

これをサンプリング定理という．アナログデータをサンプルするときには，ナイキスト周波数以上の成分が含まれないようにしておかなければならない．

以下では離散的なスペクトルを主に考えるので，$X_D(\omega)$ の添字 D を省略して単に $X(\omega)$ を書くことにする．

積のフーリエ変換　二つの時系列 x_j, y_j の積

$$z_j = x_j y_j \tag{9.13}$$

のフーリエ変換を $Z(\omega)$ とすると

$$Z(\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\sigma)Y(\omega-\sigma)d\sigma \tag{9.14}$$

が成り立つ．右辺は畳み込みである．z_j は x_j と y_j に関して対称であるから，右辺の X と Y は入れかえてもよい．

畳み込みのフーリエ変換　今度は

$$z_j = \sum_{k=-\infty}^{\infty} x_k y_{j-k} = \sum_{k=-\infty}^{\infty} x_{j-k} y_k \tag{9.15}$$

と定義する．右辺は離散的な畳み込みである．z_j のフーリエ変換は

$$Z(\omega) = X(\omega)Y(\omega) \tag{9.16}$$

である．

パーセバルの等式　(9.15) 式において

$$y_j = \overline{x_{-j}}$$

と置く．$\overline{\{\cdots\}}$ は複素共役を表す．このとき

$$z_j = \sum_k x_k \overline{x_{k-j}} = \sum_k x_{j+k}\overline{x_k}$$

である．これを x_k の自己相関関数という．y_j のフーリエ変換は $Y(\omega) = \overline{X(\omega)}$ であるから z_j のスペクトルは $Z(\omega) = |X(\omega)|^2$ である．よって

$$z_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} |X(\omega)|^2 e^{-ij\omega} d\omega$$

が成り立つ．ここで $j = 0$ と置けば

$$z_0 = \sum_{k=-\infty}^{\infty} |x_k|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |X(\omega)|^2 d\omega \tag{9.17}$$

が得られる．これをパーセバルの等式という．左辺は波形 x_j の全エネルギーを意味している．右辺は全エネルギーが単位周波数当たり $\Delta t|X(\omega)|^2$ の密度で分布していることを示している．

(9.2)，(9.7) 式や，通常のフーリエ変換 (9.8)，(9.9) 式の指数部の符号の取り方は一定していない．ここでの符号の取り方は物理学の慣用に従っている．工学系では上と符号を反対にしたやり方もよく用いられている．

有限フーリエ級数　ここまではデータが無限に続くことを想定していた．データが有限長のときにももちろんこれまでの議論は成立する．しかし有限個のときには話は簡単になる．

いま，N 個のデータ $x_0, x_1, \cdots, x_{N-1}$ が与えられたとする．このデータを用いて与えられた点を通る三角関数を用いた内挿式を作ることができる．データ間の補間をしなくてもよいのなら次のように簡単になる．

まず，次式で係数 c_k を計算する．

$$c_k = \sum_{j=0}^{N-1} x_j \exp\left(\frac{2\pi ijk}{N}\right) \qquad k = 0, 1, 2, \cdots, N-1 \tag{9.18}$$

この係数を用いると，もとのデータは

$$x_j = \frac{1}{N}\sum_{k=0}^{N-1} c_k \exp\left(-\frac{2\pi ijk}{N}\right) \tag{9.19}$$

によって表される．これが正しいことは上式の右辺に (9.18) 式を代入して

$$\sum_{k=0}^{N-1} \exp\left[\frac{2\pi ik(l-j)}{N}\right] = \begin{cases} N & l = j \\ 0 & l \neq j \end{cases}$$

を利用すれば示すことができる．

ところで (9.18) 式の和は (9.2) 式の和と全く同じ形をしている．いま

$$\omega_k = \frac{2\pi}{N}k$$

と置けば

$$c_k = X(\omega_k) \tag{9.20}$$

すなわち，$X(\omega)$ を間隔

$$\Delta\omega = \frac{2\pi}{N}$$

でサンプルしたものが c_k である．それでは逆変換 (9.7) 式と (9.19) 式はどのような関係になっているのであろうか．

$X(\omega)$ は周期関数であるから，積分 (9.7) 式の積分範囲を $(0, 2\pi)$ にしても同じである．この積分範囲を N 等分して台形公式で近似すれば

$$\begin{aligned}\frac{1}{2\pi}\int_0^{2\pi} X(\omega)e^{-ij\omega}d\omega &\doteqdot \frac{\Delta\omega}{2\pi}{\sum_{k=0}^{N}}' X(\omega_k)e^{-ij\omega_k} \\ &= \frac{\Delta\omega}{2\pi}\sum_{k=0}^{N-1} X(\omega_k)e^{-ij\omega_k}\end{aligned}$$

となる．$\sum'$ は $k = 0, N$ の項に台形公式からくる重み $1/2$ を掛けることを意味し，最後の等式は $X(\omega)$ が周期関数 $X(0) = X(\omega_N)$ であることから成り立つ．$X(\omega_k) = c_k$ であるから，最後の式は実は (9.19) 式に等しい．すなわち，有限長のデータの場合，離散的フーリエ変換の逆変換 (9.7) 式を台形法則で近似すれば正しい値が得られる．ただし，間隔 $\Delta\omega$ はデータの長さに応じて正しく選ばなければならない．

上の議論は N が偶数でも奇数でも成り立つが，どちらかといえば N が偶数のときに使いやすい公式である．N が奇数，$N = 2M + 1$ のときには時刻の原点をデータの中心に取った方が対称性のよい公式が得られる．

$2M+1$ 個のデータ x_j $(|j| \le M)$ が与えられたとき，離散的フーリエ変換は

$$X(\omega) = \sum_{j=-M}^{M} x_j e^{ij\omega}$$

であるから，有限フーリエ級数の係数は

$$c_k = X(\omega_k) \quad \omega_k = \frac{2\pi k}{2M} \quad |k| \le M \tag{9.21}$$

で与えられる．逆変換は

$$x_j = \frac{1}{2M}{\sum_{k=-M}^{M}}' c_k e^{-ij\omega_k} \tag{9.22}$$

となる．$\sum'$ は $j = -M$ と $j = M$ に係数 $1/2$ を掛けることを意味している．この関係は上式に (9.21) 式の c_k を代入することによって証明することができるが，$X(\omega)$ の逆変換 (9.7) 式の積分範囲 $(-\pi, \pi)$ を $2M$ 等分して台形公式で計算したものになっている．

時間軸上のエイリアシング　周波数軸上の間隔 $\Delta\omega$ を正しく選ばないとどうなるか．畳み込み (9.15) 式において x_j, y_j がともに $j = 0 \sim N-1$ の N 個であったとする．いま z_j を時間軸上の畳み込み (9.15) 式ではなく，周波数軸上の関係 (9.16) 式を用いて，逆変換

$$z_j = \frac{1}{2\pi}\int_0^{2\pi} X(\omega)Y(\omega)e^{-ij\omega}d\omega$$

によって計算することを考える．これは後で述べる FFT の計算法によって周波数軸上の計算の方が速いからである．x_j, y_j の長さに合わせて積分区間を N 等分して台形公式で計算すると図の一番下のようになる．これは中段の正しい z_j とは似ても似つかない形をしている．

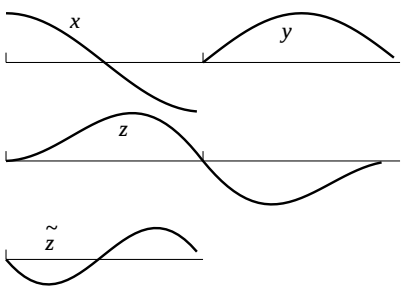


Fig. 9.3 x, y は N 個のデータ，z は x と y の畳み込み．x と y のスペクトルの積を N 点の台形公式で計算したもの．

その理由は時間軸上でサンプルしたときに周波数軸上で生じたエイリアシングとまったく同じである．z_j は $j = 0 \sim 2N-2$ に値をもつ．しかし N 等分して計算した z_j，これを $\tilde{z}_j$ とすると，これは N 個の独立な成分しかなく，周期関数 $\tilde{z}_{j+N} = \tilde{z}_j$ になっている．このため本来の z_j の $j = N \sim 2N-1$ の部分が $j = 0 \sim N-1$ に折り畳まれている．式で書けば

$$\tilde{z}_j = \sum_{n=-\infty}^{\infty} z_{j+nN}$$

である．したがって正しい z_j を求めるためには z_j の長さよりも長い数，いまの例では $2N$ を用いて台形公式を用いなければならない．

実数データ　データ x_j が実数のときには (9.18), (9.19) 式は実数だけで表すことができる．ただし，こうすると対称性が悪くなる．特に，N が偶数か奇数かで式の形が異なってしまう．ここではまず N が偶数の場合を示す．

まず，N の偶奇によらず，x_j が実数のときには (9.18) 式の c_k は

$$c_{N-k} = \overline{c_k}$$

を満たしている．したがって N が偶数のときには c_k は $k = 0 \sim N/2$ だけが独立である．c_k を実数部と虚数部にわけて

$$c_k = a_k + ib_k$$

とすると a_k, b_k は

$$\begin{aligned} a_k &= \sum_{j=0}^{N-1} x_j \cos j\omega_k \qquad \omega_k = \frac{2\pi}{N}k \\ b_k &= \sum_{j=0}^{N-1} x_j \sin j\omega_k \\ & k = 0, 1, 2, \cdots, N/2 \end{aligned} \tag{9.23}$$

で与えられる．b_0，$b_{N/2}$ はつねに 0 であるから，a_k, b_k は合計 N 個あり，これはデータの個数 N に一致する．

次に逆変換 (9.19) 式を書きかえる．k についての和二つに分けて

$$\sum_{k=0}^{N-1} c_k \cdots = \sum_{k=0}^{N/2-1} c_k \cdots + \sum_{k=N/2}^{N-1} c_k \cdots$$

とすると，第二項は

$$\begin{aligned} \sum_{k=N/2}^{N-1} c_k \cdots &= \sum_{k=1}^{N/2} c_{N-k} \exp\left[-\frac{2\pi ij(N-k)}{N}\right] \\ &= \sum_{k=1}^{N/2} \overline{c_k \exp\left(-\frac{2\pi ijk}{N}\right)} \end{aligned}$$

と書きかえられる．よって

$$\begin{aligned} x_j = \frac{2}{N}\Big[&\frac{1}{2}a_0 + \sum_{k=0}^{N/2-1} (a_k \cos j\omega_k + b_k \sin j\omega_k) \\ &+ \frac{1}{2}a_{N/2} \cos 2\pi j\Big] \end{aligned} \tag{9.24}$$

が得られた．N が奇数のときには上式の k についての和の上限を $N/2$ の整数部分とし，最後の $a_{N/2}$ の項を無視すればよい．ただし N が奇数のときには角周波数範囲 $(0, 2\pi)$ を奇数区間に分割することになるので使い勝手が悪い．

N が奇数 $2M+1$ のときには先に導いた係数 (9.21) 式から

$$\begin{aligned} a_k &= \sum_{j=-M}^{M} x_j \cos j\omega_k \qquad \omega_k = \frac{2\pi k}{2M} \\ b_k &= \sum_{j=-M}^{M} x_j \sin j\omega_k \end{aligned} \tag{9.25}$$

とすれば x_j は

$$\begin{aligned} x_j = \frac{1}{M}\Big[&\frac{1}{2}a_0 + \sum_{k=1}^{M-1} (a_k \cos j\omega_k + b_k \sin j\omega_k) \\ &+ \frac{1}{2}a_M \cos j\pi\Big] \end{aligned} \tag{9.26}$$

と表される．

ウィンドウ　$j = -\infty \sim +\infty$ で x_j が与えられているとき，実際上は (9.2) 式で計算するわけではなく，どこかで打ち切って

$$X_M(\omega) = \sum_{j=-M}^{M} x_j e^{ij\omega}$$

とするほかはない．そのために計算されたスペクトルには歪みが生じる．$X(\omega)$ と $X_M(\omega)$ の関係は簡単に導くことができる．

上式の x_j に逆変換 (9.7) 式を代入すると

$$\begin{aligned} X_M(\omega) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\sigma) \sum_{j=-M}^{M} e^{ij(\omega-\sigma)} d\sigma \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\sigma) W_M(\omega-\sigma) d\sigma \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\omega-\sigma) W_M(\sigma) d\sigma \end{aligned} \tag{9.27}$$

が得られる．ここに

$$W_M(\omega) = \sum_{j=-M}^{M} e^{ij\omega} = \frac{\sin(N+1/2)\omega}{\sin\omega/2} \tag{9.28}$$

である．

(9.27) 式の積分は畳み込み，すなわち移動平均である．移動平均の重み関数 $W_M(\omega)$ は Fig. 9.4 のような形をしている．$X_M(\omega)$ の値を求めたいと

きには，$W_M(\sigma)$ の中心を ω のところに合わせて，$X(\omega-\sigma)$ と $W_M(\sigma)$ の掛け算をして積分するという操作をする．したがって，$X_M(\omega)$ が $X(\omega)$ に等しくなるには $W_M(\sigma)$ はデルタ関数にできるでけ近くなければならない．そのためには $W_M(\sigma)$ の中心のピーク (メインローブ) の幅が狭く，高さが高く，それ以外の山谷 (サイドローブ) の振幅が小さくなければならない．この図の例では，最初のサイドローブの谷がメインローブの約 20 %もあるので，本来正であるべき $X(\omega)$ が積分の結果負になってしまうようなことも起こりかねない．§5 のギブスの振動はまさにこのために生じたものである．

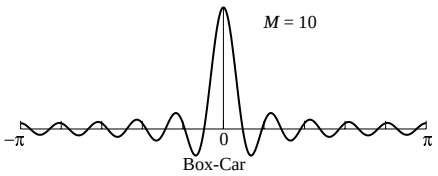


Fig. 9.4 単純打ち切りのスペクトルウィンドウ

データを単純に打ち切るかわりに少し工夫をする．いま，$|j|>M$ で 0 になる関数 w_j を用いて

$$z_j=\begin{cases} w_jx_j & |j|\le M \\ 0 & |j|>M \end{cases} \tag{9.29}$$

とすれば，積のフーリエ変換 (9.14) 式により z_j のスペクトルは

$$Z(\omega)=\frac{1}{2\pi}\int_{-\pi}^{\pi}W_M(\sigma)X(\omega-\sigma)d\sigma \tag{9.30}$$

で表される．$W_M(\omega)$ は w_j のスペクトルである．

ここでの目的はスペクトル $X(\omega)$ を推定することであるから，z_j が x_j に似ていなくても $Z(\omega)$ が $X(\omega)$ に似ていれば目的は達せられる．このために用いられる関数 w_j をデータウィンドウといい，そのフーリエ変換 $W_M(\omega)$ をスペクトルウィンドウという．はじめに示した単純に打ち切ってしまう方法はデータウィンドウとして

$$w_j=\begin{cases} 1 & |j|\le M \\ 0 & |j|>M \end{cases}$$

を用いたことに相当している．以下によく用いられるウィンドウを示す ($|j|>M$ で $w_j=0$ は以下では省略する)．

三角形ウィンドウ (Bartlett)　単純な打ち切りは四角形のウィンドウを掛けたことに相当する．四角形のかわりに三角形を用いたウィンドウ

$$w_j=1-\frac{|j|}{M}\qquad |j|\le M$$
$$W_M(\omega)=M\left(\frac{\sin M\omega/2}{M\sin\omega/2}\right)^2 \tag{9.31}$$

がある．$W_M(\omega)$ が最初に 0 になるのは $\omega=2\pi/M$ で，これは四角形ウィンドウの約 2 倍である．しかし (9.31) 式の $W_M(\omega)$ は二乗の形をしているので，Fig. 9.5 の上段に見られるようにサイドローブの高さは四角形のウィンドウに比べてはるかに低い．

ハンのウィンドウ（Cosine-bell）　三角形のウィンドウには角がある (微係数が不連続) のが気持ちが悪い．そこで滑らかな関数を用いたものにハン (人名) のウィンドウがある．

$$w_j=\frac{1}{2}\left(1+\cos\frac{\pi j}{M}\right)\qquad |j|\le M$$
$$W_M(\omega)=\frac{\sin M\omega\cos\omega/2}{2\sin\omega/2}\times\left[1-\left(\frac{\sin\omega/2}{\sin\pi/2M}\right)^2\right]^{-1} \tag{9.32}$$

これは $M\omega=\pi$ に零点があるようにみえるが，この零点は分母の零点とキャンセルするので，最初の零点は三角形ウィンドウと同じ $\omega=2\pi/M$ である．サイドローブの高さは三角形ウィンドウよりも低いが，ここでは負のサイドローブになっている (Fig. 9.5 の中段)．

ハミングのウィンドウ　ハンのウィンドウを少し一般化して

$$w_j=\begin{cases} a+\dfrac{b}{2}\left(1+\cos\dfrac{\pi j}{M}\right) & |j|<M \\ \dfrac{1}{2}a & |j|=M \end{cases}$$
$$W_M(\omega)=\frac{\sin M\omega\cos\omega/2}{\sin\omega/2}\times\left\{a+\frac{b}{2}\left[1-\left(\frac{\sin\omega/2}{\sin\pi/2M}\right)^2\right]^{-1}\right\} \tag{9.33}$$

とする．a, b はこれから決める定数である．ハンのウィンドウでは $|\omega|=5\pi/2M$ に最大のサイドロー

ブがある．そこでこれが 0 になるように a, b を決める．すなわち

$$W_M(5\pi/2M) = 0 \qquad w_0 = 1$$

から a, b を決める．第二式はほかのウィンドウと同様になるようにするためである．この式を解くことは容易であるが，$M > 10$ なら $M \to \infty$ のときの値

$$a = \frac{4}{46} \qquad \frac{b}{2} = \frac{21}{46} \tag{9.34}$$

を用いれば実用上十分である．これをハミング (人名) のウィンドウという (Fig. 9.5 の下段).

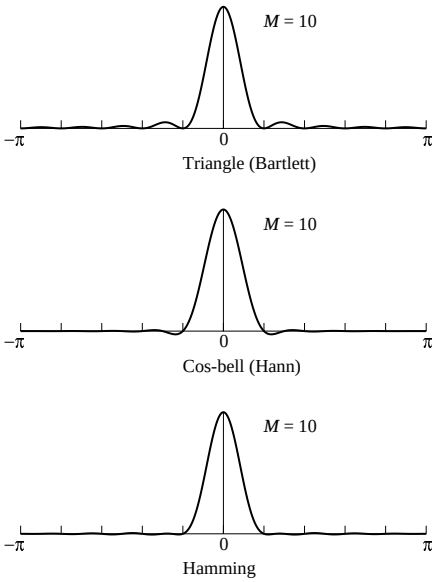


Fig. 9.5 上段：三角形ウィンドウ．中段：ハンのウィンドウ．下段：ハミングのウィンドウ．

ここでは $X(\omega)$ が未知であるが，スペクトルが既知のデータにウィンドウを掛けることがある．これはスペクトルの平滑化のためである．このときにはデータにウィンドウを掛けてからフーリエ変換してもいいが，いまは $X(\omega)$ わかっているのであるから，畳み込み (9.30) 式を台形公式で計算する，すなわち

$$Z(\omega_k) = \frac{1}{2M} \sum_{j=-M}^{M} W_M(\omega_j) X(\omega_{k-j})$$

とした方が簡単である．なぜなら，(3.31)，(3.32)，(3.33) 式の $W_M(\omega_j)$ は $|j| > 2$ ではすべて 0 になってしまうから，上の畳み込みは実は $j = 0, \pm 1$ だけの和になってしまうからである．単純な打ち切りに相当する (9.28) 式ではそうはなっていないが (Fig. 9.4 参照)，単純な打ち切りを少し補正して

$$w_j = \begin{cases} 1 & |j| < M \\ \dfrac{1}{2} & |j| = M \end{cases} \tag{9.35}$$

とすれば

$$W_M(\omega) = \frac{\sin M\omega \cos \omega/2}{\sin \omega/2}$$

となる．$j \neq 0$ なら $W_M(\omega_j) = 0$ となってしまうので，畳み込みは $j = 0$ だけ，つまり何もしないことになる．

パルツェンのウィンドウ　スペクトルを平滑化するのに僅か三点の移動平均ではいくらなんでも少なすぎる．そこでよく用いられるのが

$$w_j = \begin{cases} 1 - 6\left(\dfrac{j}{M}\right)^2 + 6\left|\dfrac{j}{M}\right|^3 & \\ \qquad\qquad 0 \le |j| \le M/2 & \\ \left(1 - \dfrac{|j|}{M}\right)^3 \qquad M/2 \le |j| \le M & \end{cases}$$

$$W_M(\omega) = \frac{3}{4} M \left(\frac{\sin M\omega/4}{M/2 \sin \omega/2}\right)^4 \tag{9.36}$$

である．ただし，w_j と $W_M(\omega)$ は正しく対応しておらず，$M \to \infty$ のときのみ正しい．これは (9.31) 式をさらに二乗したものになっているので，サイドローブは非常に低い．しかしメインローブの幅は Fig. 9.5 の二倍になっている．

FFT　データ x_j が実数のとき，ある k に対して a_k, b_k を求めるには $2N$ 回の積和と同じ数だけの三角関数の計算が必要になる．k は N 個あるから，すべての係数を計算するに要する計算量は N^2 になる．これは連立一次方程式の計算量 N^3 に比べて少ない．しかし，スペクトル解析のときの N は連立方程式のときの N に比べてはるかに大きいので，N^2 だからといって少ない計算量とはいえない．

フーリエ解析の需要は多かったので，むかしから計算法についてはいろいろな工夫が行われてきた．1965 年に Cooley and Tukey によって高速フーリエ変換アルゴリズム (FFT, Fast Fourier Transform) が発表されたとき，こんな方法は前から使っていたとか，あの方法の変形ではないか，などいろいろ言われたが，この論文の与えた衝撃ははかりしれない．FFT

の計算量は $2N\log_2 N$ であるから，$N = 2^{10} = 1024$ のとき計算量の比は $N/2\log_2 N = 1024/20 = 50$ 倍にもなる．

FFT の原理を簡単に説明する．いま，データ数 N が $N = L\cdot M$ と分解できたとする．このとき j と k は次のように書くことができる．

$$\begin{aligned} j &= l + mL \\ &\qquad 0 \le l < L, \quad 0 \le m < M \\ k &= n + pM \\ &\qquad - \le n < M, \quad 0 \le p < L \end{aligned} \tag{9.37}$$

すなわち，(l,m) を指定すれば j が一義的に決まる．これはデータ x_j が一次元の配列ではなく，L 行 M 列の二次元配列で与えられたと考えればわかりやすい．そこでデータ x_j を $x(l,m)$ と書くことにする．同様にスペクトル c_k は M 行 L 列の配列で与えられるとして $c(n,p)$ とする．$x(l,m)$ と $c(n,p)$ の行，列の数が反対になっていることに注意する．この表記を用いると

$$\frac{jk}{N} = \frac{ln}{N} + \frac{lp}{L} + \frac{mn}{M} + mp$$

となる．したがって

$$e(x) = \exp(2\pi i x)$$

と置くと，フーリエ変換 (9.18) 式に現れる指数関数の部分は

$$e\Big(\frac{jk}{N}\Big) = e\Big(\frac{ln}{N}\Big)e\Big(\frac{lp}{L}\Big)e\Big(\frac{mn}{M}\Big)$$

と書くことができる．先の式の最後の項 mp の部分は $e(mp) = 1$ によって必要ない．この表現を用いるとフーリエ変換 (9.18) 式は

$$\begin{aligned} c(n,p) = &\sum_{l=0}^{L-1} e\Big(\frac{lp}{L}\Big)\Big(\frac{ln}{N}\Big) \\ &\times \sum_{m=0}^{M-1} e\Big(\frac{mn}{M}\Big)x(l,m) \end{aligned} \tag{9.38}$$

と書くことができる．最も内側の和

$$\sum_{m=0}^{M-1} e\Big(\frac{mn}{M}\Big)x(l,m)$$

は，l を固定すると長さ M のデータのフーリエ変換である．これは二次元配列 $x(l,m)$ の l 行のフーリエ変換で，$0 \le n < M$ の M 個の係数が得られる．これを仮に

$$c'(l,n) = \sum_{m=0}^{M-1} e\Big(\frac{mn}{M}\Big)x(l,m)$$

と置く．$c'(l,n)$ は $L\times M$ の行列である．実はある行 l に対して M 個のフーリエ係数が求められれば，x の l 行は不要になるので，$c'(l,n)$ の l 行をここに上書きすることができるのであるが，ここでは別の配列として表しておく．

この行列 $c'(l,n)$ のすべての行が計算された後，(9.38) 式の外側の和

$$c''(p,n) = \sum_{l=0}^{L-1} e\Big(\frac{lp}{L}\Big)e\Big(\frac{ln}{N}\Big)c'(l,n)$$

を計算する．これは $c'(l,n)$ の n 列に回転因子 $e(ln/N)$ を掛けた長さ L のフーリエ変換である．これが $c_k = c(n,p)$ にほかならないから

$$c(n,p) = c''(p,n)$$

である．両辺の引数が反対になっている．したがって c'' から c を求めるには転置しなければならない．FFT の計算ではどこかの段階でこのような操作が必要になる．

上の計算の計算量を評価する．ただし三角関数の計算量は考えない．まず，$c'(l,m)$ の計算量は一行につき M^2 である．同様に $c''(p,n)$ の計算には一列につき L^2 である．これに回転因子の積を加えると，計算量の合計は

$$M^2L + L^2M + N = N(L + M + 1)$$

になる．これは (9.18) 式をそのまま計算したときの計算量 N^2 よりは少ない．

L や M がさらに因数分解できれば計算量はさらに減少する．データ数 N が

$$N = n_1 n_2 \cdots n_m$$

と因数分解できれば計算量は

$$N(n_1 + n_2 + \cdots + n_m)$$

になる．特に

$$N = 2^m$$

のときには計算量は

$$2mN = 2N\log_2 N \tag{9.39}$$

になる．

N が 2 の冪乗のとき $L=2$ のときには先の式から

$$c''(0,n) = c'(0,n) + e\Big(\frac{n}{N}\Big)c'(1,n)$$
$$c''(1,n) = c'(0,n) - e\Big(\frac{n}{N}\Big)c'(1,n) \quad (9.40)$$

が成り立つ．一つの c'' の計算は一回の積和で済む．ここで $c'(0,n)$ はもとのデータ x_j の偶数番目のデータだけを用いたフーリエ変換，$c'(1,n)$ は奇数番目のデータだけを用いたフーリエ変換である．ところで，c' の計算は，偶数番ごとのデータをさらに偶数番目ごとにサンプルしたデータと奇数番目ごとにサンプルしたデータのフーリエ変換で表すことができる．このように下へ下へと辿っていけば，すべての計算は (9.40) 式のように二つの値から別の二つの値を計算するという演算だけで済むことがわかる．

もう少し具体的に計算法を説明する．$N=2^3$ のとき，j, k を二進数で表す．たとえば j は

$$j = (j_2 j_1 j_0)_2 = j_2 2^2 + j_1 2^1 + j_0 2^0$$

と表す．$(\cdots)_2$ は二進数であることを意味する．同様に k も二進数で表すと

$$jk = j_0(k_2k_1k_0)_2 + 2j_1(k_1k_0)_2 + 2^2 j_2 k_0 \quad \mathrm{mod}(N)$$

となる．$\mathrm{mod}(N)$ は N の整数倍を無視することを意味している．この展開を用いると

$$\sum_{j=0}^{N-1} e\left(\frac{jk}{N}\right)x(j) = \sum_{j_0} e\Big(\frac{j_0k_2}{2}\Big)e\Big(\frac{j_0(k_1k_0)_2}{N}\Big) \times \sum_{j_1} e\Big(\frac{j_1k_1}{2}\Big)e\Big(\frac{2j_1k_0}{N}\Big) \times \sum_{j_2} e\Big(\frac{j_2k_0}{2}\Big)x(j_2j_1j_0) \quad (9.41)$$

と書くことができる．j についての和は j_0, j_1, j_2 についての和の積に分解できる．すべての和の下限は 0，上限は 1 である．

一番内側の j_2 についての和を具体的に書くと

$$k_0 = 0 \ : \ x(0j_1j_0) + x(1j_1j_0)$$
$$k_0 = 1 \ : \ x(0j_1j_0) - x(1j_1j_0)$$

である．x の引数は二進数で書いてある．この計算が終わると $x(0j_1j_0)$, $x(1j_1j_0)$ は不要になるので，ここにいま計算した値を上書きする．これを

$$x(k_0j_1j_0) = \sum_{j_2} e\Big(\frac{j_2k_0}{2}\Big)x(j_2j_1j_0)$$

と書く．ここで j_1 と j_0 はあらゆる組み合わせについて計算しなければならない．すなわち $(j_1j_0)_2$ の値は 0 から $(11)_2 = 3$ まで変化する．

j_1，j_0 についての和についても同様で

$$x(k_0k_1j_0) = \sum_{j_1} e\Big(\frac{j_1k_1}{2}\Big)e\Big(\frac{2j_1k_0}{N}\Big)x(k_0j_1j_0)$$
$$x(k_0k_1k_2) = \sum_{j_0} e\Big(\frac{j_0k_2}{2}\Big) \times e\left(\frac{j_0(k_1k_0)_2}{N}\right)x(k_0k_1j_0)$$

となる．最後の結果 $x(k_0k_1k_2)$ が求める $c(k)$ であるが，これは正しい順番に並んでいない．たとえば $c(1) = c(001)$ は x の方では $x(100) = x(4)$ の位置にある．そこで $x(k_0k_1k_2)$ を正しい $c(k_2k_1k_0)$ の位置に移しかえなければならない．

この計算にはもう一つ問題がある．最後の式では指数関数に現れる $(k_1k_0)_2$ と x の引数 $x(k_0k_1j_0)$ のビットの順番が反対になっている．したがって (k_1k_0) が与えられたとき x の引数を求めるためにはビットの順番を反対にした (k_0k_1) の値を求めなければならない．あらゆる k_0, k_1 の組み合わせに対してそのたびにビットの反転を行うのは手間がかかる．これを避けるのは簡単である．計算を始める前に $x(j)$ をビットを反転した順番に並べかえてしまう．すなわち $x(j_0j_1j_2)$ を求めておく．そうすれば上の計算は次のようにまとめられる．

$$x(j_0j_1k_0) = \sum_{j_2} e\Big(\frac{j_2k_0}{2}\Big)x(j_0j_1j_2)$$
$$x(j_0k_1k_0) = \sum_{j_1} e\Big(\frac{j_1k_1}{2}\Big)e\Big(\frac{2j_1k_0}{N}\Big)x(j_0j_1k_0)$$
$$x(k_2k_1k_0) = \sum_{j_0} e\Big(\frac{j_0k_2}{2}\Big) \times e\left(\frac{j_0(k_1k_0)_2}{N}\right)x(j_0k_1k_0) \quad (9.42)$$

最後の $x(k_2k_1k_0)$ が求めるフーリエ変換 $c(k)$ そのものである．この方法を Cooley-Tukey のアルゴリズムという．

以下に Cooley et al. (1969) のプログラムに，必要最小限の変更を加えたものを示す．

```
      Subroutine  FFT(A, M, NML)
      Complex  A(1024), U, W, T
      Data  PI/3.14159265/
      SPI = PI
      If( NML.lt.0 )  SPI =-PI
      N = 2**M
      NV2 = N/2
      J = 1
      Do  I=1, N-1
        If( I.lt.J )  then
          T = A(J)
          A(J) = A(I)
          A(I) = T
        Endif
        K = NV2
    6   If( K.ge.J )  Go to  7
          J = J - K
          K = K/2
          Go to  6
        J = J + K
      Enddo
      Do  L=1, M
        LE = 2**L
        LE1 = LE/2
        U = 1
        W = Cmplx(Cos(PI/LE1),
     +            Sin(SPI/LE1))
        Do  J=1, LE1
          Do  I=J, N, LE
            IP = I + LE1
            T = A(IP)*U
            A(IP) = A(I) - T
            A(I) = A(I) + T
          Enddo
          U = U*W
        Enddo
      Enddo
      Return
      End
```

Cooley et al. (1969) による FFT のプログラム

引数 NML は (9.18) 式のときは 1 に，逆変換 (9.19) 式のときには -1 にする．ただし (9.19) 式の係数 $1/N$ は省略している．最後の Do ループが計算の本体である．これに対して最初の Do ループは，データの順番をビットの逆順に並べかえるためのものである．

上の方法とは別に j と k の役割を入れかえて

$$jk = k_0(j_2j_1j_0)_2 + 2k_1(j_1j_0)_2 + 2^2k_0j_0 \mod(N)$$

を用いる方法もある．このときには

$$\begin{aligned}\sum_{j=0}^{N-1} e\Big(\frac{jk}{N}\Big)x(j) = \sum_{j_0} e\Big(\frac{k_2j_0}{2}\Big)\sum_{j_1} e\Big(\frac{k_1j_1}{2}\Big) \\ \times e\Big(\frac{2k_1j_0}{N}\Big)\sum_{j_2} e\Big(\frac{k_0j_2}{2}\Big) \\ \times e\Big(\frac{(j_1j_0)_2k_0}{N}\Big)x(j_2j_1j_0) \qquad (9.43)\end{aligned}$$

となる．こんどは変換が終わった後にビットの反転を行えばよい．この方法を Sande-Tukey のアルゴリズムという．

実数のフーリエ変換　上であげたプログラムはデータが複素数のときの一般的なものである．実用上はデータが実数の場合が圧倒的に多い．もちろん，虚数部を 0 にして複素数用のプログラムを用いてもよいが，それでは無駄な計算をすることになる．

実数と複素数ではデータ量が二倍違うから，実数二つが複素数一つに相当する．そこで単純に実数の偶数番目と奇数番目の一組を複素数として長さ $N/2$ の複素フーリエ変換を計算する．

$$\begin{aligned}b(k) = \sum_{j=0}^{N-1}[x(2j) + ix(2j+1)]\, e\left(\frac{2jk}{N}\right) \\ k = 0, 1, \cdots, N/2 \qquad (9.44)\end{aligned}$$

もともとのフーリエ変換を偶数番目と奇数番目に分けて書くと

$$\begin{aligned}c(k) = \sum_{j=0}^{N/2-1} x(2j)e\left(\frac{2jk}{N}\right) \\ + e\left(\frac{k}{N}\right)\sum_{j=0}^{N/2-1} x(2j+1)e\left(\frac{2jk}{N}\right)\end{aligned}$$

であるから，両式から $x(j)$ を消去すれば

$$c(k) = \frac{1}{2}\Big\{\Big[b(k) + \overline{b(N/2-k)}\Big]$$

$$-ie\left(\frac{k}{N}\right)\left[b(k)-\overline{b(N/2-k)}\right]\Big\} \quad (9.45)$$
$$k=0,1,\cdots,N/2$$

が得られる．$c(k)$ は本来 N 個あるが，残りは

$$c(N-k)=\overline{c(k)}$$

から求められる．

逆変換の場合には，$c(N/2+k)=\overline{c(N/2-k)}$ を用いると

$$x(j)=\frac{1}{N}\sum_{k=0}^{N/2-1}\Big[c(k) +e\left(-\frac{j}{2}\right)\overline{c(N/2-k)}\Big]e\left(-\frac{jk}{N}\right)$$

と書きかえることができる．そこで改めて

$$b(k)=\left[c(k)+\overline{c(N/2-k)}\right] +ie\left(-\frac{k}{N}\right)\left[c(k)-\overline{c(N/2-k)}\right] \quad (9.46)$$
$$k=0,1,\cdots,N/2$$

と定義すれば

$$x(2j)+ix(2j+1)=\frac{1}{N}\sum_{k=0}^{N/2-1}b(k) \times e\left(-\frac{2jk}{N}\right) \quad (9.47)$$

と表すことがでぃる．これも長さ $N/2$ の複素フーリエ変換である．

参考文献

Cooley, J. W. and J. W. Tukey (1965) : An algorithm for the machine computation of complex Fourier series, Math. Comput., **19**, 297-301.

Cooley, J. W., P. A. W. Lewis, and P. D. Welch (1969) : The fast Fourier transform and its applications, IEEE Trans., Education, **12**, 27-34.

10 最小二乗法

ここでは線型の最小二乗法だけを考える．最小二乗法は大きく二つのタイプにわけることができる．最も普通に最小二乗法と呼ばれているのは，観測値から，観測値と線型の関係にある未知のパラメーターを推定するもので，いわゆる逆問題の解法の一つである．もう一方の極端は直接測定の最小二乗法と呼ばれるもので，測定されたパラメーターが，与えられた制約条件を満たすように，観測値を調整するものである．これに対して最初に述べたタイプは間接測定の最小二乗法とでも呼ぶべき方法である．もちろん，両者の中間の問題，すなわち制約条件付きでパラメーターを求める最小二乗法もある．

簡単な例 t–y 平面上に分布している n 組のデータ (t_j, y_j) を t の 1 次式

$$y_j \sim a + bt_j \tag{10.1}$$

で近似する．係数 a, b は誤差の二乗和

$$S = \sum_{j=1}^{n} (y_j - a - bt_j)^2 = \min \tag{10.2}$$

が最小になるように決める．そのためには上式を a, b で偏微分して

$$\frac{\partial S}{\partial a} = -2\sum (y_j - a - bt_j) = 0$$
$$\frac{\partial S}{\partial b} = -2\sum (y_j - a - bt_j)t_j = 0$$

でなければならない．よって

$$a\sum_j 1 + b\sum_j t_j = \sum_j y_j$$
$$a\sum_j t_j + b\sum_j t_j^2 = \sum_j t_j y_j \tag{10.3}$$

が得られた．これを**正規方程式**という．この連立方程式を解けば a, b が決まる．

後で議論するように，実は正規方程式を解く方法は非常に精度が悪く，最近はあまり用いられない．上の簡単な例題でも桁落ちを避けるためには次のような計算法をとるのがよい．まず t_j と y_j の平均値

$$\overline{t} = \frac{1}{n}\sum_{j=1}^{n} t_j \qquad \overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

を計算する．これらを用いると b と a が

$$b = \frac{\sum_j (t_j - \overline{t})(y_j - \overline{y})}{\sum_j (t_j - \overline{t})^2} \qquad a = \overline{y} - b\overline{t} \tag{10.4}$$

として求められる．b は t_j と y_j の間の相関係数のような形をしているが，相関係数は

$$r = \frac{\sum_j (t_j - \overline{t})(y_j - \overline{y})}{\sqrt{\sum_j (t_j - \overline{t})^2 \sum_j (y_j - \overline{y})^2}}$$

で定義されるから，両者は違うものである．相関係数の場合には t_j と y_j とは対等に扱われているが，ここでの最小二乗法では t_j は与えられたものとして誤差は含まず，y_j への当てはめの誤差だけを小さくしている．

線型最小二乗法の一般形 上の例は (10.1) 式からわかるように，求めようとするパラメーター a，b に関して線型である．このような線型の問題を次のように一般化する．まず，n 個の観測値 y_j を要素とする列ベクトルを

$$\boldsymbol{y} = [y_1\ y_2\ \cdots\ y_n]^T$$

とする．^{T} は転置を意味する．また m 個の未知パラメーター x_j からなる列ベクトルを

$$\boldsymbol{x} = [x_1\ x_2\ \cdots\ x_m]^T$$

とする．

問題が線型であるとは，y_i が x_j の線型結合で近似されるという意味であるから，y_i は

$$y_i \sim a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{im}x_m$$

の形に表されるはずである．a_{ij} は与えられた係数である．そこで $n \times m$ の行列

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

を定義すると

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{r} \tag{10.5}$$

が成り立つ．$\boldsymbol{r}$ は当てはめの残差である．この式を観測方程式という．なお，先の例では未知ベクトル

は $\boldsymbol{x} = [a, b]^T$ であり，係数行列は

$$\boldsymbol{A} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix} \tag{10.6}$$

である．

残差の二乗和は $\boldsymbol{r}$ の L_2 ノルムを用いて

$$S = \|\boldsymbol{r}\|^2 = (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) \tag{10.7}$$

と表される．未知パラメーター $\boldsymbol{x}$ に関して変分をとると

$$\delta S = -\delta\boldsymbol{x}^T\boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) - (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})^T\boldsymbol{A}\delta\boldsymbol{x}$$

となる．S が極小値をとるためには δS が $\delta\boldsymbol{x}$ の二次以上の微少量でなければならないから，$\delta\boldsymbol{x}$ の係数を 0 と置いて

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{y} \tag{10.8}$$

が得られた．これが正規方程式である．$\boldsymbol{A}^T\boldsymbol{A}$ は $m \times m$ の正方行列，$\boldsymbol{A}^T\boldsymbol{y}$ は m 元のベクトルである．上式は m 元の連立一次方程式であるから，形式的な解は

$$\boldsymbol{x} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y} \tag{10.9}$$

で与えられる．

古典的な教科書ではこれで最小二乗問題が解けたことになっていたが，最近では正規方程式を解く方法は推奨されない．それは正規方程式の係数行列 $\boldsymbol{A}^T\boldsymbol{A}$ のたちがよくないからである．たとえば $\boldsymbol{A}$ が

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{bmatrix} \tag{10.10}$$

のとき行列

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix} 1+\varepsilon^2 & 1 & 1 \\ 1 & 1+\varepsilon^2 & 1 \\ 1 & 1 & 1+\varepsilon^2 \end{bmatrix}$$

の逆行列は

$$(\boldsymbol{A}^T\boldsymbol{A})^{-1} = \frac{1}{\varepsilon^2(3+\varepsilon^2)} \times \begin{bmatrix} 2+\varepsilon^2 & -1 & -1 \\ -1 & 2+\varepsilon^2 & -1 \\ -1 & -1 & 2+\varepsilon^2 \end{bmatrix}$$

である．ε が小さいときにはこの要素は非常に大きな値になるだけでなく，桁落ちの危険性も大きい．意地悪な例をあげる．

$$(\boldsymbol{A}^T\boldsymbol{A})^{-1}[1, 1, 1]^T = \frac{1}{3+\varepsilon^2}[1, 1, 1]^T$$

逆行列の要素が ε^{-2} のオーダーであるにもかかわらず結果が 1 のオーダーになっているのは桁落ちが起きていることを意味している．上のように解析的に計算すればこのことがわかるが，正規方程式 (10.8) を数値的に解いた場合には見過ごされてしまう．

数値例をあげる．$\boldsymbol{A}$ として 5×4 の行列

$$\boldsymbol{A} = \begin{bmatrix} 1.26 & 0.84 & 0.63 & 0.504 \\ 0.84 & 0.63 & 0.504 & 0.42 \\ 0.63 & 0.504 & 0.42 & 0.36 \\ 0.504 & 0.42 & 0.36 & 0.315 \\ 0.42 & 0.36 & 0.315 & 0.28 \end{bmatrix} \tag{10.11}$$

を用いる．観測値 $\boldsymbol{y}$ としては，$\boldsymbol{x}$ として

$$\boldsymbol{x} = [1/2, 1/3, 1/4, 1/5]^T$$

のときの計算値 $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ を用いる．これらの値を用いて $\boldsymbol{A}^T\boldsymbol{A}$，$\boldsymbol{A}^T\boldsymbol{y}$ を計算し，正規方程式 (10.8) を単精度で解いた結果は

$$\boldsymbol{x} = \begin{bmatrix} 0.41049 \\ 0.95418 \\ -0.92842 \\ 0.86204 \end{bmatrix}$$

となり，正しい値とは似ても似つかない解が得られた．これは係数行列 $\boldsymbol{A}^T\boldsymbol{A}$ の条件数が 10^6 であるから当然である．上の数値もある処理系で得られた結果で，別の処理系で計算すればまったく違う結果が得られるかもしれない．

安定化最小二乗法（damped least squares） 係数行列の条件数が大きいときには，上の例のように解が暴れてしまう．そこで残差を小さくすると同時に解 $\boldsymbol{x}$ のノルムもできるだけ小さくする．すなわち $S = \min$ ではなく

$$\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \lambda\|\boldsymbol{x}\|^2 = \min$$

から $\boldsymbol{x}$ を決める．λ は解析者が与えるパラメーターである．一般に $\boldsymbol{r}$ と $\boldsymbol{x}$ は次元の異なる量であるから，λ の大きさは問題によって大きく異なり，一般的な処方箋はない．この問題に対する観測方程式は

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} \\ \sqrt{\lambda}\boldsymbol{I}_m \end{bmatrix} \boldsymbol{x} + \boldsymbol{r} \tag{10.12}$$

である．残差 $\boldsymbol{r}$ はこんどは $n+m$ 元になる．正規方程式は

$$(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I}_m)\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{y}$$

である．$\boldsymbol{A}^T\boldsymbol{A}$ が特異に近くても，対角成分に定数 λ を加えることによって安定化をはかっている．

この方法によって正規方程式を消去法などで安定に解くことができる．しかし上の方法では大きな対角項にも小さな対角項にも同じ値 λ を加えているのであまり好ましくない．むしろ対角項を定数倍，すなわちすべての対角項を $1+\lambda$ 倍

$$(\boldsymbol{A}^T\boldsymbol{A})_{ii} \longrightarrow (1+\lambda)(\boldsymbol{A}^T\boldsymbol{A})_{ii}$$

にする方がよい．例題 (10.11) 式に対して，$\lambda = 0.001$ として正規方程式を解くと

$$\boldsymbol{x} = \begin{bmatrix} 0.46900 \\ 0.37860 \\ 0.27395 \\ 0.16619 \end{bmatrix}$$

が得られた．これは正解には程遠いが,「正しい」正規方程式を解いたものよりは正解に近い．このように，正規方程式に λ を加えることにより，手軽に解の見当をつけることができる．

最小二乗法の幾何学的意味　正規方程式 (10.8) は

$$\boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}^T\boldsymbol{r} = 0 \tag{10.13}$$

と書き表される．これは S が極小のときの残差ベクトル $\boldsymbol{r}$ が行列 $\boldsymbol{A}$ のすべての列に直交していることを意味している．$\boldsymbol{A}$ の列を $\boldsymbol{a}_j$ と書くことにすれば

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{a}_1x_1 + \boldsymbol{a}_2x_2 + \cdots + \boldsymbol{a}_mx_m$$

であるから，最小二乗法の問題はベクトル $\boldsymbol{y}$ をベクトル $\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_m$ で展開して，残差ベクトル $\boldsymbol{r}$ の長さを最小にするという問題であることがわかる．そのためには $\boldsymbol{r}$ の中に $\boldsymbol{a}_j$ の成分が少しでも含まれてはならない．すなわち

$$\boldsymbol{a}_j^T\boldsymbol{r} = 0 \quad j = 1, 2, \cdots, m$$

でなければならない．これが正規方程式 (10.13) の幾何学的意味である．

$\boldsymbol{a}_j$ が互いに直交しているときには話は簡単である．最初に $\boldsymbol{a}_1$ と $\boldsymbol{y}$ の内積から x_1 を求め，$\boldsymbol{y}$ から $\boldsymbol{a}_1$ 成分 $\boldsymbol{a}_1x_1$ を差し引く．つぎに $\boldsymbol{a}_2$ と，$\boldsymbol{a}_1$ 成分が差し引かれた $\boldsymbol{y}$ から x_2 を求め $\boldsymbol{a}_2$ 成分を差し引く，という手続きを繰り返せばよい．

グラム・シュミット法　$\boldsymbol{a}_j$ が互いに直交していない場合でも最小二乗解を簡単に求めることができる．$\boldsymbol{a}_j$ は §6 で述べた (修正) グラム・シュミット法によって正規直交化することができる．すなわち $\boldsymbol{A}$ は QR 分解できて

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$$

と書くことができる．$\boldsymbol{Q}$ は $n\times m$ 行列，$\boldsymbol{R}$ は $m\times m$ の上三角行列である．$\boldsymbol{Q}$ の列は互いに正規直交しており

$$\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_m$$

が成り立つ ($n=m$ のときには $\boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I}_n$ も成り立つ)．$\boldsymbol{A}$ の列は $\boldsymbol{Q}$ の列 $\boldsymbol{q}_j$ の線型結合で表されるから，$\boldsymbol{r}$ がすべての $\boldsymbol{a}_j$ に直交するということと，$\boldsymbol{r}$ がすべての $\boldsymbol{q}_j$ に直交するということは同等である．したがって，最小二乗の条件は

$$\boldsymbol{Q}^T\boldsymbol{r} = \boldsymbol{Q}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) = 0 \tag{10.14}$$

と書くことができる．$\boldsymbol{A}$ の QR 分解を上式に代入すれば

$$\boldsymbol{R}\boldsymbol{x} = \boldsymbol{Q}^T\boldsymbol{y} \tag{10.15}$$

が得られる．右辺は $\boldsymbol{y}$ を $\boldsymbol{q}_j$ で展開したときの展開係数にほかならない．これは $\boldsymbol{y}$ を $\boldsymbol{A}$ の後に付け加えて，$m+1$ 列の行列としてグラム・シュミット法を行えば，$\boldsymbol{R}$ の最後の列として求められる．上の三角方程式は後退代入法によって簡単に解くことができる．これが正規方程式を経ないで最小二乗解を求める一つの方法である．

例題 (10.11) 式の場合，QR 分解は次のようになる．

$$\begin{aligned}
\boldsymbol{q}_1 &= \frac{1}{\sqrt{1+\varepsilon^2}}[1,\,\varepsilon,\,0,\,0]^T \\
\boldsymbol{q}_2 &= \frac{1}{\sqrt{(1+\varepsilon^2)(2+\varepsilon^2)}}[\varepsilon,\,-1,\,1+\varepsilon^2,\,0]^T \\
\boldsymbol{q}_3 &= \frac{1}{\sqrt{(2+\varepsilon^2)(3+\varepsilon^2)}}[\varepsilon,\,-1,\,-1,\,2+\varepsilon^2]^T \\
r_{11} &= \sqrt{1+\varepsilon^2} \quad r_{12} = r_{13} = \frac{1}{\sqrt{1+\varepsilon^2}} \\
r_{22} &= \varepsilon\sqrt{\frac{2+\varepsilon^2}{1+\varepsilon^2}} \\
r_{23} &= \frac{\varepsilon}{\sqrt{(1+\varepsilon^2)(2+\varepsilon^2)}} \\
r_{33} &= \varepsilon\sqrt{\frac{3+\varepsilon^2}{2+\varepsilon^2}}
\end{aligned}$$

ε が小さいとき，r_{33} と r_{22} が ε のオーダーの量であるから，x_3 と x_2 を求めるときには ε による割り算が出てくる．しかし ε^2 による割り算が現れる正規方程式を用いた方法に比べればたちがよい．

最初の例題 (10.1) 式に対応する係数行列は (10.6) 式である．これを QR 分解すると

$$\begin{aligned}
\boldsymbol{q}_1 &= \frac{1}{\sqrt{n}}[1\ 1\ \cdots\ 1]^T \\
\boldsymbol{q}_2 &= \frac{1}{\sqrt{\sum(t_j-\bar{t})^2}}[t_1-\bar{t}\ t_2-\bar{t}\ \cdots\ t_n-\bar{t}]^T \\
r_{11} &= \sqrt{n} \quad r_{12} = \sqrt{n}\,\bar{t} \\
r_{22} &= \sqrt{\sum(t_j-\bar{t})^2} \\
\boldsymbol{q}_1^T\boldsymbol{y} &= \sqrt{n}\,\overline{\boldsymbol{y}} \\
\boldsymbol{q}_2^T\left[\boldsymbol{y}-\boldsymbol{q}_1(\boldsymbol{q}_1^T\boldsymbol{y})\right] &= \frac{\sum(t_j-\bar{t})(y_j-\overline{y})}{\sum(t_j-\bar{t})^2}
\end{aligned}$$

が得られる．この分解を用いて解を求めると (10.4) 式が得られる．

$\boldsymbol{A}$ の $m+1$ 列目に $\boldsymbol{y}$ を入れておいて修正グラム・シュミット法で QR 分解を行うときに，$m+1$ 列目からは新しい $\boldsymbol{q}_j$ が求められるたびにその成分が差し引かれていくから，最後に残った $\boldsymbol{a}_{m+1}^{(m)}$ は残差 $\boldsymbol{r}$ にほかならない．しかし計算を最後までやってしまうと $\boldsymbol{a}_{m+1}^{(m)}$ は正規化されてしまう．正規化に用いられたノルムの二乗，$r_{m+1,m+1}^2$ が残差の二乗和 S にほかならない．

例題 (10.11) の場合，グラム・シュミット法を単精度で用いても精度一杯の解が得られる．

ハウスホールダー法　行列を QR 分解するもう一つの方法にハウスホールダー変換を用いる方法がある (§6)．$n\times m$ の行列 $\boldsymbol{A}$ に，列に 0 を導入するハウスホールダー変換 (6.9) 式を左から掛けると，$\boldsymbol{A}$ は上三角行列に変換される．

$$\boldsymbol{P}_m\boldsymbol{P}_{m-1}\cdots\boldsymbol{P}_1\boldsymbol{A} = \boldsymbol{Q}^T\boldsymbol{A} = \begin{bmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{bmatrix}$$

$\boldsymbol{R}$ は $m\times m$ の上三角行列，$\boldsymbol{0}$ は $(n-m)\times m$ の零行列である．$\boldsymbol{Q}$ はグラム・シュミット法と違って $n\times n$ の直交行列である．すなわち

$$\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I}_n$$

が成り立っている．しかしハウスホールダー法では $\boldsymbol{Q}$ を計算しておかなくても解は求められる．

同様に観測ベクトル $\boldsymbol{y}$ にハウスホールダー変換を行ったものを

$$\boldsymbol{z} = \begin{bmatrix}\boldsymbol{z}^{(1)}\\ \boldsymbol{z}^{(2)}\end{bmatrix} = \boldsymbol{Q}^T\boldsymbol{y}$$

とする．$\boldsymbol{z}^{(1)}$ は m 元のベクトル，$\boldsymbol{z}^{(2)}$ は $n-m$ 元のベクトルである．そうすると観測方程式は

$$\begin{bmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{bmatrix}\boldsymbol{x} - \begin{bmatrix}\boldsymbol{z}^{(1)}\\ \boldsymbol{z}^{(2)}\end{bmatrix} = \boldsymbol{Q}^T\boldsymbol{r} \tag{10.16}$$

となる．直交変換 $\boldsymbol{Q}$ はベクトルのノルムを保存するから，$\boldsymbol{r}$ のノルムを最小にすることと $\boldsymbol{Q}^T\boldsymbol{r}$ のノルムを最小にすることは同じである．ところが上式の $\boldsymbol{x}$ を変化させても左辺の下の $n-m$ 個の成分は変化しない ($\boldsymbol{z}^{(2)}$ のままである)．したがって $\boldsymbol{r}$ を最小にするには上の m 本の式から

$$\boldsymbol{R}\boldsymbol{x} = \boldsymbol{z}^{(1)} \tag{10.17}$$

でなければならない．この式は後退代入法によってとくことができる．なお，残差の二乗和は

$$\|\boldsymbol{r}\|^2 = \|\boldsymbol{z}^{(2)}\|^2$$

で与えられる．

ハウスホールダー法は $\boldsymbol{Q}$ を計算する必要がなく，また $\boldsymbol{R}$ と $\boldsymbol{z}$ は $\boldsymbol{A}$ と $\boldsymbol{y}$ に上書きできるので，メモリーも少なく計算も速い．精度の面ではグラム・シュミット法と甲乙はつけがたい．

重みつき最小二乗法　これまで観測方程式の残差 $\boldsymbol{r}$ の各成分は同等なものとして取り扱ってきた．観測値の中には誤差の大きなものもあるし，小さなものもある．これらを同等に扱うのは合理的ではない．さらに重要なことは，すべての観測値が同じ種類のものではないこともあるということである．たとえば，$\boldsymbol{y}$ のある成分は地震波の走時，別の成分は表面波の位相速度，重力異常などであるかもしれない．このように次元の異なる量の残差の二乗和を作っても意味がない．

観測値 y_i の誤差の分散を σ_i^2 とする．これは残差の二乗 r_i^2 と同じ次元をもっている．そこで単なる残差の二乗和 (10.7) 式ではなく，σ_i^{-2} の重みを付けた

$$S_w = \sum_{i=1}^{n} \frac{r_i^2}{\sigma_i^2} \tag{10.18}$$

を最小にする．右辺の各項は無次元量であるから加え合わせても意味がある．また，観測誤差の小さな残差には大きなウェイトが掛けられているので，その意味でも合理的である．

この重みつき最小二乗法を通常の最小二乗法と同じ形にするために対角行列

$$\boldsymbol{C}_y^{-1/2} = \begin{bmatrix} \sigma_1^{-1} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-1} & \ddots & \vdots \\ \vdots & \cdots & \ddots & 0 \\ 0 & \cdots & \cdots & \sigma_n^{-1} \end{bmatrix} \tag{10.19}$$

を定義する．これを用いると (10.18) 式の S_w は

$$S_w = (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})^T \boldsymbol{C}_y^{-1} (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) \tag{10.20}$$

と書くことができる．そこで

$$\boldsymbol{y}' = \boldsymbol{C}_y^{-1/2}\boldsymbol{y} \quad \boldsymbol{A}' = \boldsymbol{C}_y^{-1/2}\boldsymbol{A} \tag{10.21}$$

と定義すれば，重みつきの残差 S_w は

$$S_w = (\boldsymbol{y}' - \boldsymbol{A}'\boldsymbol{x})^T (\boldsymbol{y}' - \boldsymbol{A}'\boldsymbol{x}) \tag{10.22}$$

となる．これは重みなしの単純な最小二乗法の二乗和 (10.7) 式と全く同じ形である．したがってこれまでの $\boldsymbol{A}$ を $\boldsymbol{A}'$ で，$\boldsymbol{y}$ を $\boldsymbol{y}'$ で置きかえればまったく同じ議論が成り立つ．

推定の誤差　観測値 $\boldsymbol{y}$ には誤差が含まれているから，それを用いて得られた最小二乗解にも誤差が含まれている．

y が確率変数のときその期待値を

$$\overline{y} = E[y]$$

と書くことにする．y の分散は

$$\sigma_y^2 = E[(y - \overline{y})^2]$$

で定義される．$\boldsymbol{y}$ が n 元のベクトルのとき，その期待値は，各成分の期待値からなるベクトル $\overline{\boldsymbol{y}}$ である．$\boldsymbol{y}$ の分散は

$$\boldsymbol{C}_y = E[(\boldsymbol{y} - \overline{\boldsymbol{y}})(\boldsymbol{y} - \overline{\boldsymbol{y}})^T] \tag{10.23}$$

で定義される．ノルムと違って後ろの項が転置になっている．したがって $\boldsymbol{C}_y$ は $n \times n$ の正方行列であり，各成分は

$$(\boldsymbol{C}_y)_{ij} = E[(y_i - \overline{y}_)(y_j - \overline{y}_j)]$$

で定義される．これは y_i と y_j の間の共分散であるから，$\boldsymbol{C}_y$ は共分散行列である．もし y_i と y_j が無相関なら

$$(\boldsymbol{C}_y)_{ij} = 0 \quad i \neq j$$

であるから，$\boldsymbol{C}_y$ は対角行列になる．

つぎに最小二乗法で求められた推定値 $\boldsymbol{x}$ の分散を計算する．はじめに重みがない場合の (10.7) 式を考える．ここでは確率変数 $\boldsymbol{y}$ から求められる推定値であることを強調して (10.9) 式を

$$\begin{aligned} \hat{\boldsymbol{x}} &= \boldsymbol{B}\boldsymbol{y} \\ \boldsymbol{B} &= (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T \end{aligned} \tag{10.24}$$

と書くことにする．$\hat{\boldsymbol{x}}$ の期待値は

$$E[\hat{\boldsymbol{x}}] = E[\boldsymbol{B}\boldsymbol{y}] = \boldsymbol{B}\overline{\boldsymbol{y}}$$

であるから，$\hat{\boldsymbol{x}}$ の分散は

$$\begin{aligned} \boldsymbol{C}_{\hat{x}} &= E[(\hat{\boldsymbol{x}} - \overline{\boldsymbol{x}})(\hat{\boldsymbol{x}} - \overline{\boldsymbol{x}})^T] \\ &= E[\boldsymbol{B}(\boldsymbol{y} - \overline{\boldsymbol{y}})(\boldsymbol{y} - \overline{\boldsymbol{y}})^T\boldsymbol{B}^T] \\ &= \boldsymbol{B}E[(\boldsymbol{y} - \overline{\boldsymbol{y}})(\boldsymbol{y} - \overline{\boldsymbol{y}})^T]\boldsymbol{B}^T \\ &= \boldsymbol{B}\boldsymbol{C}_y\boldsymbol{B}^T \end{aligned} \tag{10.25}$$

と表される．この関係を**誤差の伝播公式**という．もし y_i が同種の観測量で互いに無相関で，しかも分散がすべて等しく σ_0^2 であるなら

$$\boldsymbol{C}_y = \sigma_0^2 \boldsymbol{I}_n$$

であるから，推定値 $\hat{\boldsymbol{x}}$ の共分散行列は

$$\boldsymbol{C}_{\hat{x}} = \sigma_0^2 \boldsymbol{B}\boldsymbol{B}^T = \sigma_0^2 (\boldsymbol{A}^T\boldsymbol{A})^{-1} \tag{10.26}$$

と簡単になる．観測値が無相関でもパラメーターの推定値相互間には相関がある．

ここに現れた逆行列 $(\boldsymbol{A}^T\boldsymbol{A})^{-1}$ は，$\boldsymbol{A}$ が QR 分解できていれば簡単に計算することができる．

$$\begin{aligned}(\boldsymbol{A}^T\boldsymbol{A})^{-1} &= (\boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{R})^{-1} \\ &= (\boldsymbol{R}^T\boldsymbol{R})^{-1} = \boldsymbol{R}^{-1}\boldsymbol{R}^{-T}\end{aligned}$$

であるから，$\boldsymbol{R}^{-1}$ さえ計算すればよい．$\boldsymbol{R}^{-T}$ は $\boldsymbol{R}^{-1}$ の転置行列である．$\boldsymbol{R}$ はグラム・シュミット法でもハウスホールダー法で計算してもよい．

$\boldsymbol{R}$ は上三角行列であるから逆行列 $\boldsymbol{R}^{-1}$ の要素は後退代入法によって簡単に求めることができる．$\boldsymbol{R}$ の要素を R_{ij}，$\boldsymbol{R}^{-1}$ の要素を R_{ij}^{-1} と書くことにすると

$$\sum_{k=1}^{j} R_{ik}R_{kj}^{-1} = \delta_{ij} \quad i \le j$$

が成り立つ．R_{ij}^{-1} は下の行から計算する．すなわち $i = m, m-1, \cdots, 1$ に対して

$$\begin{aligned}R_{ii}^{-1} &= \frac{1}{R_{ii}} \\ R_{ij}^{-1} &= -R_{ii}^{-1}\sum_{k=i+1}^{j} R_{ik}R_{kj}^{-1} \\ &\qquad j = i+1, i+2, \cdots, m\end{aligned} \tag{10.27}$$

とする．$i = m$ のときには二番目の式は不要である．この計算法では R_{ij}^{-1} を R_{ij} に上書きすることができる．

重みつき最小二乗法 (10.22) の場合は，推定値は

$$\begin{aligned}\hat{\boldsymbol{x}} &= \boldsymbol{B}'\boldsymbol{y} \\ \boldsymbol{B}' &= (\boldsymbol{A}'^T\boldsymbol{A}')^{-1}\boldsymbol{A}'^T \\ &= (\boldsymbol{A}^T\boldsymbol{C}_y^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{C}_y^{-1}\end{aligned} \tag{10.28}$$

と書くことができる．$\boldsymbol{A}'$ は (10.21) 式である．これより

$$\boldsymbol{C}_{\hat{x}} = \boldsymbol{B}'\boldsymbol{C}_y\boldsymbol{B}'^T = (\boldsymbol{A}'^T\boldsymbol{A}')^{-1} \tag{10.29}$$

と簡単に表される．重み付き最小二乗法を解くの場合，$\boldsymbol{A}'$ を QR 分解することになるから，改めて

$$\boldsymbol{A}' = \boldsymbol{Q}'\boldsymbol{R}'$$

とすれば

$$\boldsymbol{C}_{\hat{x}} = \boldsymbol{R}'^{-1}\boldsymbol{R}'^{-T} \tag{10.30}$$

が得られる．

計算値の分散　$\hat{\boldsymbol{x}}$ が求められると観測値の計算値が

$$\hat{\boldsymbol{y}} = \boldsymbol{A}\hat{\boldsymbol{x}} \tag{10.31}$$

から求められる．この計算値の共分散行列 $\boldsymbol{C}_{\hat{y}}$ は，誤差の伝播法則から

$$\boldsymbol{C}_{\hat{y}} = \boldsymbol{A}\boldsymbol{C}_{\hat{x}}\boldsymbol{A}^T$$

と表される．$\boldsymbol{C}_{\hat{x}}$ として重みつきの (10.29) 式を用い

$$\boldsymbol{A}' = \boldsymbol{C}_y^{-1/2}\boldsymbol{A} = \boldsymbol{Q}'\boldsymbol{R}'$$

に注意すれば

$$\boldsymbol{C}_{\hat{y}} = \boldsymbol{C}_y^{1/2}\boldsymbol{Q}'\boldsymbol{Q}'^T\boldsymbol{C}_y^{1/2} \tag{10.32}$$

となる．$\boldsymbol{Q}'$ は $n \times m$ の行列であるから $\boldsymbol{Q}'\boldsymbol{Q}'^T$ は単位行列にはならない．$\boldsymbol{C}_y$ が対角行列のときには，対角成分を σ_i^2，$\boldsymbol{Q}'$ の成分を q'_{ij} とすれば，$\boldsymbol{C}_{\hat{y}}$ の対角成分 $(\boldsymbol{C}_{\hat{y}})_{ii}$，すなわち $\hat{y}_i$ の分散は

$$(\boldsymbol{C}_{\hat{y}})_{ii} = \sigma_i^2 \sum_{j=1}^{m} q'^2_{ij}$$

と書くことができる．和は $\boldsymbol{Q}'$ の行のノルムの二乗である．$\boldsymbol{Q}'$ は正規直交行列であるから，列のノルムは 1，行のノルムは 1 よりも小さい ($\boldsymbol{Q}'$ が $n \times n$ のときは行のノルムも 1 になる)．よって

$$(\boldsymbol{C}_{\hat{y}})_{ij} \le \sigma_i^2 = (\boldsymbol{C}_y)_{ii} \tag{10.33}$$

すなわち，計算値の分散は観測値の分散より小さい (大きくはない)．

$\hat{\boldsymbol{x}}$ の共分散は $\boldsymbol{R}$ あるいは $\boldsymbol{R}'$ だけから計算できるから，ハウスホールダー法で十分であるが，計算値 $\hat{\boldsymbol{y}}$ の共分散は $\boldsymbol{Q}$ あるいは $\boldsymbol{Q}'$ が必要であるから，グラム・シュミット法の方が便利である．

相関がある観測値 これまで観測値間には相関がないとしてきた．相関があるときには共分散行列 $\boldsymbol{C}_y$ は対角行列ではなくなる．このときでも，(10.20) 式によって重みつきの残差の二乗和を定義して，これを最小にするような解を求めることができる．問題は (10.21) 式などに現れる $\boldsymbol{C}_y^{1/2}$ などの量である．

$\boldsymbol{C}_y$ は対称行列であるから，たとえば下三角行列 $\boldsymbol{L}$ の積

$$\boldsymbol{C}_y = \boldsymbol{L}\boldsymbol{L}^T \tag{10.34}$$

$$\boldsymbol{L} = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ l_{n1} & \cdots & \cdots & l_{nn} \end{bmatrix}$$

に分解することができる．$\boldsymbol{L}$ はクラウトの方法 (§4) を対称行列に適用したコレスキーの方法で簡単に解くことができる．$\boldsymbol{C}_y$ の要素を改めて σ_{ij} とすると，この方法は $k = 1, 2, \cdots, n$ について

$$\begin{aligned} l_{kk} &= \sqrt{\sigma_{kk}^2 - \sum_{j=1}^{k-1} l_{kj}^2} \\ l_{ik} &= \frac{1}{l_{kk}}\Big(\sigma_{ik}^2 - \sum_{j=1}^{k-1} l_{ij} l_{kj}\Big) \\ & \qquad i = k+1, k+2, \cdots, n \end{aligned} \tag{10.35}$$

とすればよい．$\boldsymbol{C}_y$ は正定値行列であるから，第一行目の根号の中が負になることはない．

こうして分解ができると

$$\boldsymbol{C}_y^{-1} = (\boldsymbol{L}\boldsymbol{L}^T)^{-1} = \boldsymbol{L}^{-T}\boldsymbol{L}^{-1}$$

であるから，(10.20) 式は

$$S_w = \|\boldsymbol{L}^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})\|^2$$

となる．したがって (10.21) 式の変換は

$$\boldsymbol{y}' = \boldsymbol{L}^{-1}\boldsymbol{y} \quad \boldsymbol{A}' = \boldsymbol{L}^{-1}\boldsymbol{A}$$

とすればよいことがわかる．要するに，これまでの $\boldsymbol{C}_y^{-1/2}$ を $\boldsymbol{L}^{-1}$ で置きかえればよい．ただし，これまでは $\boldsymbol{C}_y^{1/2}$ を対称行列と仮定したが，もちろん $\boldsymbol{L}$ は対称ではないからそのための修正が必要である．

最尤法 最小二乗法は統計学的には最尤法に基づいているが，これまでは議論を単純化するためにこの点についてはまったく触れないできた．

x を平均 μ，分散 σ^2 の正規分布に従う確率変数であるとき，確率分布の密度関数は

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{(x-\mu)^2}{2\sigma^2}\Big] \tag{10.36}$$

で表される．いま，一つの観測値 x_1 が与えられたとき，x_1 が平均 μ に近いほど $p(x_1|\mu, \sigma^2)$ の値は大きくなる．逆に x_1 を固定して未知パラメーター μ の関数と考えたとき

$$L(\mu, \sigma^2|x_1) = p(x_1|\mu, \sigma^2) \tag{10.37}$$

は μ が x_1 に近いほど大きな値をとる．この関数を尤度関数という．

同じ母集団から独立に n 個の観測値 $x_1, x_2, \cdots, x_n$ が得られたとする．このときの尤度関数は

$$\begin{aligned} L(\mu, \sigma^2|x_1, x_2, \cdots, x_n) &= p(x_1|\mu, \sigma^2) \\ &\quad \times p(x_2|\mu, \sigma^2) \cdots p(x_n|\mu, \sigma^2) \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\Big] \end{aligned} \tag{10.38}$$

で定義される．これを μ の関数と見たとき，最大になるところの μ が最もありそうな平均値である．このように尤度が最大になるように未知パラメーターを決める方法を最尤法という．

正規分布の場合，尤度関数そのものよりも対数尤度関数

$$\begin{aligned} l(\mu, \sigma|x_1, x_2, \cdots, x_n) &= \log L(\mu, \sigma^2|x_1, \cdots) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \end{aligned} \tag{10.39}$$

を用いる方が便利である．μ に関して L が最小になるのは

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \min$$

のときで，これは明らかに最小二乗法である．上式を μ で微分して 0 と置けば

$$\sum_{i=1}^{n}(x_i - \mu) = 0$$

したがってよく用いられている標本平均

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{10.40}$$

は最尤法による推定値にほかならない.

分散 σ^2 も最尤法で求めることができる. (10.39) 式を σ^2 で微分して 0 と置くと

$$-\frac{n}{\sigma^2}+\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2=0$$

であるから, σ^2 の推定値として

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 \tag{10.41}$$

が得られた. これも標本分散としてよく用いられている.

x_i は確率的に変化する量であるから, $\hat{\mu}$ や $\hat{\sigma}^2$ も観測値の組ごとに変化する. そこで確率的な平均値を求めると

$$E[\hat{\mu}]=\mu \tag{10.42}$$

が得られる. このように, 期待値が真の値になるような推定量を不偏推定量 (unbiased estimator) という. これに対して, 計算は面倒だが $\hat{\sigma}^2$ の期待値を計算すると

$$E[\hat{\sigma}^2]=\frac{n-1}{n}\sigma^2 \tag{10.43}$$

となって σ^2 に一致しない. すなわち (10.41) 式の $\hat{\sigma}^2$ は不偏推定量ではない. そこで σ^2 の推定量として

$$\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 \tag{10.44}$$

もよく用いられる. これは不偏推定量ではあるが, 最尤推定量ではない.

ついでにいうと, $\hat{\mu}$ の分散は

$$E[(\hat{\mu}-\mu)^2]=\frac{1}{n}\sigma^2$$

であるから, サンプル数 n が無限大の極限では分散は 0, すなわち $\hat{\mu}$ は正確な値に漸近する. このような推定量を一致推定量という.

推定値の相関係数 最小二乗法のに話をもどすと, 未知パラメーター $\boldsymbol{x}$ の推定量 $\hat{\boldsymbol{x}}$ の分散が $\boldsymbol{C}_{\hat{x}}$ であるということは, 正規分布を仮定すると対数尤度関数が

$$l(\boldsymbol{x})=-\frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{x}})^T\boldsymbol{C}_{\hat{x}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{x}})+\text{定数} \tag{10.45}$$

で表されることを意味している. $\hat{\boldsymbol{x}}$ は最小二乗解を, $\boldsymbol{C}_{\hat{x}}$ はその分散を表しており, (10.26) 式, あるいは (10.29) 式で与えられている. 最後の定数は $\boldsymbol{x}$ によらない定数である. この式から $\boldsymbol{x}=\hat{\boldsymbol{x}}$ のときに尤度が最大になるが (そのように $\hat{\boldsymbol{x}}$ を決めた), その周辺で尤度は一様に減少していない.

わかりやすくするために二次元のときを考え

$$\boldsymbol{C}_{\hat{x}}=\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix} \tag{10.46}$$

と置く. σ_1^2, σ_2^2 はそれぞれ $\hat{x}_1, \hat{x}_2$ の分散であり, ρ は $\hat{x}_1$ と $\hat{x}_2$ の間の相関係数である. この共分散行列の逆行列は

$$\boldsymbol{C}_{\hat{x}}^{-1}=\frac{1}{1-\rho^2}\begin{bmatrix}1/\sigma_1^2 & -\rho/\sigma_1\sigma_2\\ -\rho/\sigma_1\sigma_2 & 1/\sigma_2\end{bmatrix}$$

である. いま

$$\delta x_i=x_i-\hat{x}_i$$

と置くと

$$\delta S=(\boldsymbol{x}-\hat{\boldsymbol{x}})^T\boldsymbol{C}_{\hat{x}}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{x}})=\frac{1}{1-\rho^2}\Big[\Big(\frac{\delta x_1}{\sigma_1}\Big)^2 -2\rho\Big(\frac{\delta x_1\delta x_2}{\sigma_1\sigma_2}\Big)+\Big(\frac{\delta x_2}{\sigma_2}\Big)^2\Big]$$

が得られる. これは最小二乗解 $\hat{\boldsymbol{x}}$ の周りの相対的な尤度を表したものである.

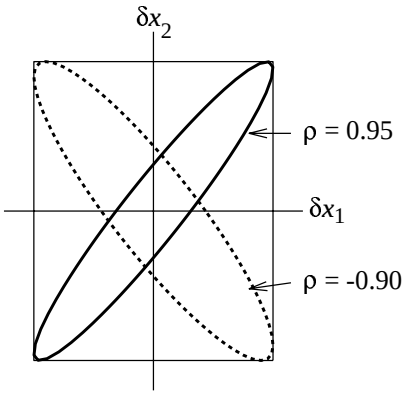


Fig. 10.1 対数尤度関数の等高線

Fig. 10.1 は $\sigma_2/\sigma_1=2.5/2.0$ のときの $\delta S=1$ の等高線を $\rho=0.95$(実線), $\rho=-0.9$(点線) に対して示したものである. 図の中心が尤度が最大の点である. 等高線が細長く延びているということは, 二つ

のパラメーター x_1 と x_2 がまったく独立に決まるのではなく，互いに相関をもって決まっていることを示している．たとえば $\rho = 0.95$ のときには δx_1 を増加させると同時に δx_2 を増加させても尤度はほとんど変わらないから，これも一つの解と考えることができる．したがって，最小二乗解を求めたときには解の分散行列を計算して，パラメーター間の相関係数を計算しておくことが望ましい．これは係数行列の QR 分解 R の逆行列だけから計算できるからたいした手間ではない．

直接観測　直接観測の最小二乗法については基本的なところだけを述べるにとどめる．

三角形の三つの角 x_1, x_2, x_3 を測定したとする．これらの和は π にならなければならない．しかし観測誤差のためにそうはなっていない．そこで角 x_i の誤差を e_i として，条件

$$\sum_{i=1}^{3}(x_i + e_i) = \pi$$

の下で誤差の二乗和が最小

$$\sum_{i=1}^{3}|e_i|^2 = \min$$

という条件から e_i を決めることができる．e_i が求められると $x_i + e_i$ が角度の推定値になる．

この問題を一般化する．m 個のパラメーター x_j に対して n 組の線型の条件

$$b_{i1}x_1 + b_{i2}x_2 + \cdots + b_{im}x_m = f_i \qquad i = 1, 2, \cdots, n$$

課せられているとする．ここで $n < m$，すなわち条件の数はパラメータの数よりも少ないものとする．これを行列で書いて

$$\boldsymbol{B}\boldsymbol{x} = \boldsymbol{f} \tag{10.47}$$

とする．$\boldsymbol{B}$ は $n \times m$ の与えられた行列，$\boldsymbol{f}$ は n 次元の与えられたベクトルである．上式は $\boldsymbol{x}$ に誤差が含まれないときに成り立つ式である．

実際に測定を行って得られた値を改めて $\boldsymbol{x}$ とする．これには誤差が含まれているから上式は成り立たない．そこで誤差を $\boldsymbol{e}$ として $\boldsymbol{x}+\boldsymbol{e}$ が上式を満たすようにする．

$$\boldsymbol{B}(\boldsymbol{x} + \boldsymbol{e}) = \boldsymbol{f} \tag{10.48}$$

$n < m$ であるから観測値 $\boldsymbol{x}$ に対して上式を満足する $\boldsymbol{e}$ は無限に存在するが，そのなかで誤差の二乗和が最小のもの

$$\|\boldsymbol{e}\|^2 = \min \tag{10.49}$$

を解とする．

(10.48) 式の条件の下で $\|\boldsymbol{e}\|^2$ を最小にするというのは条件付極値問題である．これを解く定石はラグランジェの未定係数法である．式を簡略にするために (10.48) 式を

$$\boldsymbol{B}\boldsymbol{e} = \boldsymbol{f} - \boldsymbol{B}\boldsymbol{x} \equiv \boldsymbol{y} \tag{10.50}$$

と書き直しておく．右辺の $\boldsymbol{y}$ は観測値 $\boldsymbol{x}$ から計算できる量である．この表現を用いると，いまの問題ではもとの問題を解くかわりに極値問題

$$\boldsymbol{e}^T\boldsymbol{e} + \boldsymbol{\lambda}^T(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{e}) + (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{e})^T\boldsymbol{\lambda} = \min \tag{10.51}$$

を解けばよいことになる．ここに $\boldsymbol{\lambda}$ は n 次元の未定係数ベクトルで，これも極値を求めるときの変数である．

上式を $\boldsymbol{e}$ について変分をとって 0 と置けば

$$\delta\boldsymbol{e}^T\boldsymbol{e} + \boldsymbol{e}^T\delta\boldsymbol{e} - \boldsymbol{\lambda}^T\boldsymbol{B}\delta\boldsymbol{e} - \delta\boldsymbol{e}^T\boldsymbol{B}^T\boldsymbol{\lambda} = 0$$

が得られる．これが $\delta\boldsymbol{e}$ によらずに成り立つためには

$$\boldsymbol{e} = \boldsymbol{B}^T\boldsymbol{\lambda}$$

でなければならない．一方，(10.51) 式を $\boldsymbol{\lambda}$ について変分をとった式からは当然 (10.50) 式が得られる．いま求めた $\boldsymbol{e}$ が (10.50) 式を満たすためには

$$\boldsymbol{B}\boldsymbol{e} = \boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\lambda} = \boldsymbol{y}$$

でなければならない．この式から $\boldsymbol{\lambda}$ が決まり，解は最終的には

$$\boldsymbol{e} = \boldsymbol{B}^T\left(\boldsymbol{B}\boldsymbol{B}^T\right)^{-1}\boldsymbol{y} \tag{10.52}$$

となる．したがって $\boldsymbol{x}$ の推定値は

$$\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{B}^T\left(\boldsymbol{B}\boldsymbol{B}^T\right)^{-1}(\boldsymbol{f} - \boldsymbol{B}\boldsymbol{x}) \tag{10.53}$$

となる．右辺に現れる $\boldsymbol{x}$ は観測値である．

最初にあげた例では

$$\boldsymbol{B} = [1\ 1\ 1] \quad f = \pi \quad y = \pi - x_1 - x_2 - x_3$$

であるから

$$\boldsymbol{B}\boldsymbol{B}^T = 3$$

したがって解は

$$\hat{x_i} = x_i + \frac{1}{3}y$$

になる．すなわち，誤差を三つの角に等分に割り振ったのが最も確からしい．

(10.52) 式は間接測定の場合でいえば正規方程式を用いて解を導いたものであるから，パラメーターの数が多くなると精度が悪くなる．そこで間接測定の場合のように観測方程式そのものを用いた解がほしくなる．いろいろな方法が考えられるが，たとえば QR 分解を用いることにして

$$\boldsymbol{B}^T = \boldsymbol{Q}\boldsymbol{R} \tag{10.54}$$

と分解する．ここに $\boldsymbol{Q}$ は $m \times n$ の直交行列で $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_n$，$\boldsymbol{R}$ は $n \times n$ の上三角行列である．この分解を用いると (10.52) 式は

$$\boldsymbol{e} = \boldsymbol{Q}\boldsymbol{R}^{-T}\boldsymbol{y} \tag{10.55}$$

となる．

参考文献

中川 徹・小柳義夫 (1982)：最小二乗法による実験データ解析，東京大学出版会．

11 特異値分解と一般逆行列

§6 で $n \times m$ $(n \geq m)$ 行列 $\boldsymbol{A}$ が

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$$

の形に分解できることを示した．そこでは分解のアルゴリズムだけに注目していたが，ここでは一般の場合を考え，$n \geq m$ の条件も外しても分解が可能であること，連立方程式 $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ の解がある意味では常に存在することなどを示す．

対称行列の固有値問題　はじめに対称行列の固有値問題の復習をする．

$\boldsymbol{S}$ を $n \times n$ の実対称行列

$$\boldsymbol{S}^T = \boldsymbol{S} \tag{11.1}$$

とする．$\boldsymbol{S}$ が複素数行列のときにはエルミート行列であれば，すなわち $\boldsymbol{S}^T = \overline{\boldsymbol{S}}$ が成り立てば，以下の議論はほとんど平行して成り立つが，ここでは実行列のみを取り扱う．

対称行列の固有値，固有ベクトルは実数である．固有値 λ_i に属する固有ベクトルを $\boldsymbol{u}_i$ とすると

$$\boldsymbol{S}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i \quad i = 1, 2, \cdots, n \tag{11.2}$$

が成り立つ．異なる固有値に属する固有ベクトルは互いに直交する．すなわち

$$\boldsymbol{u}_j^T \boldsymbol{u}_i = 0 \quad \lambda_i \neq \lambda_j$$

が成り立つ．固有値が重根のときには多重度の数だけの固有ベクトルが存在し，それらを互いに直交化することができる．そこで固有ベクトルを正規化すれば

$$\boldsymbol{u}_j^T \boldsymbol{u}_i = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{11.3}$$

とすることができる．

固有ベクトルを横に並べた $n \times n$ 正方行列を

$$\boldsymbol{U} = [\boldsymbol{u}_1 \ \boldsymbol{u}_2 \ \cdots \ \boldsymbol{u}_n] \tag{11.4}$$

とする．正規直交性 (11.3) 式により，$\boldsymbol{U}$ は直交行列

$$\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_n = \boldsymbol{U}\boldsymbol{U}^T \tag{11.5}$$

である．ただし $\boldsymbol{I}_n$ は $n \times n$ の単位行列である．$\boldsymbol{U}$ を用いると固有値問題 (11.2) 式は

$$\boldsymbol{S}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Lambda} \tag{11.6}$$

と書くことができる．ここに $\boldsymbol{\Lambda}$ は対角成分が固有値の $n \times n$ の対角行列

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix} \tag{11.7}$$

である．直交関係 (11.5) 式を用いると

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T \quad \boldsymbol{U}^T\boldsymbol{S}\boldsymbol{U} = \boldsymbol{\Lambda} \tag{11.8}$$

の関係が導かれる．すなわち $\boldsymbol{S}$ が直交行列と対角行列の積に分解できた．

(11.8) 式は固有値に 0 が含まれていても成り立つ一般的な関係である．もし固有値のうち r 個が 0 でなく，残りの $n-r$ 個が 0 であるとき，$\boldsymbol{S}$ のランク (階数) は r であるという．このとき固有値と固有ベクトルを並べかえて

$$\lambda_i \neq 0 \quad 1 \leq i \leq r$$
$$\lambda_i = 0 \quad r+1 \leq i \leq n$$

のように番号を付けかえる．(11.6) 式の右辺は

$$\boldsymbol{U}\boldsymbol{\Lambda} = [\lambda_1\boldsymbol{u}_1 \ \lambda_2\boldsymbol{u}_2 \ \cdots \ \lambda_r\boldsymbol{u}_r \ \boldsymbol{0} \ \cdots \ \boldsymbol{0}]$$

であるから，後ろの $n-r$ 列は 0 になる．これに $\boldsymbol{U}^T$ を右から掛けるとき，$\boldsymbol{U}^T$ の下の $n-r$ 行には，いいかえれば $\boldsymbol{U}$ の最後の $n-r$ 列には 0 が掛けられる．したがって $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ の計算では $\boldsymbol{U}$ の後ろの $n-r$ 列は必要がないことになり

$$\boldsymbol{U}_r = [\boldsymbol{u}_1 \ \boldsymbol{u}_2 \ \cdots \ \boldsymbol{u}_r]$$

とすれば

$$\boldsymbol{S} = \boldsymbol{U}_r\boldsymbol{\Lambda}_r\boldsymbol{U}_r^T \tag{11.9}$$

と書けることになる．$\boldsymbol{\Lambda}_r$ は 0 でない固有値からなる対角行列，すなわち (11.7) 式で $n = r$ と置いたものである．$\boldsymbol{U}_r$ には後ろの $n-r$ 列が欠けているので

$$\boldsymbol{U}_r^T\boldsymbol{U}_r = \boldsymbol{I}_r$$

は成立するが，$r < n$ のときには

$$\boldsymbol{U}_r\boldsymbol{U}_r^T \neq \boldsymbol{I}_n \quad (r < n)$$

である．

連立方程式の解 対称行列 $\boldsymbol{S}$ を係数行列とする連立方程式

$$\boldsymbol{S}\boldsymbol{x} = \boldsymbol{y} \tag{11.10}$$

を解きたい．そのために $\boldsymbol{x}$ と $\boldsymbol{y}$ を固有ベクトル $\boldsymbol{u}_i$ で展開する．

$$\begin{aligned} \boldsymbol{x} &= \sum_i x'_i \boldsymbol{u}_i \quad x'_i = \boldsymbol{u}_i^T \boldsymbol{x} \\ \boldsymbol{y} &= \sum_i y'_i \boldsymbol{u}_i \quad y'_i = \boldsymbol{u}_i^T \boldsymbol{y} \end{aligned} \tag{11.11}$$

これをもとの方程式 (11.10) に代入すれば

$$\boldsymbol{S}\boldsymbol{x} = \sum_i x'_i \boldsymbol{S}\boldsymbol{u}_i = \sum_i x_i \lambda_i \boldsymbol{u}_i$$

であるから

$$\sum_{i=1}^{n} \lambda_i x'_i \boldsymbol{u}_i = \sum_{i=1}^{n} y'_i \boldsymbol{u}_i$$

となる．$\boldsymbol{u}_i$ は互いに直交しているから，上式は

$$\lambda_i x'_i = y'_i \quad 1 \le i \le n \tag{11.12}$$

を意味している．もしすべての固有値が 0 でないなら

$$x'_i = \frac{y'_i}{\lambda_i}$$

によって x'_i が求められるから，これを (11.11) 式に代入すれば解 $\boldsymbol{x}$ が得られることになる．

$$\boldsymbol{x} = \sum_{i=1}^{n} \frac{\boldsymbol{u}_i^T \boldsymbol{y}}{\lambda_i} \boldsymbol{u}_i = \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T\boldsymbol{y} \quad (r = n) \tag{11.13}$$

最後の式は分解 (11.8) 式を (11.10) 式に代入しても導くことができる．

固有値に 0 が含まれるときには，先と同じように番号付けをすれば (11.12) 式の $i > r$ の部分は

$$0 \cdot x'_i = y'_i \quad r < i \le n$$

を意味する．したがって $y'_i = 0$ でなければ方程式は不能になる．すなわち，方程式 (11.10) が解けるためには

$$y'_i = 0 \quad i > r$$

でなければならない．この条件を**適合条件** (compatibility condition) という．この条件を書きかえれば

$$\boldsymbol{u}_i^T \boldsymbol{y} = 0 \quad i > r \tag{11.14}$$

である．$\boldsymbol{S}$ が正則でないときには (11.10) 式の右辺 $\boldsymbol{y}$ を勝手に与えても解は存在せず，上の適合条件を満たす右辺に対してだけ解が存在する．いいかえれば，$\boldsymbol{y}$ は $\boldsymbol{u}_1 \sim \boldsymbol{u}_r$ の張る空間内になければならない．ただしこの条件が満たされていたとしても解は一義的ではなく，ある解に 0 固有値の固有ベクトルを加えたもの，すなわち

$$\boldsymbol{x} = \sum_{i=1}^{r} \frac{\boldsymbol{u}_i^T \boldsymbol{y}}{\lambda_i} \boldsymbol{u}_i + \sum_{i=r+1}^{n} x'_i \boldsymbol{u}_i \tag{11.15}$$

も解になっている．ここでの x'_i は任意の定数である．なぜなら，上式に $\boldsymbol{S}$ を左から掛けると $i > r$ の $\boldsymbol{u}_i$ の項は消えてしまうからである．

長方形行列の分解 任意の $n \times m$ 行列 $\boldsymbol{A}$ から対称行列

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{A} \\ \boldsymbol{A}^T & \boldsymbol{0} \end{bmatrix} \tag{11.16}$$

を定義する．$\boldsymbol{S}$ は $(n+m) \times (n+m)$ 対称行列であるから，固有値問題

$$\boldsymbol{S}\boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i \tag{11.17}$$

には重根を含めて $n+m$ 組の解が存在する．固有ベクトル $\boldsymbol{w}_i$ は $n+m$ 次元であるから，はじめの n 成分を $\boldsymbol{u}_i$，残りの m 成分を $\boldsymbol{v}_i$ とする．すなわち

$$\boldsymbol{w}_i = \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{v}_i \end{bmatrix}$$

とすると，固有値問題 (10.17) 式は

$$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{A} \\ \boldsymbol{A}^T & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{v}_i \end{bmatrix} = \lambda_i \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{v}_i \end{bmatrix}$$

すなわち

$$\boldsymbol{A}\boldsymbol{v}_i = \lambda_i \boldsymbol{u}_i \quad \boldsymbol{A}^T \boldsymbol{u}_i = \lambda_i \boldsymbol{v}_i \tag{11.18}$$

が得られる．これは連立した固有値問題とでも呼ぶべき方程式である．しかし $\boldsymbol{u}_i$，$\boldsymbol{v}_i$ はそれぞれ消去することができて

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{v}_i = \lambda_i^2 \boldsymbol{v}_i \quad \boldsymbol{A}\boldsymbol{A}^T\boldsymbol{u}_i = \lambda_i^2 \boldsymbol{u}_i \tag{11.19}$$

が得られる．$\boldsymbol{A}^T\boldsymbol{A}$ は $m \times m$ 正方行列であるから第一の固有値問題には m 組の解がある．同様に第二の固有値問題には n 組の解がある．

しかしこれでは数が合わない．固有値問題 (11.17) では独立な $\boldsymbol{w}_i$ は $n+m$ 個あるはずであるから，(11.19) 式の解から $\boldsymbol{w}_i$ を作るには数が足りない．これは $\boldsymbol{S}$ が特別な形をしているからである．

固有値問題 (11.18) の固有系を $(\lambda_i, \boldsymbol{u}_i, \boldsymbol{v}_i)$ と表現することにする．すぐにわかるように $(\lambda_i, \boldsymbol{u}_i, \boldsymbol{v}_i)$ が固有系なら，$(-\lambda_i, \boldsymbol{u}_i, -\boldsymbol{v}_i)$ も固有系である．そこで (11.18) の正の固有値が p 個あるとして，改めて

$$(\lambda_i, \boldsymbol{u}_i, \boldsymbol{v}_i) \quad \lambda_i > 0 \qquad i = 1, 2, \cdots, p \tag{11.20}$$

とする．そうすると負の固有値の固有系は

$$(-\lambda_i, \boldsymbol{u}_i, -\boldsymbol{v}_i) \qquad i = 1, 2, \cdots, p \tag{11.21}$$

で表される．これで $2p$ 組の固有系が定まった．残りの $n+m-2p$ 個の固有値は 0 である．固有値 0 に対する固有ベクトルは $\boldsymbol{u}_j$ と $\boldsymbol{v}_j$ をカップルさせて解く必要はない．すなわち (11.19) 式から

$$\begin{aligned} &\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{u}_j = 0 \qquad j = p{+}1, p{+}2, \cdots, n \\ &\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{v}_j = 0 \qquad j = p{+}1, p{+}2, \cdots, m \end{aligned}$$

を解くと，残りの固有系は

$$\begin{aligned} &(0, \boldsymbol{u}_j, 0) \qquad j = p{+}1, p{+}2, \cdots, n \\ &(0, 0, \boldsymbol{v}_j) \qquad j = p{+}1, p{+}2, \cdots, m \end{aligned} \tag{11.22}$$

で表される．以上 (11.20)，(11.21)，(11.22) 式で合計 $p+p+(n-p)+(m-p) = n+m$ 組の固有系が定められたことになる．このようにして定められた固有ベクトル $\boldsymbol{w}_i$ は互いに直交しているから

$$\boldsymbol{w}_i^T\boldsymbol{w}_j = 0 \qquad i \neq j$$

が成り立つ．$\boldsymbol{w}_i$，$\boldsymbol{w}_j$ として (11.20) 式のタイプの解を選ぶとこの直交関係は

$$\boldsymbol{u}_i^T\boldsymbol{u}_j + \boldsymbol{v}_i^T\boldsymbol{v}_j = 0 \qquad i \neq j$$

となる．一方，$\boldsymbol{w}_j$ として (11.21) 式のタイプを選ぶと

$$\boldsymbol{u}_i^T\boldsymbol{u}_j - \boldsymbol{v}_i^T\boldsymbol{v}_j = 0 \qquad i \neq j$$

となる．したがって (11.20)，(11.21) 式のタイプの $\boldsymbol{u}_i$，$\boldsymbol{v}_i$ はそれぞれが直交している．また固有値 0 に属する固有ベクトルも互いに直交するようにできる．そこで以下では $\boldsymbol{w}_i$ を正規直交化するのではなく，$\boldsymbol{u}_i$，$\boldsymbol{v}_i$ それぞれを正規化することにする．すなわち

$$\boldsymbol{u}_i^T\boldsymbol{u}_j = \delta_{ij} \qquad \boldsymbol{v}_i^T\boldsymbol{v}_j = \delta_{ij} \tag{11.23}$$

ここまでは $\boldsymbol{S}$ の固有系を求めるのに腐心してきたのであったが，われわれの関心は $\boldsymbol{S}$ ではなく $\boldsymbol{A}$ である．上で求めた $\boldsymbol{u}_i$，$\boldsymbol{v}_i$ を横にならべて作られた行列を

$$\begin{aligned} \boldsymbol{U} &= [\boldsymbol{u}_1\ \boldsymbol{u}_2\ \cdots\ \boldsymbol{u}_p\ \boldsymbol{u}_{p+1}\ \cdots\ \boldsymbol{u}_n] \\ \boldsymbol{V} &= [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \cdots\ \boldsymbol{v}_p\ \boldsymbol{v}_{p+1}\ \cdots\ \boldsymbol{v}_m] \end{aligned} \tag{11.24}$$

とする．$\boldsymbol{U}$ は $n \times n$ の直交行列，$\boldsymbol{V}$ は $m \times m$ の直交行列で，いずれも最初の p 列は正の固有値に属する固有ベクトルであり，残りの列は固有値 0 に属する．これを用いると固有値問題 $\boldsymbol{A}\boldsymbol{v}_i = \lambda_i\boldsymbol{u}_i$ はまとめて

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Lambda} \tag{11.25}$$

と書くことができる．ここに $\boldsymbol{\Lambda}$ は対角成分が正の固有値 $\lambda_1, \lambda_2, \cdots, \lambda_p$，残りは 0 のの $n \times m$ 対角行列である．$\boldsymbol{U}$，$\boldsymbol{V}$ は直交行列であるから，上式に右から $\boldsymbol{V}^T$ を掛け，また左から $\boldsymbol{U}^T$ を掛ければ

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T \qquad \boldsymbol{U}^T\boldsymbol{A}\boldsymbol{V} = \boldsymbol{\Lambda} \tag{11.26}$$

が得られる．これで任意の行列の特異値分解ができたことになる．

しかしこれでもまだ冗長性が残っている．対称行列のときと同様に考えると，0 固有値に属する固有関数は結果には現れてこないので，正の固有値に属するベクトルだけから

$$\begin{aligned} \boldsymbol{U}_p &= [\boldsymbol{u}_1\ \boldsymbol{u}_2\ \cdots\ \boldsymbol{u}_p] \\ \boldsymbol{V}_p &= [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \cdots\ \boldsymbol{v}_p] \end{aligned} \tag{11.27}$$

を作り，正の固有値から $p \times p$ の対角行列

$$\boldsymbol{\Lambda}_p = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_p \end{bmatrix} \tag{11.28}$$

とすると，特異値分解は

$$\boldsymbol{A} = \boldsymbol{U}_p\boldsymbol{\Lambda}_p\boldsymbol{V}_p^T \tag{11.29}$$

と表される．$\boldsymbol{U}_p$，$\boldsymbol{V}_p$ は直交性

$$\boldsymbol{U}_p^T\boldsymbol{U}_p = \boldsymbol{I}_p \qquad \boldsymbol{V}_p^T\boldsymbol{V}_p = \boldsymbol{I}_p$$

を満たすが，正方行列ではないから前後を入れかえた式，たとえば $\boldsymbol{U}_p\boldsymbol{U}_p^T = \boldsymbol{I}_n$ は一般には成り立たない．

対角行列 $\boldsymbol{\Lambda}$ の対角要素は正または 0 で，合計

$$r = \min(n, m)$$

個ある．これらの要素 $\lambda_i \geq 0\,(1 \leq i \leq r)$ を行列 $\boldsymbol{A}$ の**特異値**という．p は行列 $\boldsymbol{A}$ のランク (階数) と呼ばれる．$\boldsymbol{A}$ が零行列でないかぎり $p \geq 1$ であり

$$1 \leq p \leq r = \min(n, m)$$

が成り立つ．$p = r$ のときをフルランク (最大階数) といい，$p < r$ のときはランク落ちがあるという．

特異値の最大と最小の比

$$\kappa(\boldsymbol{A}) = \frac{\max\limits_{1\leq i\leq r} \lambda_i}{\min\limits_{1\leq i\leq r} \lambda_i} \tag{11.30}$$

が行列 $\boldsymbol{A}$ の条件数である．ランク落ちがあるときには条件数は無限大である．ランク落ちがないとき，$\boldsymbol{A}^T\boldsymbol{A}$ や $\boldsymbol{A}\boldsymbol{A}^T$ の固有値は (11.19) 式から特異値の二乗であるから

$$\kappa(\boldsymbol{A}^T\boldsymbol{A}) = \kappa(\boldsymbol{A}\boldsymbol{A}^T) = \kappa(\boldsymbol{A})^2$$

が成り立つ．最小二乗法の問題で正規方程式を用いて解を求めると精度が悪くなるのはこのためである．

例題　$\boldsymbol{A}$ として

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{bmatrix}$$

をとると $\boldsymbol{A}^T\boldsymbol{A}$，$\boldsymbol{A}\boldsymbol{A}^T$ の固有値はそれぞれ

$$\boldsymbol{A}^T\boldsymbol{A} \ :\ \lambda^2 = 2 + \varepsilon^2,\ \varepsilon^2$$
$$\boldsymbol{A}\boldsymbol{A}^T \ :\ \lambda^2 = 2 + \varepsilon^2,\ \varepsilon^2,\ 0$$

である．したがって $\boldsymbol{A}$ のランクは $p = 2$ で，$\boldsymbol{u}$ の 1 個のベクトルが固有値 0 に属している．特異値は

$$\lambda = \sqrt{2 + \varepsilon^2}, \varepsilon$$

固有ベクトルは

$$\boldsymbol{U} = \begin{bmatrix} 2 & 0 & \varepsilon \\ \varepsilon & 1 & -1 \\ \varepsilon & -1 & -1 \end{bmatrix}$$

$$\boldsymbol{V} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

である．ただし，固有ベクトルは正規化はされていない．

連立方程式の解　$n \times m$ 行列 $\boldsymbol{A}$ を係数行列とする連立方程式

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} \tag{11.31}$$

を解くために，対称行列のときに行ったように $\boldsymbol{x}$ と $\boldsymbol{y}$ を展開する．対称行列のときとは異なり，$\boldsymbol{x}$ は m 次元ベクトルであるから m 次元の完全系 $\boldsymbol{v}_i$ で，n 次元ベクトル $\boldsymbol{y}$ は $\boldsymbol{u}_i$ で展開しなければならない．

$$\boldsymbol{x} = \sum_{i=1}^{m} x_i'\boldsymbol{v}_i \qquad \boldsymbol{y} = \sum_{i=1}^{n} y_i'\boldsymbol{u}_i \tag{11.32}$$

x_i' は $\boldsymbol{x}$ を $\boldsymbol{v}_i$ で展開したときの展開係数，y_i' は $\boldsymbol{y}$ を $\boldsymbol{u}_i$ で展開したときの展開係数である．たとえば $\boldsymbol{y}$ が与えられたとき，展開係数 y_i' は

$$y_i' = \boldsymbol{u}_i^T\boldsymbol{y}$$

で計算できる．

$\boldsymbol{x}$ の展開を (11.31) 式の左辺に代入すると

$$\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{m} x_i'\boldsymbol{A}\boldsymbol{v}_i = \sum_{i=1}^{m} x_i'\lambda_i\boldsymbol{u}_i$$

となるから，両辺の $\boldsymbol{u}_i$ の係数を比較すると

$$\lambda_i x_i' = y_i' \qquad i = 1, 2, \cdots, p \tag{11.33}$$
$$0 = y_i' \qquad i = p{+}1, p{+}2, \cdots, n \tag{11.34}$$

となる．これが (11.31) 式が解けるための必要十分条件である．係数行列が正方行列でなくても (11.31) 式は解けることに注意しなければならない．

(11.33) 式からは $\boldsymbol{x}$ の p 個の展開係数が求められるが，残りは決まらない．決まった係数だけを (11.32) 式に代入したものを $\boldsymbol{x}^\dagger$ とすると，これは

$$\begin{aligned}\boldsymbol{x}^\dagger &= \sum_{i=1}^{p} \frac{y_i'}{\lambda_i}\boldsymbol{v}_i = \sum_{i=1}^{p} \frac{\boldsymbol{u}_i^T\boldsymbol{y}}{\lambda_i}\boldsymbol{v}_i \\ &= \boldsymbol{V}_p\boldsymbol{\Lambda}_p^{-1}\boldsymbol{U}_p^T\boldsymbol{y}\end{aligned} \tag{11.35}$$

と表される．

(11.34) 式は解が存在するための適合条件である．これが成り立たないと $x_i' = y_i'/0\ (i > p)$ となって，

問題はいわゆる不能になる．いいかえれば，解が存在するためには右辺 $\boldsymbol{y}$ を勝手に選ぶことはできず，$\boldsymbol{y}$ には $\boldsymbol{u}_i$ $(i > p)$ 成分が含まれていてはならない．この条件は

$$\boldsymbol{u}_i^T \boldsymbol{y} = 0 \quad i = p+1, p+2, \cdots, n \tag{11.36}$$

と書くことができる．

任意の $\boldsymbol{x}$，$\boldsymbol{y}$ に対して，(11.31) 式の残差は

$$\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{p} (y_i' - \lambda_i x_i')\boldsymbol{u}_i + \sum_{i=p+1}^{n} y_i' \boldsymbol{u}_i$$

と表される．$\boldsymbol{x}$ が (11.31) 式の解であるための必要条件は (11.33) 式が成り立つことであるから，解に対しては少なくとも

$$\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} = \sum_{i=p+1}^{n} y_i' \boldsymbol{u}_i \tag{11.37}$$

がなりたつ．さらに適合条件 (11.34) 式が成り立てば，残差は 0 になる．(11.31) 式の解が存在するというのはこのような意味である．

係数 x_i' $(i > p)$ は決まらない．$\boldsymbol{A}$ を掛けることによってこの成分が消えてしまうからである．これは逆にいえば

$$\boldsymbol{x} = \boldsymbol{x}^\dagger + \sum_{i=p+1}^{m} x_i' \boldsymbol{v}_i \tag{11.38}$$

もこの方程式の解になっているということである．ここで係数 x_i' は任意である． この $\boldsymbol{x}$ に対しても残差は (11.37) 式で与えられる．したがって残差が最小になる $\boldsymbol{x}$ も一義的ではない．しかし $\boldsymbol{x}^\dagger$ は残差を最小にする $\boldsymbol{x}$ のなかでノルムが最小になっている．これは一義的に決まる．

以下では n, m, p の大小関係ごとにコメントを付け加える．

正方行列 ($\boldsymbol{n = m}$) $n = m$ の場合は最も馴染みのある問題，通常の連立一次方程式の問題である．

$p = n = m$ のときを最大階数 (フルランク) という．このときにはすべての固有値が 0 でないから，$\boldsymbol{A}$ は正則行列であり，(11.33) 式は $i = n$ まで成立して解は (11.35) 式

$$\boldsymbol{x} = \boldsymbol{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T\boldsymbol{y}$$

である．この解は任意の $\boldsymbol{y}$ に対して成り立ち，もちろん一義的である．この式を (11.31) 式の解の計算に用いることはできるが，固有値，固有ベクトル計算に手間がかかることの上に，丸め誤差が大きくなるので避けた方がよい．

$p < n$ のときをランク落ちがあるという．ランク落ちの最も簡単な例は，左辺の係数が同じ式が複数個ある場合である．このときには (11.38) 式は残差 (11.37) を最小にする．適合条件 (11.36) 式が成り立っていれば残差が 0 になるから，これは解になっている．しかし一義的ではない．残差を最小にし，しかも解のノルムを最小にするのは $\boldsymbol{x}^\dagger$ である．

$\boldsymbol{n > m}$ の場合 $n > m$ のときには未知数の数より方程式の数のほうが多い，いわゆる最小二乗法の問題の場合である．この場合には $p = m$ がフルランクで，このときには (11.38) 式の第二項は現れないから，$\boldsymbol{x}^\dagger$ が残差を最小にする解で，これは一義的に決まる．ランク落ちがある場合には (11.38) 式の第二項が現れるので，最小二乗解は一義的でなくなる．ランク落ちがあってもなくても，適合条件 (11.36) が成り立てば残差は 0 になる．

$\boldsymbol{n < m}$ の場合 §10 の直接観測の式 (10.50) のように，未知数の数よりも方程式の数の方が少ないときには，一義的な解が存在しそうにないことは直感的に明らかである．フルランクは $p = n$ のときで，このときには残差 (11.37) が 0 になる．しかしこのときでも (11.38) 式の第二項は存在するから，解は一義的でない．ランク落ちがあると残差は 0 でなくなる．

一般逆行列 ある行列 $\boldsymbol{A}$ に対して

$$\begin{aligned} &\text{(i)} \quad \boldsymbol{A}\boldsymbol{X}\boldsymbol{A} = \boldsymbol{A} \\ &\text{(ii)} \quad \boldsymbol{X}\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X} \\ &\text{(iii)} \quad (\boldsymbol{A}\boldsymbol{X})^T = \boldsymbol{A}\boldsymbol{X} \\ &\text{(iv)} \quad (\boldsymbol{X}\boldsymbol{A})^T = \boldsymbol{X}\boldsymbol{A} \end{aligned} \tag{11.39}$$

を満足する行列 $\boldsymbol{X}$ を，ムーア・ペンローズの意味での一般逆行列という．$\boldsymbol{A}$ が正則な正方行列のとき，その逆行列 $\boldsymbol{A}^{-1}$ が上の 4 条件を満たしていることは明らかである．

先に導いた解 (11.35) は，正確には (11.31) 式を満たしていないが，残差のノルムを最小にし，しか

もノルムが最小になるという意味での，広い意味の (11.31) 式の解になっている．そこで

$$\boldsymbol{A}^{\dagger} = \boldsymbol{V}_p \boldsymbol{\Lambda}_p^{-1} \boldsymbol{U}_p^T \tag{11.40}$$

を $\boldsymbol{A}$ の一種の逆行列と考え，これをランチョスの自然逆行列と呼ぶ．実はこれが (11.39) 式の 4 条件を満たすことは実際に $\boldsymbol{X} = \boldsymbol{A}^{\dagger}$ に対して左辺を計算してみればわかる．

$\boldsymbol{A}^{\dagger}$ の中には特異値の逆数が現れる．定義により $\boldsymbol{\Lambda}_p$ に含まれる特異値は 0 ではないから代数的には問題ない．しかし特異値を数値的に求めた場合には，本来 0 であるべき固有値が計算誤差のために非常に小さな特異値として現れることもある．また元来の特異値の中にも小さなものがあるかもしれない．小さな特異値があれば計算誤差が拡大されてしまうから，むしろ初めから小さな特異値を除いてしまった方が数値計算の誤差は小さくなるかもしれない．

そこで特異値を大きな方から順番を付け，大きい方から q 個の特異値と対応する固有ベクトルから，(11.40) 式と同様にして

$$\boldsymbol{A}_q^{\dagger} = \boldsymbol{V}_q \boldsymbol{\Lambda}_q^{-1} \boldsymbol{U}_q^T \tag{11.41}$$

を定義する．$\boldsymbol{A}^{\dagger}$ は $m \times n$ の行列である．これを $\boldsymbol{A}$ の**有効一般逆行列**と呼ぶことにする．次にこれが一般逆行列の条件を満たしているかどうかを検討する．

$\boldsymbol{A}$ の分解 (11.29) 式を用いると

$$\boldsymbol{A}_q^{\dagger}\boldsymbol{A} = (\boldsymbol{V}_q \boldsymbol{\Lambda}_q^{-1} \boldsymbol{U}_q^T)(\boldsymbol{U}_p \boldsymbol{\Lambda}_p \boldsymbol{V}_p^T)$$

であるが，$q \times p$ 行列 $\boldsymbol{U}_q^T \boldsymbol{U}_p$ は $\boldsymbol{u}_i$ の正規直交性により $q \times q$ の単位行列の右に $p - q$ 列の 0 ベクトルを付け加えたものである．その結果

$$\boldsymbol{U}_q^T \boldsymbol{U}_p \boldsymbol{\Lambda}_p \boldsymbol{V}_p^T = \boldsymbol{\Lambda}_q \boldsymbol{V}_q^T$$

となる．同様な計算を行なった結果をまとめると次のようになる．

$$\begin{aligned}
&\text{(i)} \quad \boldsymbol{A}\boldsymbol{A}_q^{\dagger}\boldsymbol{A} = \boldsymbol{U}_q \boldsymbol{\Lambda}_q \boldsymbol{V}_q^T \\
&\text{(ii)} \quad \boldsymbol{A}_q^{\dagger}\boldsymbol{A}\boldsymbol{A}_q^{\dagger} = \boldsymbol{A}_q^{\dagger} \\
&\text{(iii)} \quad \boldsymbol{A}\boldsymbol{A}_q^{\dagger} = \boldsymbol{U}_q \boldsymbol{U}_q^T \\
&\text{(iv)} \quad \boldsymbol{A}_q^{\dagger}\boldsymbol{A} = \boldsymbol{V}_q \boldsymbol{V}_q^T
\end{aligned} \tag{11.42}$$

$q = p$ のときには (11.29) 式より上の (i) の右辺は $\boldsymbol{A}$ になるがそれ以外は $\boldsymbol{A}$ にならない．しかし $q < p$ のときでも (ii) 〜(iv) は条件 (11.39) を満たしている．したがって有効一般逆行列 $\boldsymbol{A}_q^{\dagger}$ は厳密には一般逆行列ではないが，一般逆行列に非常に近い．

分解能，情報量　有効一般逆行列を用いた (11.31) 式の解を改めて

$$\boldsymbol{x}^{\dagger} = \boldsymbol{A}_q^{\dagger}\boldsymbol{y} \tag{11.43}$$

と書くことにする．$\boldsymbol{x}$ が真の解であったとすれば (11.31) 式を満たしているはずであるから

$$\boldsymbol{x}^{\dagger} = \boldsymbol{R}_q \boldsymbol{x} \qquad \boldsymbol{R}_q = \boldsymbol{A}_q^{\dagger}\boldsymbol{A} \tag{11.44}$$

と書くことができる．これは真の解と計算された解の関係を表している．$\boldsymbol{x}^{\dagger}$ の i 成分は，真の $\boldsymbol{x}$ の成分を $\boldsymbol{R}_q$ の i 行目を重みとして平均したものになっている．$\boldsymbol{x}^{\dagger}$ が真の解に等しくなるためには $\boldsymbol{R}_q$ は単位行列でなければならない．行列 $\boldsymbol{R}_q$ を分解能行列と呼ぶ．われわれの $\boldsymbol{A}_q^{\dagger}$ の定義では (11.42) 式から

$$\boldsymbol{R}_q = \boldsymbol{V}_q \boldsymbol{V}_q^T$$

であるから，$q = m$ のとき以外は $\boldsymbol{R}_q$ が単位行列になることはない．

解 $\boldsymbol{x}^{\dagger}$ に対応する $\boldsymbol{y}$ の計算値は

$$\boldsymbol{y}^{\dagger} = \boldsymbol{A}\boldsymbol{x}^{\dagger} \tag{11.45}$$

である．これは (11.43) 式を用いて書きかえると

$$\boldsymbol{y}^{\dagger} = \boldsymbol{S}_q \boldsymbol{y} \qquad \boldsymbol{S}_q = \boldsymbol{A}\boldsymbol{A}_q^{\dagger} \tag{11.46}$$

となる．$\boldsymbol{S}_q$ を情報量行列と呼ぶ．$\boldsymbol{y}$ と $\boldsymbol{y}^{\dagger}$ の関係は，$\boldsymbol{x}$ と $\boldsymbol{x}^{\dagger}$ の関係と同じである．$\boldsymbol{S}_q$ が単位行列に近いほど $\boldsymbol{y}$ と $\boldsymbol{y}^{\dagger}$ の対応がよくなる．$\boldsymbol{A}_q^{\dagger}$ が有効一般逆行列のときには (11.42) 式から

$$\boldsymbol{S}_q = \boldsymbol{U}_q \boldsymbol{U}_q^T \tag{11.47}$$

で表される．これは $q = n$ のときに単位行列になるが，最小二乗法の場合には $n > m \geq p \geq q$ であるから，$\boldsymbol{S}_q$ が単位行列になることはあり得ない．

解の分散　(11.31) 式を最小二乗法の観測方程式と考えたとき，$\boldsymbol{x}$ はモデルのパラメーター，$\boldsymbol{A}\boldsymbol{x}$ はモデルから期待される観測値，$\boldsymbol{y}$ は実際の観測値である．$\boldsymbol{y}$ は確率変数であるから，$\boldsymbol{x}$ の推定量 (11.43)

式も確率変数である．§10 で導いた誤差の伝播法則を用いると，$\boldsymbol{x}^\dagger$ の共分散行列 $\boldsymbol{C}_{x^\dagger}$ は

$$\boldsymbol{C}_{x^\dagger} = \boldsymbol{A}_q^\dagger \boldsymbol{C}_y (\boldsymbol{A}_q^\dagger)^T \tag{11.48}$$

で表される．$\boldsymbol{C}_y$ は §10 と同様に観測値 $\boldsymbol{y}$ の共分散行列である．

簡単のために観測値 $\boldsymbol{y}$ の成分間には相関がなく，i 成分の分散は σ_i^2，すなわち $\boldsymbol{C}_y$ は σ_i^2 を対角成分とする対角行列とする．このとき $\boldsymbol{x}^\dagger$ の i 成分の分散は

$$(\boldsymbol{C}_{x^\dagger})_{ii} = \sum_{k=1}^{n} \Big(\sum_{j=1}^{q} \frac{v_{ij} u_{kj}}{\lambda_j} \Big)^2 \sigma_k^2 \tag{11.49}$$

と書き表される．ここに u_{ij}, v_{ij} は $\boldsymbol{u}_j$, $\boldsymbol{v}_j$ の成分を表し，たとえば

$$\boldsymbol{u}_j = [u_{1j}\ u_{2j}\ \cdots\ u_{nj}]^T$$

である．分解能行列や情報量行列を単位行列に近づけるためにはできるだけ沢山の特異値を取り込む方がよい．しかしあまり沢山取り込むと，小さな特異値が含まれてしまうので，上式により解の分散が大きくなる．したがって q は分散，分解能行列，情報量行列のかね合いによって決めなければならない．

例題　最初の例題は非常に人工的な行列

$$\boldsymbol{A} = \begin{bmatrix} 1 & 6 & 11 \\ 2 & 7 & 12 \\ 3 & 8 & 13 \\ 4 & 9 & 14 \\ 5 & 10 & 15 \end{bmatrix}$$

を用いる．この行列の特異値を倍精度で計算した結果は

$$\lambda = \begin{bmatrix} 35.127223 \\ 2.646397 \\ 5.15\text{e-}16 \end{bmatrix}$$

となった．最後の特異値は代数的には 0 である特異値が丸め誤差のためにこうなったと考えることができる．したがってこの行列はランク落ちがあって $p = 2$ であると判断される．

そこで $q = p = 2$ として分解能行列 $\boldsymbol{R}_2$，情報量行列 $\boldsymbol{S}_2$ を計算すると

$$\boldsymbol{R}_2 = \begin{bmatrix} 0.83 & 0.33 & -0.16 \\ 0.33 & 0.33 & 0.33 \\ -0.16 & 0.33 & 0.83 \end{bmatrix}$$

$$\boldsymbol{S}_2 = \begin{bmatrix} 0.6 & 0.4 & 0.2 & 0.0 & -0.2 \\ 0.4 & 0.3 & 0.2 & 0.1 & 0.0 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.0 & 0.1 & 0.2 & 0.2 & 0.4 \\ -0.2 & 0.0 & 0.2 & 0.4 & 0.6 \end{bmatrix}$$

になる．$\boldsymbol{R}_2$ の要素はすべて循環小数で，たとえば 0.16 とあるのは 0.16666$\cdots$ の略である．この行列はランク落ちがあるために分解能行列，情報量行列ともに非常に性質が悪くなっている．

二番目の例は §10 でも用いた行列

$$\boldsymbol{A} = \begin{bmatrix} 1.26 & 0.84 & 0.63 & 0.504 \\ 0.84 & 0.63 & 0.504 & 0.42 \\ 0.63 & 0.504 & 0.42 & 0.36 \\ 0.504 & 0.42 & 0.36 & 0.315 \\ 0.42 & 0.36 & 0.315 & 0.28 \end{bmatrix}$$

である．この行列の特異値は

$$\lambda = \begin{bmatrix} 2.5582006 \\ 0.1799580 \\ 6.2085986\text{e-}3 \\ 9.9670848\text{e-}5 \end{bmatrix}$$

である．これはフルランクであり ($p = 4$)，条件数は

$$\kappa = 2.6 \times 10^4$$

であるからそれほど大きくはない．しかし $\boldsymbol{A}^T\boldsymbol{A}$ の条件数はこの二乗の 7×10^8 になるから，正規方程式を単精度で解くには条件数が大きすぎる．

この行列はフルランクであるから $q = p = 4$ に選べば分解能行列は単位行列になる．情報量行列がどのようなものであるかを見るためにあえて書き上げると次のようになる．ただし $\boldsymbol{S}_q$ は対称行列であるから右上だけを示してある．

$$\boldsymbol{S}_4 = \begin{bmatrix} 0.99982 & 0.00211 & -0.00739 & 0.00986 & -0.00444 \\ & 0.97465 & 0.08873 & -0.11831 & 0.05324 \\ & & 0.68944 & 0.41408 & -0.18634 \\ & & & 0.44789 & 0.24845 \\ & & & & 0.88820 \end{bmatrix}$$

§10 と同様に

$$\boldsymbol{x} = [1/2,\ 1/3,\ 1/4,\ 1/5]^T$$

から「観測値」を $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ によって計算する．この行列に対する $\boldsymbol{U}$ や $\boldsymbol{V}$ は示していないが，$\boldsymbol{A}_4^\dagger$ を計算し，$\boldsymbol{x}^\dagger$ を計算すると次のようになった．

$$\boldsymbol{x}^\dagger = \begin{bmatrix} 0.50000119 \\ 0.33332490 \\ 0.25001625 \\ 0.19999081 \end{bmatrix}$$

この解はグラム・シュミット法やハウスホールダー法で計算したものよりも精度が悪い．これは連立方程式の解を逆行列を通して計算すると，丸め誤差のために消去法よりも精度が悪くなるのと同様な理由である．

ランチョスの名前は前にも出てきたかと思うが，非常にユニークな発想をする人である．参考文献は，題目には微分演算子となっているが，内容は微分演算子，行列，積分演算子などを一つの見方で押し通している．特異値分解の項はこの教科書によるところが多い．

参考文献

Lanczos, C. (1961) : Linear Differential Operators, Van Nostrand.

12 整列法（ソーティング）

ここで「整列」というのは，与えられた数値データを大きい順に (降順)，あるいは小さい順 (昇順) に並べかえることを意味している．これは数値「計算」とはいえないかもしれないが，数値計算の一部としてこのような並べかえが必要になる例は多い．

たとえば地震動から震度を求める計算がある．気象庁が定めている計算法は複雑であるが，簡単にいえば，たとえば測定間隔が 1/100 秒のときには加速度値を大きさの順に並べた 30 番目の加速度値から震度が決まる．

整列法は計算科学の重要なテーマの一つであり，これだけで一冊の本が書けるくらいであるが，以下に代表的な整列法を述べる．どちらでも同じことであるから，大きさの順に並べる方法だけを考える．

> Knuth といえば今では TEX の創始者として有名であるが，計算科学の創始者のひとりでもあり，'The Art of Computer Programming' 十何巻をひとりで書いたというプログラミングの名人でもある．この本の第三巻が整列法にあてられている (全巻の和訳がある)．
> TEX といえば，このノートははじめ PC9801 の上で MicroTEX を使って書いたものだった．

挿入法 (straight insertion)　適当な名前がないのでこう呼んでおくが，誰でも考え付く最も単純な方法である．いま，n 個のデータ $a(1), a(2), \cdots, a(n)$ が与えられたとして，そのうちの最初の k 個を大きさの順に並べ替えたものを改めて $a(1), a(2), \cdots, a(k)$ とする．すなわち

$$a(1) \geq a(2) \geq \cdots \geq a(k)$$

が成り立っているとする．次のデータ $a(k{+}1)$ を取ってきたときに，これをどの位置に挿入すればいいのかが問題である．$a(k) \geq a(k{+}1)$ ならなにもしないで次のデータ $a(k{+}2)$ にいけばいい．$a(k) < a(k{+}1)$ なら $a(k)$ を $a(k{+}1)$ の位置に移動して次に $a(k{-}1)$ と $a(k{+}1)$ の比較をする．このように順次上位のデータとの比較という手続きを繰り返せば，どこかで当面問題にしている $a(k{+}1)$ を入れるべき位置が見つかる．これはプログラムで見た方がわかりやすい．

```
      do  k=1, n-1
        ak=a(k+1)
        do  i=k, 1, -1
          if( a(i).lt.ak )  then
            a(i+1) = a(i)
          else
            goto  1
          endif
        enddo
1       a(i+1)=ak
      enddo
```

挿入法のプログラム

このプログラムの内側のループでは最大 k 回のデータの比較と挿入を行っている．これを平均 $k/2$ 回とすると計算量の見積もりは

$$\sum_{k=1}^{n-1} \frac{k}{2} = \frac{1}{4}n(n-1) \sim \frac{1}{4}n^2 \qquad (12.1)$$

となる．n が 100 程度ならともかく，1000 にもなるとこの計算量はばかにならない．しかし上のプログラムは短くて，必要なところに埋め込むことができるから，n が 100 程度なら使ってもいいだろう．

比較の回数を減らすことは簡単にできる．$a(k{+}1)$ が $a(1)$ と $a(k)$ の間に入っていることがわかっているとき，端から比較を行うのではなく，その中央にある

$$a(l) \quad \text{ただし} \quad l = \frac{k+1}{2}$$

との比較を行う．l の端数は切捨てでよい．もし $a(l) > a(k+1)$ ならこんどは $a(l)$ と $a(k)$ の中央との比較を行う．このように区間を半分にしながら比較を行っていけば，最後には $a(k{+}1)$ を入れるべき場所が決まる．

この比較の回数は $\log_2 k$ である．したがって比較の総数は

$$\sum_{k=1}^{n-1} \log_2 k = \log_2 (n-1)!$$

になる．階乗に対するスターリングの公式を用いると

$$\log_2 (n-1)! \sim n \log_2 n$$

となる．たとえば $n = 10^3 \doteq 2^{10}$ のとき，$n^2/4$ と上の見積もりとの比は 25:1 になる．ただし，この方法ではデータの移動の回数は減少しない．プログラムは次のようになる．

```
      do  k=1, n-1
        ak=a(k+1)
        if( a(k).ge.ak )  goto  3
        if( a(1).lt.ak )  then
          i=1
        else
          il=1
          ir=k
1         i=(il+ir+1)/2
          if( i.eq.ir )  goto  2
          if( a(i).gt.ak )  then
            il=i
          else
            ir=i
          endif
        goto  1
        endif
2       do  j=k, i, -1
          a(j+1)=a(j)
        enddo
        a(i)=ak
3     enddo
```

二分法を用いた挿入法

シェル法　単純な挿入法に比べて少しましな方法にシェル法がある．この方法の計算量は最悪の場合でも $n^{3/2}$ である．したがって中程度の大きさのデータに用いることができる．

データ $a(i)$ を奇数番目と偶数番目とでそれぞれに並べかえて

$$a(1) \geq a(3) \geq a(5) \geq \cdots$$
$$a(2) \geq a(4) \geq a(6) \geq \cdots$$

が成り立っていたとする．この状態では $a(i) \geq a(i+1)$ にはなっていないが，ほとんどのデータはあるべき位置の近くにいる．したがって全体 $a(1), a(2), a(3), \cdots$ を並べかえようとしたときに，データを移動する回数は少なくなる．

一つおきに整列する前は，一般に n_k おきにデータを整列させる．すなわち

$$a(1) \geq a(n_k+1) \geq a(2n_k+1) \geq \cdots$$
$$a(2) \geq a(n_k+2) \geq a(2n_k+2) \geq \cdots$$
$$\cdots$$
$$a(n_k) \geq a(2n_k) \geq a(3n_k) \geq \cdots$$

が成り立つようにする．この整列には単純な挿入法を用いる．n_k は最後には 1 になる数列であるが，推奨されているのは

$$n_k = \frac{3^k - 1}{2}$$

すなわち

$$n_1 = 1 \quad n_{k+1} = 3n_k + 1 \tag{12.2}$$

で生成される数列である．$n > n_k$ となる最大の n_k から始めて飛び飛びの整列を繰り返せばよい．

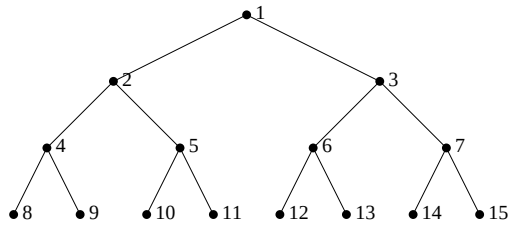
```
      nk=1
      do  k=1, n
        nk=3*nk+1
        if( nk.gt.n )  goto  1
      enddo
1     nk=nk/3
      if( nk.le.0 )  return
      do  i=nk+1, n
        ai=a(i)
        do  j=i, nk+1, -nk
          if( a(j-nk).lt.ai )  then
            a(j)=a(j-nk)
          else
            goto  2
          endif
        enddo
2       a(j)=ai
      enddo
    goto  1
```

シェル法

ヒープソート　先に用いた二分法を利用した整列法として代表的なものにヒープ・ソートがある．ヒープ (heap) は藁などを積み重ねた山のことをいうが，ここでは

$$a(i/2) \leq a(i) \quad 1 \leq i \leq n \tag{12.3}$$

を満たすような数列をいう．ここで $i/2$ は整数の割り算 (小数以下は切り捨て) である．このような数列は二分木の形に書くとわかりやすい．



この図では要素を番号だけで表している．$a(1)$ は $a(2), a(3)$ より小さい (正確には，等しいか小さい)．$a(2)$ は $a(4), a(5)$ より小さく，$a(3)$ は $a(6), a(7)$ より小さい．しかし $a(2)$ と $a(3)$ はどちらが大きいとも小さいとも決まっていない．この図からはまたひとつ上のレベルに上がるには添字を 2 で割ればよいこと，下のレベルに下りるには添字を 2 倍すればよいことがわかる．

一旦このようなヒープができると整列は簡単である．$a(1)$ が最も小さいからこれと $a(n)$ を交換する．新しい $a(1)$ は条件 (12.3) を満たしていないから $a(2)$ と $a(3)$ のうちの小さい方と $a(1)$ を交換する．たとえばそれが $a(3)$ であったとすると，$a(6)$ と $a(7)$ の小さいと $a(3)$ を交換する，というようなことを繰り返していけばいつかは $a(n)$ の落ち着き場所が決まる．このような操作をシフトダウンという．データの数が一つ少なくなるが，ヒープの性質は保たれている．一番上にある新しい $a(1)$ は全体のデータの二番目に小さな要素である．そこでつぎにはこれと $a(n-1)$ を取りかえて，再びシフトダウンを行う．こういう操作を繰り返すと $a(1)$ には常に最小の要素が作られ，$a(n)$ から逆順には小さな順にデータが並ぶことになる．

それではそもそもヒープをどうやってつくるか．$n = 15$ の場合を例にとれば，$a(14)$ と $a(15)$ のうちの小さい方と $a(7)$ を比較して小さい方を $a(7)$ に入れる．同様に $a(12), a(13), a(6)$ のうちの小さい方を $a(6)$ に入れる．これを左端まで行うと二分木の最低のレベルとその一つ上のレベルでは (12.3) 式が成り立っている．次に $a(3)$ を取り出し $a(6), a(7)$ と比較する．$a(3)$ が最も小さければ何もしなくてよいが，たとえば $a(7)$ の方が小さければ $a(3)$ と $a(7)$ を入れかえなければならない．しかし新しい $a(7)$ が $a(14), a(15)$ よりも小さいという保証はないので比較して入れかえをする必要がある．このようにシフトダウンしながら適当な位置を捜すのは整列するときと同様である．これらの計算は二分木を上下にたどっていくものであるから，計算量は

$$n \log_2 n \tag{12.4}$$

である．

二分木を作るときも，整列を行うときも，同じシフトダウンのプログラムを用いることが可能である．こうするとプログラムは少し複雑になるが次のようになる．

```
      il=n/2+1
      ir=n
      if( il.gt.1 )  then
        il=il-1
        ai=a(il)
      else
        ai=a(ir)
        a(ir)=a(1)
        ir=ir-1
        if( ir.eq.1 )  then
          a(1)=ai
          return
        endif
      endif
      i=il
      j=2*il
    2 if( j.gt.ir )  goto  3
      if( j.lt.ir .and. a(j).gt.a(j+1) )
     *   j=j+1
      if( ai.gt.a(j) )  then
        a(i)=a(j)
        i=j
        j=2*j
      else
        j=ir+1
      endif
      goto  2
    3 a(i)=ai
      goto  1
```

ヒープソート

> Fortran のプログラムはからみあったスパゲッティのように，ロジックの道筋がたどれない，という悪口をいわれることが多い．たしかに C 言語は論理的には明解であるが，上のような簡単なプログラムでも，C で書くと括弧と括弧の対応がとりにくくなり，for や do で break したときにどこに抜け出るのかがわかりにくい．そこにいくと Fortran では抜けていく先が明示的に示されているので，わかりやすい．もっともこれは Fortran になれた老人の繰言である．

クイックソート　この方法は最も速いといわれている方法である．数列の中からある要素 x を選び出し，交換を繰り返して x の左側の要素はすべて $a(i) \geq x$ を満足し，x の右側の要素はすべて $a(j) \leq x$ を満たすようにする．次に左側の部分列から別の要素 y を選んで $a(i) \geq y$ の部分列と $a(j) \leq x$ の部分列に分ける．右側の部分列についても同様なことを行う．こうして生成された部分列の内部では整列されていないが，部分列間では大きさの順に並んでいる．たとえば

$$(8, 10, 7, 9) \geq (5, 3, 7) \geq (1, 3, 0, 2)$$

では左側の部分列の要素は右側の部分列の要素より大きいか等しい．そこで部分列の長さが十分短くなったら，たとえば 10 よりも短くなったら，単純な挿入法で並べ替えてやればよい．

部分列に分解するにはつぎのようにする．まずある要素 x を選び，ポインター i を 1 から増やしながら最初に $a(i) < x$ となる要素を見つける．一方，ポインター j は n から 1 づつ減らしながら最初に $a(j) > x$ となる要素を見つける．$a(i)$ は左側の部分列に入れるべき要素ではないし，$a(j)$ は右側の部分列に入れるべきではないので，これらを交換する．引き続き i を増やし，j を減らして交換を続けていくと二つのポインターが交差する．ここが x の入るべき位置である．

クイックソートでは部分列の初めと終わりのポインターを記憶しておくために最大 $2\log_2 n$ 個のメモリーが必要である．しかし $n = 10^6 \doteq 2^{20}$ のときでもこの量は 40 個であるから，たいした量ではない．

クイックソートのプログラムは上に述べたことをそのまま丁寧にコーディングすればよい．プログラムは長くなるが下に示す．なお，このプログラムは Numerical Recipes in C を参考にして最近書き直したものである．

```
      ir=n
      il=1
      is=0
    1 if( ir-il.lt.m )  then
        do  j=il+1, ir
          aj=a(j)
          do  i=j-1, il, -1
            if( a(i).ge.aj )  goto  2
            a(i+1)=a(i)
          enddo
          i=il-1
    2     a(i+1)=aj
        enddo
        if( is.eq.0 )  return
        ir= istck(is)
        il=istck(is-1)
        is=is-2
      else
        k=(il+ir)/2
        tmp=a(k)
        a(k)=a(il+1)
        a(il+1)=tmp
        if( a(il+1).lt.a(ir) )  then
          tmp=a(il+1)
          a(il+1)=a(ir)
          a(ir)=tmp
        endif
        if( a(il).lt.a(ir) )  then
          tmp=a(il)
          a(il)=a(ir)
          a(ir)=tmp
        endif
        if( a(il+1).lt.a(il) )  then
          tmp=a(il+1)
          a(il+1)=a(il)
          a(il)=tmp
        endif
        i=il+1
        j=ir
        aj=a(il)
    3   do  k=il+1, ir
          if( a(i).lt.aj )  goto  4
          i=i+1
```

```
      enddo
4     do  k=il+1, ir
        if( a(j).gt.aj )  goto  5
        j=j-1
      enddo
5     if( j.ge.1 )  then
        tmp=a(i)
        a(i)=a(j)
        a(j)=tmp
        goto  3
      endif
      a(il)=a(j)
      a(j)=aj
      is=is+2
      if( is.gt.nstck )  then
        write(6,*)  'stack full'
        return
      endif
      if( ir-i+1.ge.j-1 )  then
        istck(is)=ir
        istck(is-1)=i
        ir=j-1
      else
        istck(is)=j-1
        istck(is-1)=il
        il=i
      endif
    endif
    goto  1
```

クイックソート

ここで istck は部分列の左と右の位置を入れておくための配列で，nstck はその大きさである．先にも述べたように，nstck としては 50 もとっておけば十分である．また部分列の長さが m 以下になったら挿入法で整列することにしているが，m としては 10 以下の値をとればよい．

上位 m 番目までの整列　最初にあげた例のように，1000 個のデータすべての順番が必要なわけではなく，大きさの順に 30 番目までが必要となる場合がある．このようなときにはすべてを並べかえたあとで 30 番目までをとるのはばかげている．

いま n 個のデータのうちの上から m 番目までが必要であるとする．$m \ll n$ のときには単純な挿入法を用いても問題はないであろう．そこで最初にあげたプログラムをほんの少し修正した

```
      do  k=1, n-1
        ak=a(k+1)
        kx=min(k, m)
          do  i=kx, 1, -1
            if( a(i).lt.ak )  then
              a(i+1)=a(i)
            else
              goto  1
            endif
          enddo
          i=0
1         a(i+1)=ak
        enddo
```

上位 m 番目までの整列

とすればよい．二分法を用いればもっと速くなる．

震度をリアルタイムで計測するあるシステムでは，実際に 1000 個のデータを全部並べかえてから 30 番目を選ぶという方法をとっていた．

インデックスの作成　これまでは初めに与えられた配列 $a(i)$ の順序を大きさの順に実際に並べ替えてしまったが，もとのままに残しておきたいことがある．たとえば縦方向に学生の氏名，横方向には一人一人の学生の数学の点数，物理の点数，英語の点数などが並んでいたとする．数学の点数の順に並べ替えを行ってしまうと，横方向の対応が壊れてしまう．

そこでもとの配列は残したままで成績の順番と名簿の番号の対応表を作ることにする．

i	1	2	3	4	5
$a(i)$	45	68	93	70	50
$indx(i)$	3	4	2	5	1

$indx(i)$ は成績が i 番目の学生の学生番号を表している．上の例では成績が 1 番の学生は学生番号 3，成績が 2 番目の学生は学生番号 4，$\cdots$ である．

この表を作るのは簡単である．初めに $indx(i) = i$ としておく．配列 $a(i)$ の入れかえを行うときには $a(i)$ ではなく $indx(i)$ の方を入れかえる．実際の配列の要素は $indx(i)$ を用いて捜すことができる．シェルソートのときには次のように変更すればよい．

```
      do  i=1, n
        indx(i)=i
      enddo
      nk=1
      do  k=1, n
        nk=3*nk+1
        if( nk.gt.n )  goto  1
      enddo
    1 do  k=1, n
        nk=nk/3
        if( nk.le.0 )  return
        do  i=nk+1, n
          ai=a(indx(i))
          ia=indx(i)
          do  j=i, nk+1, -nk
            if( a(indx(j-nk)).lt.ai )
     *        then
              indx(j)=indx(j-nk)
            else
              goto  2
            endif
          enddo
    2     indx(j)=ia
        enddo
      enddo
```

インデックスを用いたシェル法

ヒープソートやクイックソートのときでもわずかの変更でインデックスを用いたソーティングを行うことができる．ただしインデックスを用いると配列要素へのアクセスに手間取るので，計算スピードは遅くなる．

計算法の比較　計算量の比較をするために，実行された実数の代入文の合計を測定した．データとしては同じ乱数を用い，$n = 1000$ と $n = 5000$ の場合の結果を下表に示す．ただしこれはただ一回の試行の結果である．

方法	$n = 1000$	$n = 5000$	行数
Insertion	247,286	6,180,687	10
Shell	19,230	136,715	20
Heap	12,068	72,063	30
Quick	8,452	50,699	70

挿入法ははじめに示した見積もり $n^2/4$ に非常に近い値になっている．シェル法は $n^{3/2}$ に比例し，ヒープソートは $n \log_2 n$ に比例している．クイックソートはこの例では $n^{1.1}$ に比例しているようにみえるが，二例だけからはなんともいえない．確かにクイックソートは最も速いが，ヒープソートに比べて劇的に速いというわけではない．

> 現在では計算機のスピードが速くなったため，$n = 10,000$ になっても，単純な挿入法は例外として，シェル法，ヒープソート，クイックソートの差は測定誤差の範囲内である．また最近の計算機では，PC といえども，同時に複数のタスクが走っているので，実行のたびに計算時間が変わってしまう．実行時間ではなく，代入の回数を比較の対象にしたのはそのためである．

上の表の最後の欄は，例として示したプログラムの行数である．10 行のプログラムのかわりに 70 行の手間をかければ，スピードは数十倍にもなる．

13 線型フィルターと z 変換

本節では線型のディジタルフィルターを取り扱う．線型という意味は，入力が α 倍になると出力も α 倍になる，二つの入力が同時に入ったときの出力はそれぞれが独立に入ったときの出力の和になる，というような意味である．入力の絶対値をとるフィルターや，二乗を作るフィルターなどもあるが，これらは線型フィルターではない．

移動平均 データに含まれるランダムなノイズを消すために，移動平均がよく用いられる．これは最も簡単なディジタルフィルターの例である．原データを x_j とすると，その前後の値 x_{j-k} に重み w_k を掛けて加え合わせたものが移動平均である．したがって平滑化されたデータ y_j は

$$y_j = \sum_{k=-M}^{M} w_k x_{j-k} \tag{13.1}$$

で表される．w_k は有限長である必要はないが，移動平均であるからにはデータ x_j が定数のときには y_j も定数でなければならないから，条件

$$\sum_{j=-\infty}^{\infty} w_j = 1 \tag{13.2}$$

を満たしていなければならない．

上式 (13.1) は畳み込みにほかならない．したがって y_j のスペクトルは w_j のスペクトルと x_j のスペクトルの積で表せそうであるが，そのためには入力 x_j は (9.1) 式の条件 $\sum |x_j| < \infty$ を満足していなければならない．ここでは微動のように

$$\sum_{j=-\infty}^{\infty} |x_j| \to \infty$$

のようなデータも考えたいので，§9 の議論はそのままでは使えない．

しかし移動平均 (13.1) がどのような特性をもっているかは知ることができる．いま，入力が単振動

$$x_j = e^{-ij\omega}$$

であるとする．これは (9.1) 式を満たしてはいない．ω は §9 で定義した無次元の角周波数である．このときの出力は (13.1) 式から

$$y_j = \sum_k w_k e^{-i(j-k)\omega} = e^{-ij\omega} \sum_k w_k e^{ik\omega}$$

であるから，入力と出力の比は

$$\frac{\text{出力}}{\text{入力}} = \frac{y_j}{x_j} = \sum_k w_k e^{ik\omega} = W(\omega) \tag{13.3}$$

と表されることがわかる．$W(\omega)$ は w_j のフーリエ変換にほかならない．これをフィルターの周波数応答と呼ぶ．

w_j が実数でも $W(\omega)$ は一般に複素数である．これを極表示して振幅と位相角にわけて

$$W(\omega) = |W(\omega)| e^{i\phi(\omega)} \tag{13.4}$$

とする．入力が単振動のときの出力は

$$y_j = |W(\omega)| e^{-i(j\omega - \phi)}$$

となる．すなわち出力の振幅は入力の $|W(\omega)|$ 倍になり，位相は $\phi(\omega)$ だけ遅れる．位相の遅れを時間に直すと

$$\tau_p(\omega) = \frac{\phi(\omega)}{\omega} \tag{13.5}$$

である．これを位相遅延時間 (phase delay time) という．τ_p の単位はサンプルの間隔 Δt である．

零位相 移動平均をとったときに，x_j の山や谷の位置がずれてしまっては困る．そのようなことが起こらないためには位相遅れ ϕ が 0 でなければならない．w_j が実数のとき，ϕ が 0 になるためには w_j が偶関数

$$w_{-j} = w_j \tag{13.6}$$

であればよい．w_j が偶関数であれば

$$W(\omega) = \sum_{j=-\infty}^{\infty} w_j e^{ij\omega} = w_0 + 2\sum_{j=1}^{\infty} w_j \cos j\omega$$

であるから，たしかに ϕ は 0 になる (上式でも $W(\omega) < 0$ になることがあるが，単に符号の変化だけのときには零位相とする)．

ここで思いだすのは §9 で導いたデータウィンドウである．そこでの w_j はすべて (13.6) 式を満たしている．ただしそれらは (13.2) 式を満たしてはいないが，この式を満たすようにスケーリングするのは容易である．なお，(13.2) 式が満たされているとき

$$W(0) = 1$$

が成り立つから，この関係を用いてもスケーリングすることもできる．

Fig. 9.5 に示したウィンドウはすべて $|\omega| > 2\pi/M$ で $W_M(\omega)$ が実質的に 0 になっている．これはいいかえれば，周期がサンプル間隔を単位として M よりも短い成分が除かれるということを意味している．したがって，これらのデータウィンドウは移動平均の重みとして用いることができる．

線型位相　対称な重み w_j $(|j| \le M)$ の時刻をずらして

$$h_j = w_{j-M} \qquad j = 0, 1, 2, \cdots, 2M$$

を作ったとする．h_j による移動平均の結果は w_j による移動平均の結果と本質的には同じで，ただ時間軸を M だけずらしたものである．h_j の周波数応答を計算すると

$$\begin{aligned} H(\omega) &= \sum_{j=0}^{2M} h_j e^{ij\omega} = \sum_{j=0}^{2M} w_{j-M} e^{ij\omega} \\ &= \sum_{k=-M}^{M} w_k e^{i(k+M)\omega} = W(\omega) e^{iM\omega} \end{aligned}$$

となる．w_j が零位相だとすると h_j の位相遅れ時間は (13.5) 式から

$$\tau_p(\omega) = M$$

となり，ω によらない定数になる．

この逆も正しい．周波数応答の位相 $\phi(\omega)$ が ω に比例するときには，$\tau_p(\omega)$ は ω によらない定数になる．このときにはすべての周波数成分が同じ時間だけ遅れるから，出力波形 y_j は零位相の出力波形の時間軸をずらしたものになっている．したがって，零位相と線型位相は本質的に同じものである．

長周期成分の除去　これまでとは逆に，ドリフトなどの長周期成分を除くには，一旦短周期成分を除いたものを原記録から差し引けばよい．すなわち

$$y_j = x_j - \sum_k w_k x_{j-k}$$

を計算すればよい．これは形の上では重み

$$v_j = \delta_j - w_j = \begin{cases} 1 - w_j & j = 0 \\ -w_j & j \ne 0 \end{cases} \tag{13.7}$$

を用いた移動平均

$$y_j = \sum_j v_k x_{j-k}$$

にほかならない．なお，v_j の周波数応答は $W(\omega)$ を用いて

$$V(\omega) = 1 - W(\omega)$$

と表される．

一般のフィルター　一般のフィルターへの入力 x_j と出力 y_j との関係を (13.1) 式と同様に改めて

$$y_j = \sum_{k=-\infty}^{\infty} h_k x_{j-k} \tag{13.8}$$

と書くことにする．入力がインパルス δ_j のときの出力は

$$y_j = \sum_k h_k \delta_{j-k} = h_j$$

となるから，h_j はフィルターの**インパルス応答**と呼ばれる．入力が有界のときに出力が有界になるためには

$$\sum_{j=-\infty}^{\infty} |h_j| < \infty \tag{13.9}$$

であれば十分である．

移動平均と同様に h_j の離散的フーリエ変換を

$$H(\omega) = \sum_{j=-\infty}^{\infty} h_j e^{ij\omega} \tag{13.10}$$

とすれば，これは入力が単振動のときの，入力と出力の比，すなわち**周波数応答**である．振幅応答，位相遅れも同様に定義できる．(13.10) 式の逆変換は

$$h_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(\omega) e^{-ij\omega} d\omega \tag{13.11}$$

である．実は (13.10) 式の $H(\omega)$ が存在するためには (13.9) 式の条件は厳しすぎる．二乗積分の意味で $H(\omega)$ が存在するためには

$$\sum_j |h_j|^2 < \infty$$

で十分である (§9)．この条件が満たされていれば，よほど意地のわるい入力でなければ出力も有界になる．

因果的フィルター 地震計のような物理系の応答を模したフィルターは，入力が入る前には出力は現れるはずはないから，インパルス応答は

$$h_j = 0 \quad j < 0 \tag{13.12}$$

でなければならない．このようなフィルターを因果的 (causal)，あるいは物理的に実現可能 (physically realizable) なフィルターという．因果的なフィルターは零位相ではない．

群遅延時間 入力 x_j がスペクトルをもっていたとして，角周波数 $\omega - \Delta\omega/2$ から $\omega + \Delta\omega/2$ の範囲の成分だけを取り出す．これは x_j の逆変換 (13.11) 式から

$$x_j(\omega) = \frac{1}{2\pi}\int_{\omega-\Delta\omega/2}^{\omega+\Delta\omega/2} |X(\sigma)|e^{-i(j\sigma-\theta)}d\sigma$$

と書くことができる．$X(\sigma)$ は x_j のスペクトル，$\theta(\sigma)$ は x_j の位相スペクトルである．この範囲ではスペクトルはあまり変化しないとして

$$|X(\sigma)| \doteq |X(\omega)|$$
$$\theta(\sigma) \doteq \theta(\omega) + \theta'(\omega)(\sigma - \omega)$$

で近似する．$\theta'(\omega) = d\theta(\omega)/d\omega$ である．これを先の積分に代入して積分すると

$$x_j(\omega) \doteq \frac{\Delta\omega}{2\pi}X(\omega)e^{-ij\omega} \cdot \frac{\sin(j-\theta')\Delta\omega/2}{(j-\theta')\Delta\omega/2}$$

が得られる．最後の因数は sinc 関数にほかならない．

この式から x_j の角周波数 ω の成分は，時刻

$$t_g(\omega) = \frac{d\theta(\omega)}{d\omega} \tag{13.13}$$

に振幅が最大になることがわかる．この時刻を群到着時刻 (group arrival time) と呼ぶことがある．同様なことを出力 y_j (13.8) 式について行う．$H(\omega)$ の位相を $\phi(\omega)$ とすると，y_j の位相は $\theta + \phi$ になる．これは x_j に比べて $\phi(\omega)$ だけ遅れているが，これを時間に直したものが (13.5) 式の $\tau_p(\omega)$ である．一方，$y_j(\omega)$ の振幅が最大になる時刻は $j = \theta' + \phi'$ であるが，これは $x_j(\omega)$ に比べて

$$\tau_g(\omega) = \frac{d\phi(\omega)}{d\omega} \tag{13.14}$$

だけ遅れる．これが群遅延時間 (group delay time) である．零位相フィルターの場合には，位相遅れ，群遅れともに 0 である．

インパルス応答の打ち切り (13.1) 式ではインパルス応答を先に与えたが，周波数応答を与えてインパルス応答を求めることもできる．一例として，無次元角周波数 ω_c 以下の振動だけを通す，理想的なローパスフィルターを考える．この周波数応答を

$$H(\omega) = \begin{cases} 1 & |\omega| < \omega_c \\ 0 & |\omega| > \omega_c \end{cases} \tag{13.15}$$

とすれば，インパルス応答はフーリエ逆変換 (13.11) 式を用いて

$$h_j = \frac{1}{2\pi}\int_{-\omega_c}^{\omega_c} e^{-ij\omega}d\omega = \frac{\omega_c}{\pi}\frac{\sin j\omega_c}{j\omega_c} \tag{13.16}$$

となる．

上のインパルス応答は無限に続き，しかも振幅は $1/j$ でしか減衰しない．実際に $\pm\infty$ までの和をとることができないから，(13.8) 式を有限項の和

$$y_j = \sum_{k=-M}^{M} h_k x_{j-k}$$

で近似したとすれば，このフィルターの実際の周波数応答は $H(\omega)$ ではなく

$$H_M(\omega) = \sum_{j=-M}^{M} h_j e^{ij\omega}$$

である．これは離散的フーリエ変換を求めるときのデータの打ち切りと同じ形をしている．

Fig. 13.1 は一例である．サンプル間隔 $\Delta t = 0.01$s のとき，f_c =10Hz 以上の波を完全に遮断するフィルターのインパルス応答を (13.16) 式で計算した．ここに，$\omega_c = 2\pi f_c \Delta t = 0.2\pi$ である．図の上段にインパルス応答 h_j を $j = 0$ から $j = 20$ まで示してある．h_j は偶関数であるから $j \geq 0$ だけを示してある．このインパルス応答を $|j| = 20$ で打ち切ったときの周波数応答は下段に実線で示してある．この応答は最初に仮定した応答 (13.15) 式 (細い破線) とはかけ離れている．

そこでスペクトルを求めるときのように，インパルス応答 h_j にデータウィンドウ w_j を掛けた $w_j h_j$ をインパルス応答として用いる．w_j として三角形ウィンドウ (9.31) 式を用いたときの応答が破線で示してある．この応答は約 6Hz まで平坦で，それ以

降単調に減衰している．これは山や谷がある単純な打ち切りに比べて，ローパスフィルターとしてははるかに望ましい性質である．ここでの結論は，理想的なインパルス応答をそのまま用いるよりも，ウィンドウを掛けた方がよいということである．

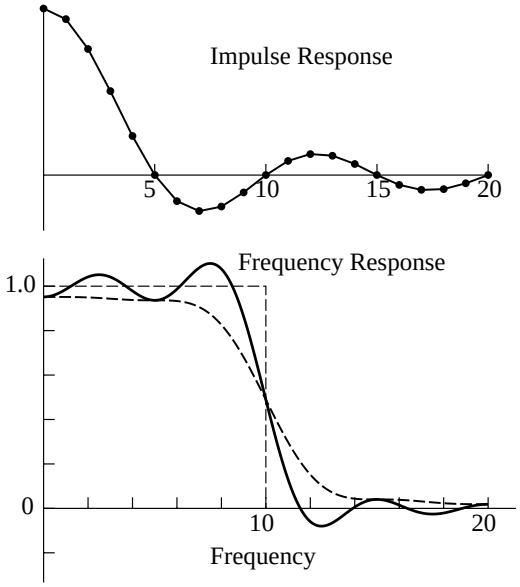


Fig. 13.1 ローパスフィルターの例．上段：インパルス応答．下段：周波数応答．実線は単純な打ち切り，破線は三角形ウィンドウを掛けたときの応答.

遅延演算子と z 変換　時系列 x_j の添字を一つ戻す，すなわち時間を一単位だけ前に戻す演算子 z を

$$zx_j = x_{j-1}$$

で定義し，これを単位の遅延演算子と呼ぶ．一般に

$$z^n x_j = x_{j-n} \quad (n \text{ は整数}) \tag{13.17}$$

と定義する．

(13.9) 式を満たすフィルターに対して

$$H(z) = \sum_{k=-\infty}^{\infty} h_k z^k \tag{13.18}$$

を，このフィルターの z 変換という．これは演算子である．$H(z)$ を x_j に作用させると，(13.17) 式を用いて

$$H(z)x_j = \sum_k h_k z^k x_j = \sum_k h_k x_{j-k}$$

となる．これは (13.8) 式にほかならない．いいかえればフィルターの出力は z 変換を用いて

$$y_j = H(z)x_j \tag{13.19}$$

と書くことができる．

すでに気が付いたように，$H(z)$ の z に

$$z = e^{i\omega} \tag{13.20}$$

を代入すればフィルターの周波数応答 $H(\omega)$ が得られる．すなわち

$$H(\omega) = H\big(z = e^{i\omega}\big) \tag{13.21}$$

が成り立つ．右辺は z 変換，左辺は周波数応答を同じ記号 H で表しているが，慣れれば混乱は生じない．上式は z を複素変数と考えたときに，z 平面上の単位円上における $H(z)$ の値が周波数応答であることを意味している．

周波数応答のフーリエ逆変換がインパルス応答である．これに対応して，z 変換の逆変換は

$$h_j = \frac{1}{2\pi i}\oint_{|z|=1} H(z) z^{-j}\frac{dz}{z} \tag{13.22}$$

で表される．この積分の積分路は複素 z 平面上の単位円上の反時計回りの周回路である．これが正しいことは (13.18) 式を代入することによって形式的に確かめることができる．ただし $H(z)$ が単位円上に特異点を持つような場合には，この積分には特別の考慮をしなければならない．

積と畳み込み　二つの z 変換 $A(z)$ と $B(z)$ の積

$$C(z) = A(z)B(z) \tag{13.23}$$

も z 変換である．$A(z)$ の係数を a_j，$B(z)$ の係数を b_j とすると，$C(z)$ に対応する係数 c_j は

$$\begin{aligned} C(z) &= \sum_j a_j z^j \sum_k b_k z^k = \sum_j \sum_k a_j b_k z^{j+k} \\ &= \sum_n \Big(\sum_j a_j b_{n-j}\Big) z^n \end{aligned}$$

より

$$c_n = \sum_j a_j b_{n-j} = \sum_k a_{n-k} b_k \tag{13.24}$$

と表される．これは a_j と b_j の離散的畳み込みである．

減衰振動の z 変換 一例として，減衰振動

$$x_n = \cos n\theta e^{-\alpha n} \quad n \geq 0, \quad \alpha > 0 \tag{13.25}$$

の z 変換を計算してみると

$$\begin{aligned} X(z) &= \frac{1}{2}\sum_{n=0}^{\infty}\left(e^{(i\theta-\alpha)n} + e^{-(i\theta+\alpha)n}\right) z^n \\ &= \frac{1}{2}\left(\frac{1}{1-e^{i\theta-\alpha}z} + \frac{1}{1-e^{-i\theta-\alpha}z}\right) \\ &= \frac{1-e^{-\alpha}\cos\theta z}{1-2e^{-\alpha}\cos\theta z + e^{-2\alpha}z^2} \end{aligned} \tag{13.26}$$

となる．このように，z 変換は z の冪級数だけでなく，有理式として表すこともできる．

最小位相 二つの z 変換

$$A_1(z) = z - \alpha \qquad A_2(z) = 1 - \overline{\alpha}z \tag{13.27}$$

を考える．$\overline{\alpha}$ は α の複素共役である．簡単な計算でこれらの振幅スペクトルが

$$|A_1(\omega)|^2 = |A_2(\omega)|^2 = 1 + |\alpha|^2 - 2\mathrm{Re}(\alpha e^{-i\omega})$$

と，同じであることがわかる．もちろん位相は異なっている．

そこで位相のことを調べるためにスペクトルを振幅と位相に分けて

$$A_1(\omega) = e^{i\omega} - \alpha = |A_1(\omega)|e^{i\phi_1(\omega)}$$

とする．対数微分をとると

$$\frac{1}{A_1(\omega)}\frac{dA_1}{d\omega} = \frac{d}{d\omega}\ln|A_1(\omega)| + i\frac{d\phi_1(\omega)}{d\omega}$$

であるから，左辺を計算して虚数部をとればそれが $d\phi_1/d\omega$，すなわち $A_1(z)$ の群遅延時間が求められることになる．計算を実行すると

$$\begin{aligned} \frac{1}{A_1}\frac{dA_1}{d\omega} &= \frac{i(1-\overline{\alpha}e^{i\omega})}{|A_1(\omega)|^2} \\ \frac{1}{A_2}\frac{dA_2}{d\omega} &= \frac{i(|\alpha|^2-\overline{\alpha}e^{i\omega})}{|A_2(\omega)|^2} \end{aligned}$$

となる．先に示したように $|A_1(\omega)| = |A_2(\omega)|$ であるから

$$\frac{d\phi_1(\omega)}{d\omega} - \frac{d\phi_2(\omega)}{d\omega} = \frac{1-|\alpha|^2}{|A_1(\omega)|^2}$$

が成り立つ．

α は $A_1(z)$ の零点である．いま仮に $|\alpha| > 1$，すなわち $A_1(z)$ の零点が単位円の外部にあったとすれば

$$\frac{d\phi_1(\omega)}{d\omega} < \frac{d\phi_2(\omega)}{d\omega} \qquad |\alpha| > 1 \tag{13.28}$$

であるから，すべての周波数に対して $A_1(z)$ の群遅延時間の方が小さい．

α が実数のときには $A_1(0) = A_2(0) = 1-\alpha$ であるから

$$\phi_1(0) = \phi_2(0)$$

と選ぶことができる．これを初期条件として (13.28) 式を積分すれば

$$\phi_1(\omega) < \phi_2(\omega) \quad 0 < \omega < \pi \tag{13.29}$$

が成り立つ．したがって $A_1(z)$ は群遅延が小さいだけでなく，位相遅れも小さい．

この議論は更に一般化することができる．m 次の z 変換

$$A(z) = a_0 + a_1 z + \cdots + a_m z^m \tag{13.30}$$

には重根も含めて m 個の零点がある．これらの零点を α_j とすると $A(z)$ は

$$A(z) = a_m(z-\alpha_1)(z-\alpha_2)\cdots(z-\alpha_m)$$

と書くことができる．$A(z)$ のスペクトルは各因数 $z-\alpha_j$ のスペクトルの積であるから，ある因数 $z-\alpha_k$ を (13.27) 式の規則に従って置きかえる，すなわち $1-\overline{\alpha}_k z$ で置きかえても振幅スペクトルは変化しない．置きかえの組み合わせは重根も含めると 2^m 組ある．この中に零点がすべて単位円の外側にしか存在しない多項式，いいかえれば

$$A(z) \neq 0 \quad |z| \leq 1 \tag{13.31}$$

を満たす z 変換が存在する．これは上の議論によって同じ振幅スペクトルをもつ m 次の z 変換の中で群遅延時間が最も小さいものである．このような z 変換を最小位相という．$A(z)$ の係数 $a_0 \sim a_m$ を波形 (wavelet) と考えたとき，この波形の群到着時刻は同じスペクトルをもつ波形の中で最も早い．このような波形を最小位相波形という．逆に全ての零点が単位円の内部にあるような z 変換を最大位相と呼ぶ．これら以外の波形を混合位相という．

実数だけで考えると同じ振幅スペクトルをもつ波形の数はずっと少なくなる．(13.30) 式の係数 a_j が

すべて実数のとき，α_k が複素根なら $\overline{\alpha}_k$ も複素根である．実数だけで考えるときには，根の入れかえを行うときにこの二つをペアで入れかえなければならない．たとえば

$$\begin{aligned} A_1(z) &= (z-\alpha)(z-\overline{\alpha}) \\ &= z^2 - 2\mathrm{Re}\,\alpha\, z + |\alpha|^2 \\ A_2(z) &= (1-\overline{\alpha}z)(1-\alpha z) \\ &= 1 - 2\mathrm{Re}\,\alpha\, z + |\alpha|^2 z^2 \end{aligned}$$

は振幅スペクトルの等しい z 変換である．$|\alpha| > 1$ とすれば $A_1(z)$ は最小位相，$A_2(z)$ は最大位相である．スペクトルで見ると明らかに $A_1(0) = A_2(0)$ であるから，先と同じようにして

$$\phi_1(\omega) < \phi_2(\omega)$$

が成り立つ．すなわち，実係数で考えたときには最小位相は実際に位相が最小になっている．

Fig. 13.2 に示した例は 7 次 ($m = 7$) の波形である．この波形の零点は 3 実根，4 複素根であるが，実係数であるから，複素根は二組の複素共役根になっている．したがって実質的に独立な零点は 5 組であるから，独立な波形は $2^5 = 32$ 個ある．Fig. 13.2 にはそのうちの 16 個の波形が示してある．残りの 16 個はこれらの波形の時間軸を反転したものである．

Fig. 13.2 の左上の 0 番の波形が最小位相である．この波形では先頭部にエネルギーの集中が見られる．8 番の波形もエネルギーが先頭部にあるように見えるが，仔細に見れば後部の振幅が大きくなっていることがわかる．しかし一つの波形だけを見て，それが最小位相かどうかを判断することは難しい．

最小位相の零点を全て単位円の内部にもってきたものが最大位相である．これは作り方から明らかなように，最小位相の時間軸を反転させた波形になっている．

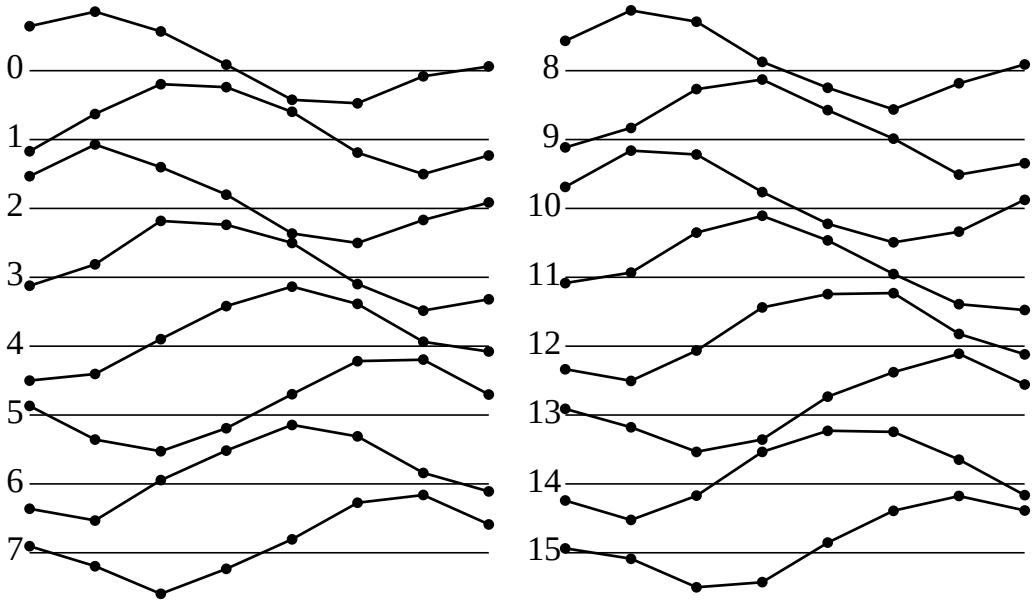


Fig. 13.2 同じ振幅スペクトルをもつ 7 次の波形．左上 (0 番) が最小位相．

逆フィルター　地震計の応答の z 変換を $B(z)$ とする．地震計は物理系であるから，$B(z)$ は z の 0 および正の冪で展開される (因果的)．ここでは $B(z)$ は n 次多項式とする．

地震計の出力を x_j とする．これは地動を u_j とすると

$$x_j = B(z)u_j$$

が成り立つ．観測の目的は u_j を求めることである．x_j から u_j を求めるフィルターが逆フィルターである．

$$u_j = \frac{1}{B(z)}x_j \tag{13.32}$$

$1/B(z)$ に対応するインパルス応答を求めるには，逆変換 (13.22) 式を用いればよいが，この計算は容易ではない．ここではもっとわかりやすい方法を用いる．

$B(z)$ は n 次の多項式であるから n 個の零点をもっている．そこで

$$B(z) = b_n(z-\beta_1)(z-\beta_2)\cdots(z-\beta_n)$$

と書く．あまり本質的ではないが，零点に重根はないと仮定する．そうすると部分分数の定理により $1/B(z)$ は

$$\frac{1}{B(z)} = \sum_{k=1}^{n} \frac{B_k}{z-\beta_k} \tag{13.33}$$

のように展開することができる．B_k は定数で，この展開は一義的である．いま，ある零点 β_j の絶対値が 1 よりも大きかったとすると，これに対応する部分分数は

$$\begin{aligned}\frac{1}{z-\beta_j} &= -\frac{1}{\beta_j}\frac{1}{1-\beta_j^{-1}z} \\ &= -\frac{1}{\beta_j}\Big(1+\beta_j^{-1}z+\beta_j^{-2}z^2+\cdots\Big) \qquad |\beta_j|>1\end{aligned}$$

となって，この展開は収束する (z^n の係数が $n\to+\infty$ で発散しない)．一方，$|\beta_k|<1$ の零点に対しては

$$\begin{aligned}\frac{1}{z-\beta_k} &= \frac{1}{z}\frac{1}{1-\beta_k z^{-1}} \\ &= z^{-1}+\beta_k z^{-2}+\beta_k^2 z^{-3}+\cdots \qquad |\beta_k|<1\end{aligned}$$

となってこの展開も収束する．以上をまとめれば，部分分数展開 (13.33) 式の各項を上の規則に従って展開して和を作れば

$$\frac{1}{B(z)} = \sum_{j=-\infty}^{\infty} h_j z^j$$

の係数 h_j が求められることになる．これが逆フィルターのインパルス応答にほかならない．

h_j の作り方から，単位円外の零点からインパルス応答の 0 および正の部分が生じ，単位円内の零点から負の部分が生じることがわかる．ところで逆フィルターに負の部分 h_j $(j<0)$ があるということは，観測という立場から見ると望ましいことではない．

j を現在の時刻とする．現在の観測値 x_j には現在の地動 u_j，および過去の地動 $u_{j-1}, u_{j-2}, \cdots$ の影響が含まれている．したがって，現在の地動 u_j を求めるために現在および過去の観測値 $x_j, x_{j-1}, x_{j-2}, \cdots$ を利用するというのが最も自然な考え方である．しかし単位円内に零点があると $k<0$ の h_k が生じるから，u_j を求めるために現在，過去の観測値のほかに，未来の観測値も必要になる．このようなことが起きるのは，$B(z)$ に対応するインパルス応答の後ろの部分に振幅の大きいところがあり，現在の u_j の影響が将来の x_k $(k>j)$ に強く現れるからである．このような測定系が望ましいものではないのはいうまでもない．

以上をまとめると，測定系 $B(z)$ の逆が因果的であるためには $B(z)$ は最小位相でなければならない．$B(z)$ が単に因果的であっても逆が因果的であるとは限らない．

漸化式 (再帰的) フィルター　$B(z)$ が最小位相のときには逆 $1/B(z)$ のインパルス応答を求めなくてもフィルタリングの操作を行うことができる．測定系への入力 u_j が $j<0$ で 0 とする．当然，出力 x_j も $j<0$ で 0 になる．入出力の関係 $x_j = B(z)u_j$ を書きかえて

$$b_0 u_j + \sum_{k=1}^{n} b_k u_{j-k} = x_j \qquad j \geq 0$$

とし，今度は x_j が与えられた入力，u_j が求めたい出力と読みかえる．$B(z)$ の零点がすべて単位円の外にあるときには，上式を

$$\begin{aligned}u_j = \frac{1}{b_0}\Big(x_j - \sum_{k=1}^{n} b_k u_{j-k}\Big) \\ j = 0, 1, 2, \cdots\end{aligned} \tag{13.34}$$

のように正の向きに計算しても安定である．このことは §8 で示しておいたが，次のように考えてもよい．$B(z)$ が因果的であるとき，x_j は現在および過去の u_j から決まる．さらに $B(z)$ が最小位相であれば u_j は現在および過去の u_j で表される．上式は現在の x_j と過去の u_j で書き表されているからこのことと矛盾しない．逆に，$B(z)$ が最小位相でないときには u_j は現在，過去および未来の u_j によって書き表されるはずであるが，上式には未来の u_j が含まれていないので，この式から計算した u_j は正しい値ではない ($j\to\infty$ で発散してしまう)．

(13.34) 式はさらに一般化することができる．いま，二つの z 変換

$$A(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_m z^m$$
$$B(z) = 1 + b_1 z + b_2 z^2 + \cdots + b_n z^n$$

が与えられたとして，フィルター

$$y_j = \frac{A(z)}{B(z)} x_j \tag{13.35}$$

を考える．比だけが問題であるから $b_0 = 1$ としてある．$B(z)$ が最小位相であるとすれば，(13.34) 式と同様に

$$y_j = \sum_{k=0}^{m} a_k x_{j-k} - \sum_{k=1}^{n} b_k y_{j-k} \qquad j = 0, 1, 2, \cdots \tag{13.36}$$

によって計算しても正しい結果が得られる．この計算は $A(z)/B(z)$ のインパルス応答を求めてから入力 x_j との畳み込みを計算するよりもはるかに高速である．

簡単な例

$$y_j = \frac{1}{1 - z/2}\, x_j$$

を考えてみる．このフィルターのインパルス応答は

$$h_k = 2^{-k} \qquad k \geq 0$$

であるから，24 ビットの精度で畳み込みを計算するためには一つの出力当たり少なくとも 24 回の積和が必要になる．しかし

$$y_j = x_j + 0.5 y_{j-1}$$

で計算すれば一つの出力当たり一回の積和で済む．

(13.36) 式は過去の出力が右辺であたかも入力のように扱われている．これはフィードバック回路と同じである．

バターワース・フィルター　ここではアナログフィルターを考える．アナログの角周波数を σ とし，振幅応答が

$$|H(\sigma)|^2 = \frac{1}{1 + (\sigma/\sigma_c)^{2n}} \tag{13.37}$$

で与えられるフィルターを n 次のバターワースのフィルターという．n は正の整数である．このフィルターの振幅応答の二乗は $\sigma = 0$ で 1，$\sigma = \pm\sigma_c$ で 1/2(半値点)，$\sigma \to \pm\infty$ で 0 になるから，$H(\sigma)$ はローパスフィルターである．n が大きくなればなるほど $\sigma = \pm\sigma_c$ における減衰の勾配が急になり，n が無限大では

$$|H(\sigma)|^2 \longrightarrow \begin{cases} 1 & |\sigma| < \sigma_c \\ 0 & |\sigma| > \sigma_c \end{cases} \qquad n \longrightarrow \infty$$

となって，理想的なローパスフィルターになる．

(13.37) 式では振幅応答だけが定義されていて，位相についてはわからない．そこで $n = 1$ の簡単な場合について考えてみる．このときには (13.37) 式の分母は簡単に因数分解できて

$$|H(\sigma)|^2 = \frac{1}{[1 - i(\sigma/\sigma_c)][1 + i(\sigma/\sigma_c)]}$$

となる．したがって $H(\sigma)$ は

$$H(\sigma) = \frac{1}{1 - i(\sigma/\sigma_c)} \quad \text{または} \quad \frac{1}{1 + i(\sigma/\sigma_c)}$$

のどちらかである．$H(\sigma)$ が決まればそれをフーリエ逆変換

$$h(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\sigma) e^{-i\sigma t} d\sigma$$

することによってインパルス応答 $h(t)$ が得られる．

そこでためしに前者

$$H(\sigma) = \frac{1}{1 - i(\sigma/\sigma_c)}$$

を選んでみる．このフーリエ変換は複素 σ 平面の下半平面の $\sigma = -i\sigma_c$ に極をもっている．フーリエ逆変換は実軸上の積分であるが，これに半円 R^+ または R^- を付け加えた周回積分を用いて計算するのが常道である．まず $t < 0$ のときには，半円の半径を無限大にしたとき指数部 $e^{-i\sigma t}$ が 0 に収束するためには，半円として上半平面上の R^+ を選ばなけれ

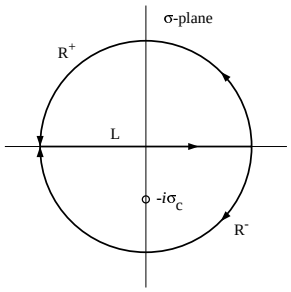


Fig. 13.3 フーリエ逆変換の積分路

ばならない．上半平面には $H(\sigma)$ の極はないから，コーシーの定理から

$$\oint_{\mathrm{L+R^+}} = 0$$

が成り立つ．半円の半径を無限大にした極限では半円上の積分は 0 になる．したがって実軸上の積分も 0 になる．これは

$$h(t) = 0 \quad t < 0$$

を意味している．一方，$t > 0$ のときには下半平面上の半円 R^- を用いた周回積分を行わなければならない．このときには積分路の内部に極があるから

$$\oint_{\mathrm{L+R^-}} = -2\pi i\,\mathrm{Res}$$

となる．Res は $H(\sigma)$ の留数である．負号は積分路が時計回りであるからである．留数を計算すると

$$\mathrm{Res} = i\sigma_c$$

であるから，$t > 0$ のときの逆変換

$$h(t) = \sigma_c e^{-\sigma_c t} \quad t > 0$$

が得られた．すなわちこの選び方では因果的なフィルターが得られる．

これとは反対に

$$H(\sigma) = \frac{1}{1 + i(\sigma/\sigma_c)}$$

を選んだとすると，今度はは極が上半平面にくるから

$$h(t) = 0 \quad t > 0$$

となってしまう．この選び方だと未来の入力を知らないと現在の出力がわからないということになってしまう．とちらの選び方が物理的であるかは明らかである．

一般の n のときには $|H(\sigma)|^2$ の分母は $2n$ 個の零点をもっている．具体的には

$$\sigma/\sigma_c = \exp\Big(\frac{2k+1}{2n} i\pi\Big) \qquad (13.38)$$
$$k = 0, 1, 2, \cdots, 2n-1$$

で与えられる．これらの零点は複素 z 平面上の半径 σ_c の円周上に等間隔に並んでおり，そのうちの n 個は上半平面上にあり，n 個が下半平面上にある．先の議論から明らかなように，$H(\sigma)$ が因果的，すなわち $t < 0$ で $h(t) = 0$ になるためには $H(\sigma)$ の極が上半平面上にあってはならない．そこで上の零点のうち下半平面上にあるものだけを集めて $H(\sigma)$ を作る．

$$H(\sigma) = \prod_{k=1}^{n} \Big[e^{i\theta_k} - i(\sigma/\sigma_c) \Big]^{-1} \qquad (13.39)$$
$$\theta_k = \frac{2k-n-1}{2n}\pi$$

後の計算の便宜上，分母に i を掛けてある．これが因果的であり，しかも (13.37) 式を満たしていることはその作られ方から明らかである．なおここに現れた $e^{i\theta_k}$ は単位円の右半分上の点である．

双一次変換　(13.39) 式はアナログフィルターの周波数応答であるから，そのままディジタルフィルターに用いることはできない．フーリエ逆変換からインパルス応答 $h(t)$ を求め，これを間隔 Δt でサンプルすれば一応ディジタルフィルターのインパルス応答になるが，このディジタルフィルターの周波数応答は，エイリアシングのために元の周波数応答 $H(\sigma)$ とは違ってしまう．しかしちょっとした工夫で，これをディジタルフィルターに変換することができる．

変換

$$\sigma = \frac{2}{i\Delta t}\,\frac{z-1}{z+1} \qquad (13.40)$$

は複素 z 平面上の単位円の内部を複素 σ 平面の上半平面に，単位円の外部を下半平面に写像する．Δt はサンプル間隔である．たとえば z 平面の原点 $z = 0$ は σ 平面の点 $\sigma = 2i/\Delta t$ に，$z = \infty$ は $\sigma = -2i/\Delta t$ に変換される．特に z 平面の単位円

$$z = e^{i\omega} \quad |\omega| < \pi$$

は σ 平面の実軸

$$\sigma = \frac{2}{\Delta t}\tan\frac{\omega}{2} \qquad (13.41)$$

に変換される．ω が小さいところでは右辺を展開すると

$$\sigma \doteq \frac{\omega}{\Delta t}$$

であるから，アナログの無次元角周波数 $\sigma\Delta t$ とディジタルの無次元角周波数 ω が近似的に等しいことがわかる．

そこでこの変換 (13.40) 式を (13.39) 式に代入する．結果は z の有理式となるから，これはディジタルフィルターの z 変換と考えることができる．このフィルターは因果的である．なぜなら，(13.39) 式の下半平面上の極は，z 平面上の単位円の外部に写像されるからである．その周波数応答は，アナログフィルター $H(\sigma)$ の σ 軸を $-\pi < \omega < \pi$ に圧縮したものになっており，$\sigma = 0$ 付近の応答の歪は少ない．したがって得られた z 変換もローパスフィルターである

(13.40) 式を (13.39) 式に代入するとき，θ_k が

$$\theta_{n-k+1} = -\theta_k$$

を満たしているから，k と $n-k+1$ の項は互いに複素共役になっている．これらのペアを一組として計算を行えば実係数の z 変換が次のように得られる．

$$\begin{aligned} H(z) &= \prod_{k=1}^{M} \frac{(1+z)^2}{b_0^{(k)} + b_1^{(k)} z + b_2^{(k)} z^2} \qquad (13.42)\\ b_0^{(k)} &= 1 + \frac{4}{\sigma_c \Delta t}\cos\theta_k + \Big(\frac{2}{\sigma_c \Delta t}\Big)^2 \\ b_1^{(k)} &= 2\Big[1 - \Big(\frac{2}{\sigma_c \Delta t}\Big)^2\Big] \\ b_2^{(k)} &= 1 - \frac{4}{\sigma_c \Delta t}\cos\theta_k + \Big(\frac{2}{\sigma_c \Delta t}\Big)^2 \\ M &= \begin{cases} n/2 & n \text{ が偶数} \\ (n+1)/2 & n \text{ が奇数} \end{cases} \end{aligned}$$

ただし，n が奇数のときは最後の $k = M$ の項は

$$\begin{aligned} &\frac{1+z}{b_0^{(M)} + b_1^{(M)} z} \qquad (13.43)\\ &b_0^{(M)} = 1 + \frac{2}{\sigma_c \Delta t} \\ &b_1^{(M)} = -\frac{2}{\sigma_c \Delta t} \end{aligned}$$

で置きかえなければならない．

出力の計算法 (13.42) 式の分子，分母の積を実行すると，$H(z)$ は z の n 次式の比になるから，入力が x_j のときの出力 y_j は (13.36) 式によって計算することができる．しかしせっかく因数分解できているものを掛け算してしまうと係数に丸め誤差が入ってしまうので，できるなら (13.42) 式そのままを用いて出力の計算を行ないたい．

そこで (13.42) 式を改めて

$$\begin{aligned} H(z) &= G \prod_{k=1}^{M} H_k(z) \qquad (13.44)\\ H_k(z) &= \frac{1 + a_1^{(k)} z + a_2^{(k)} z^2}{1 + b_1^{(k)} z + b_2^{(k)} z^2} \end{aligned}$$

と書きかえる．分子分母の定数項を 1 にしたのは少しでも計算量を少なくするためである．出力は次のように計算できる．

$$\begin{aligned} y_j^{(1)} &= H_1(z) G x_j \\ y_j^{(2)} &= H_2(z) y_j^{(1)} \\ &\cdots\cdots \\ y_j^{(M)} &= H_M(z) y_j^{(M-1)} \end{aligned}$$

最後の出力 $y_j^{(M)}$ が求める y_j である．各ステップは

$$\begin{aligned} y_j^{(k)} &= y_j^{(k-1)} + a_1^{(k)} y_{j-1}^{(k-1)} + a_2^{(k)} y_{j-2}^{(k-1)} \\ &\quad - b_1^{(k)} y_{j-1}^{(k)} - b_2^{(k)} y_{j-2}^{(k)} \qquad (13.45)\\ &\qquad k = 1, 2, \cdots, M \end{aligned}$$

となる．ただし $y_j^{(0)} = G x_j$ である．

Fig. 13.4 は 6 次のバターワース・フィルターの応答である．パラメーターとしては Fig. 13.1 と同様に $\Delta t = 0.01$s，$f_c = 10$Hz に選んである．ただしここでの f_c は振幅応答が $1/\sqrt{2}$ になる周波数を意味している．係数 (13.42) 式の決定に必要な $\sigma_c \Delta t$ は (13.41) 式の関係を用いて

$$\sigma_c \Delta t = 2 \tan \frac{\omega_c}{2} \qquad \omega_c = 2\pi f_c \Delta t$$

から決めてある．このため振幅応答は $f = 10$Hz で正確に $|H(\omega)| = 1/\sqrt{2}$ になっている．振幅応答は $f = 8$Hz くらいまではほとんど 1 で，その後ゆるやかに減少している．$M = 3$ のフィルターでは一個の出力当たり 13 回の積和ので済むので，約 40 回の積和を必要とする Fig. 13.1 のフィルターよりも効率的であり，しかも性能がよい．

しかしバターワース・フィルターは零位相ではないので，出力に群遅延が生じる．Fig. 13.4 の下段に Δt を単位とした群遅延時間を示してある．群遅延時間は周波数 0 で約 6，それから緩やかに増加し，遮断周波数 f_c で最大の約 10 になる．単位はサンプル間隔 Δt である．

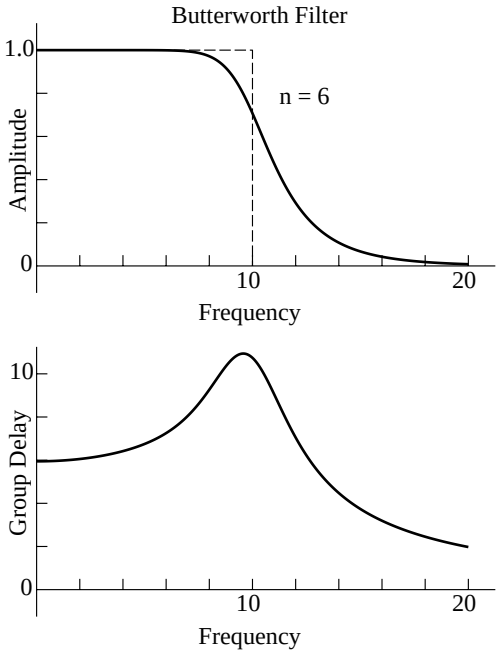


Fig. 13.4 バターワース・フィルターの周波数応答．上段：振幅応答，下段：群遅延時間．単位は Δt.

零位相化　一般に (13.35) 式で定義されるフィルターは零位相ではない．Fig. 13.4 で示したローパスフィルターでは最大 10 ポイントの群遅れが生じている．この遅れを無視していいかどうかは問題にもよるが，どうしても零位相にしたければ，同じフィルターを往復して掛ければよい．すなわち，フィルター $H(z)$ を正方向に掛けた出力 y_j の時間を反転したものに同じフィルターを掛けるのである．これは

$$z_j = \sum_{k=0}^{m} a_k y_{j+k} - \sum_{j=1}^{n} b_k z_{j+k} \qquad (13.46)$$
$$j = N, N-1, N-2, \cdots$$

という演算で表される．ただし，$j > N$ で $y_j = 0$ と仮定している．この演算は z 変換で書くと

$$z_j = H(z^{-1}) y_j = H(z^{-1}) H(z) x_j$$

で表されるから，同じフィルターを往復掛けたときの周波数応答は

$$H(-\omega) H(\omega) = |H(\omega)|^2$$

となる．すなわち位相は 0 になるが振幅応答は $|H(\omega)|$ ではなく，$|H(\omega)|^2$ になることに注意しなければならない．

参考文献

斎藤正徳 (1978)：漸化式ディジタル・フィルターの自動設計，物理探鉱，**31**, 240-263.

14 パワースペクトルの推定

パワー有限のデータ 離散的データの二乗和が有限なら，すなわち全エネルギーが有限なら，離散的フーリエ変換が存在する (§9)．しかし世の中には微動のように，始めも終わりもないようなデータがいくらもある．本節ではそのようなデータのスペクトル解析を行う．なお本節では，時系列を $x(t)$ のように連続変数 t の関数のように表示するが，ここでの t は時刻のインデックスで整数値のみをとる変数である．

あまり本質的ではないが，ここでは後の記述を簡単にするために平均が 0 の時系列のみを考える．ここでいう平均とは時間平均の意味で

$$E_t[x(t)] \equiv \lim_{T\to\infty} \frac{1}{2T+1} \sum_{t=-T}^{T} x(t) = 0 \qquad (14.1)$$

である．$E_t[\ ]$ は時間平均の演算を意味する．さらに以下で考えるデータ $x(t)$ は

$$E_t[|x(t)|^2] = \lim_{T\to\infty} \frac{1}{2T+1} \sum_{t=-T}^{T} |x(t)|^2 < \infty \qquad (14.2)$$

を満たしているものと仮定する．二乗の時間平均はパワーに相当するから，この条件を満たすデータをパワー有限のデータと呼ぶことにする．離散的なデータのときには，この条件は $x(t)$ が有界であることとほとんど同値である．エネルギー有限のデータも上の条件を満たしていることはいうまでもないが，この場合にはパワーが 0 であるから，パワースペクトル解析の対象とする意味がない．

> データに定常確率過程を仮定して，平均はアンサンブル平均をとるのが正統的なやり方である．しかし実際にデータの平均を計算する際には，さらにエルゴード性を仮定してアンサンブル平均を時間平均で置きかえるという手間をかけている．ここでは中間をとばして，時間平均だけで議論を進めることにする．

自己相関関数 $x(t)$ がパワー有限のとき

$$r_{xx}(\tau) = E_t[x(t+\tau)\overline{x(t)}] \qquad (14.3)$$

によって $x(t)$ の自己相関関数を定義する．引数 τ を自己相関関数のラグという．$r_{xx}(0)$ は (14.2) 式によって $x(t)$ のパワーにほかならない．自己相関関数は

$$r_{xx}(-\tau) = \overline{r_{xx}(\tau)} \qquad (14.4)$$

の対称性を満たしている．したがって $x(t)$ が実数なら $r_{xx}(\tau)$ は偶関数である．

a_k を任意に数列とすると次の恒等式が成り立つ．

$$E_t\Big[\Big|\sum_k a_k x(t+k)\Big|^2\Big] \geq 0 \qquad (14.5)$$

左辺を展開すると

$$\sum_{k,l} a_k E_t\Big[x(t+k)\overline{x(t+l)}\Big]\overline{a_l} = \sum_{k,l} a_k r_{xx}(k-l)\overline{a_l}$$

となるから，任意の数列 a_k に対して自己相関関数は

$$\sum_{k,l} a_k r_{xx}(k-l)\overline{a_l} \geq 0 \qquad (14.6)$$

を満足する．この条件を満たす $r_{xx}(\tau)$ を**正定値** (positive definite) であるという．上式には等号も含まれているので，厳密には非負定値 (non-negative definite) と呼ぶべきであるが，簡単に正定値と呼んでおく．

正定値性は自己相関関数の基本的な性質であり，逆に正定値性によって自己相関関数を定義することができる．$r_{xx}(\tau)$ が条件 (14.6) 式を満たしているとき，a_k として $a_0 = 1$，ある n に対して $a_n = \lambda$ でそれ以外は全て 0 という数列を選ぶと，(14.6) 式は

$$r_{xx}(0) + \lambda r_{xx}(n) + \overline{\lambda} r_{xx}(-n) + |\lambda|^2 r_{xx}(0) \geq 0$$

となる．左辺が実数になるためには

$$r_{xx}(-n) = \overline{r_{xx}(n)}$$

でなければならない．すなわち (14.4) 式が導かれた．ここではまだ λ が任意であるから，ある n に対して λ の偏角を $r_{xx}(n)$ の偏角の符号を反対にしたものを選ぶと

$$\lambda r_{xx}(n) = \mu |r_{xx}(n)|$$

とすることができる．μ は実数である．そうすると先の不等式は

$$r_{xx}(0) + 2\mu |r_{xx}(n)| + \mu^2 r_{xx}(0) \geq 0$$

となる．これが任意の μ に対して成り立つためには判別式から

$$|r_{xx}(n)| \le r_{xx}(0) \tag{14.7}$$

でなければならない．すなわち，自己相関関数の絶対値はラグが 0 のときが最大である．上式で等号が成り立つのは，もとの式 (14.5) に戻って考えれば，$x(t)$ が周期関数のときであることがわかる．

与えられたデータから実際に自己相関関数を計算するときにも，正定値性が保障されるような計算法をとらなければ，その後の計算で問題が生じる．$t=0$ から $t=N$ までのデータ $x(t)$ が与えられたとき，このデータの自己相関関数を

$$r_{xx}(\tau) \doteq \frac{1}{N-\tau+1}\sum_{t=0}^{N-\tau} x(t+\tau)\overline{x(t)} \quad \tau \ge 0$$

で推定すのはいかにももっともらしい．データの与えられていない $t<0$，$t>N$ については $x(t)=0$ として，積和の項数で平均しているからである．しかし上の $r_{xx}(\tau)$ は正定値ではない．もっと単純にラグ τ によらず常にデータ数 $N+1$ で割って

$$r_{xx}(\tau) \doteq \frac{1}{N+1}\sum_{t=0}^{N-\tau} x(t+\tau)\overline{x(t)} \quad \tau \ge 0 \tag{14.8}$$

とすれば，これは正定値になっている．

例　単振動の自己相関関数　$x(t)$ が単振動

$$s(t) = e^{-i\omega_0 t}$$

のときの自己相関関数は

$$r_{ss}(\tau) = E_t[e^{-i\omega_0(t+\tau)}e^{i\omega_0 t}] = e^{-i\omega_0\tau}E_t[1]$$

したがって

$$r_{ss}(\tau) = e^{-i\omega_0\tau} \tag{14.9}$$

となって同じ角周波数の単振動になる．

例　ホワイトノイズの自己相関関数　$\xi(t)$ が平均 0，分散 σ_ξ^2 のランダムな数列であるとする．すなわち

$$E_t[\xi(t)] = 0 \qquad E_t[|\xi(t)|^2] = \sigma_\xi^2 \tag{14.10}$$

さらに $\xi(t)$ は互いに無相関であると仮定する．これは

$$E_t[\xi(t+\tau)\overline{\xi(t)}] = 0 \quad \tau \ne 0 \tag{14.11}$$

という意味である．これらの仮定から $\xi(t)$ の自己相関関数は

$$r_{\xi\xi}(\tau) = \sigma_\xi^2\delta_\tau \tag{14.12}$$

と表される．δ_τ は離散的デルタ関数である．このようなランダムな数列をホワイトノイズという．ここでは平均と分散しか与えられておらず，分布関数には触れられていないので，同じホワイトノイズといっても，ガウス分布のホワイトノイズ，一様分布のホワイトノイズなどがある．

パワースペクトル　自己相関関数 $r_{xx}(\tau)$ の離散的フーリエ変換

$$R_{xx}(\omega) = \sum_{\tau=-\infty}^{\infty} r_{xx}(\tau)e^{i\omega\tau} \tag{14.13}$$

を $x(t)$ のパワースペクトルという．もちろん，これが存在するためには $r_{xx}(\tau)$ がエネルギー有限の関数でなければならない．逆変換は

$$r_{xx}(\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} R_{xx}(\omega)e^{-i\omega\tau}d\omega \tag{14.14}$$

であるが，ここで $\tau=0$ とすれば

$$r_{xx}(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} R(\omega)d\omega$$

となる．$r_{xx}(0)$ はパワーにほかならないから，上式は $R_{xx}(\omega)$ が単位周波数当たりのパワーという意味をもっていることがわかる．その意味で $R_{xx}(\omega)$ を $x(t)$ のパワースペクトル密度という．普通は単にパワースペクトルという．

パワーは正の量であるから，$R_{xx}(\omega)$ がパワースペクトルの密度という意味をもっているためには，$R_{xx}(\omega)$ は ω によらず正でなければならない．(14.6) 式の a_k として

$$a_k = e^{ik\omega} \qquad -N \le k \le N$$

を選ぶと，この不等式は

$$0 \le \sum_{k,l=-N}^{N} e^{ik\omega}r_{xx}(k-l)e^{-il\omega} = \sum_{k=-N}^{N}\sum_{n=k-N}^{k+N} r_{xx}(n)e^{in\omega}$$

となるが ($n = k - l$ と置いてある)，和の順序を入れかえると

$$= \sum_{n=-2N}^{2N} r_{xx}(n) e^{in\omega} \sum_{k} 1$$

となる．二重和の範囲は Fig. 14.1 に示す菱形の領域であるから，k についての和の範囲は

$$k = \begin{cases} -N \sim n+N & n \le 0 \\ n-N \sim N & n \ge 0 \end{cases}$$

である．したがって k についての和は

$$\sum_{k} 1 = 2N + 1 - |n|$$

になる．よって

$$\sum_{n=-2N}^{2N} \Big(1 - \frac{|n|}{2N+1}\Big) r_{xx}(n) e^{in\omega} \ge 0$$

が得られた．ここで N を無限大にすれば左辺は $R_{xx}(\omega)$ に収束するから (本当は証明が必要であるが)

$$R_{xx}(\omega) \ge 0 \tag{14.15}$$

が導かれた．すなわちパワースペクトルは周波数によらず 0 または正である．

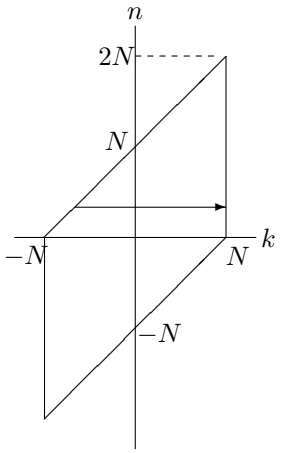


Fig. 14.1 k，n についての和の領域.

ホワイトノイズのパワースペクトル　ホワイトノイズの自己相関関数は (14.12) 式で与えられているので，そのパワースペクトルは

$$R_{\xi\xi}(\omega) = \sigma_\xi^2 \sum_{\tau} \delta_\tau e^{i\tau\omega} = \sigma_\xi^2 \tag{14.16}$$

となる．これは ω によらず一定である．ホワイト (白色) という名前はここからきている．

単振動のパワースペクトル　単振動の自己相関関数 (14.9) 式はエネルギー有限ではないから，通常の意味でのフーリエ変換は存在しない．しかしデルタ関数を許すことにすれば，形式的なパワースペクトルが得られる．

デルタ関数 $\delta(\omega - \omega_0)$ は

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \delta(\omega - \omega_0) e^{-i\omega\tau} d\omega = \frac{1}{2\pi} e^{-i\omega_0\tau} \tag{14.17}$$

を満たしている．したがって指数関数の形式的なフーリエ変換は

$$\frac{1}{2\pi} \sum_{\tau} e^{-i\omega_0\tau} e^{i\omega\tau} = \delta(\omega - \omega_0) \tag{14.18}$$

となるべきである．この和はもちろん収束しないが，(14.17) 式とのペアで意味のある関係である．この関係を用いると，単振動のパワースペクトルは

$$R_{ss}(\omega) = 2\pi\delta(\omega - \omega_0) \tag{14.19}$$

と表される．これは線スペクトルである．

相互相関関数　$x(t)$，$y(t)$ がともにパワー有限のデータであるとき

$$r_{xy}(\tau) = E_t[x(t+\tau)\overline{y(t)}] \tag{14.20}$$

によって $x(t)$ と $y(t)$ の間の相互相関関数を定義する．x と y の順序は重要である．これを反対にすると

$$E_t[y(t+\tau)\overline{x(t)}] = \overline{E_t[x(t-\tau)\overline{y(t)}]}$$

であるから

$$r_{yx}(\tau) = \overline{r_{xy}(-\tau)} \tag{14.21}$$

が成り立つ．

シュワルツの不等式　λ を任意の定数とするとき

$$E_t[|x(t+\tau) + \lambda y(t)|^2] \ge 0 \tag{14.22}$$

は恒等式である．左辺を展開すれば

$$0 \le r_{xx}(0) + \overline{\lambda} r_{xy}(\tau) + \lambda \overline{r_{xy}(-\tau)} + |\lambda|^2 r_{yy}(0)$$

となる．前と同様に

$$\overline{\lambda} r_{xy}(\tau) = \mu |r_{xy}(\tau)|$$

と選べば

$$r_{xx}(0) + 2\mu |r_{xy}(\tau)| + \mu^2 r_{yy}(0) \geq 0 \quad (14.23)$$

が得られる．これが任意の μ について成り立つためには

$$|r_{xy}(\tau)|^2 \leq r_{xx}(0) r_{yy}(0) \quad (14.24)$$

でなければならない．上式が等号をとるときには (14.23) 式がある μ で 0 になる．これは (14.22) 式がある λ で 0 になることを意味しているから，(14.24) 式がある τ で等号をとるのは，ある定数 λ が存在してすべての t に対して

$$x(t+\tau) + \lambda y(t) = 0$$

が成り立つとき，すなわち $x(t)$ と $y(t)$ が時刻の原点移動 τ を行うと一時独立でなくなるときである．(14.7) 式は (14.24) 式の特別な場合である．

相互相関関数を

$$\rho_{xy}(\tau) = \frac{r_{xy}(\tau)}{\sqrt{r_{xx}(0) r_{yy}(0)}} \quad (14.25)$$

によって正規化することができる．これは統計量の相関係数に相当するものである．(14.24) 式によりこれは

$$|\rho_{xy}(\tau)| \leq 1 \quad (14.26)$$

を満たしている．

クロスパワースペクトル $r_{xy}(\tau)$ のフーリエ変換

$$R_{xy}(\omega) = \sum_{\tau=-\infty}^{\infty} r_{xy}(\tau) e^{i\omega\tau} \quad (14.27)$$

を $x(t)$ と $y(t)$ の間のクロスパワースペクトルという．パワースペクトルと違って，クロスパワースペクトルは一般に複素数である．クロスパワースペクトルに対応して，単なるパワースペクトルをオートパワースペクトルと呼ぶことがある．

クロススペクトルに対してもシュワルツの不等式を導くことができる．恒等式

$$E_t\Big[\Big|\sum_{k=-N}^{N} \big[x(t+k) + \lambda y(t+k)\big] e^{ik\omega}\Big|^2\Big] \geq 0$$

を展開すると

$$\sum_{k,l=-N}^{N} \Big[r_{xx}(k-l) + \overline{\lambda} r_{xy}(k-l) + \lambda r_{yx}(k-l) + |\lambda|^2 r_{yy}(k-l)\Big] e^{i(k-l)\omega} \geq 0$$

となるから，前と同様に $n = k - l$ と変数変換して和の順序を入れかえると

$$\sum_{n=-2N}^{2N} \Big(1 - \frac{|n|}{2N+1}\Big) \big[r_{xx}(n) + 2\mu r_{xy}(n) + \mu^2 r_{yy}(n)\big] e^{in\omega} \geq 0$$

となる．ここで N を無限大にすると

$$R_{xx}(\omega) + 2\mu R_{xy}(\omega) + \mu^2 R_{yy}(\omega) \geq 0$$

が得られる．これが任意の μ に対して成り立つためには

$$\big|R_{xy}(\omega)\big|^2 \leq R_{xx}(\omega) R_{yy}(\omega) \quad (14.28)$$

でなければならない．式 (14.7)，(14.24)，(14.28) はすべてシュワルツの不等式である．

相互相関関数と同じようにクロススペクトルを正規化した

$$\gamma_{xy}(\omega) = \frac{R_{xy}(\omega)}{\sqrt{R_{xx}(\omega) R_{yy}(\omega)}} \quad (14.29)$$

をコヒーレンシーという．これはシュワルツの不等式から

$$\big|\gamma_{xy}(\omega)\big| \leq 1 \quad (14.30)$$

を満足する．

入出力間のクロススペクトル 入力データを $x(t)$，これをフィルター $h(t)$ に通したときの出力を $y(t)$ とする．すなわち

$$y(t) = \sum_k h(k) x(t-k) \quad (14.31)$$

とする．まず出力と入力の間の相互相関関数を計算する．

$$\begin{aligned} r_{yx}(\tau) &= E_t[y(t+\tau)\overline{x(t)}] \\ &= \sum_k h(k) r_{xx}(\tau - k) \end{aligned}$$

すなわち

$$r_{oi}(\tau) = \sum_k h(k) r_{ii}(\tau - k) \tag{14.32}$$

が得られる．i は入力を，o は出力を意味している．上式は $h(t)$ と $r_{ii}(\tau)$ の畳み込みにほかならないから，出力と入力の間のクロススペクトルは直ちに

$$R_{oi}(\omega) = H(\omega) R_{ii}(\omega) \tag{14.33}$$

と求められる．$H(\omega)$ はフィルターの周波数応答である．

次に出力の自己相関関数を計算する．

$$r_{oo}(\tau) = E_t\Big[\sum_k h(k)x(t+\tau-k)\overline{\sum_l h(l)x(t-l)}\Big]$$

より

$$\begin{aligned} r_{oo}(\tau) &= \sum_{k,l} h(k) r_{ii}(\tau + l - k)\overline{h(l)} \\ &= \sum_l r_{oi}(\tau + l)\overline{h(l)} \end{aligned} \tag{14.34}$$

となる．これは $r_{oi}(\tau)$ と $\overline{h(k)}$ との，エネルギー有限の関数としての相互相関関数である．したがって出力のパワースペクトルは

$$R_{oo}(\omega) = \overline{H(\omega)} R_{oi}(\omega) = |H(\omega)|^2 R_{ii}(\omega) \tag{14.35}$$

と表される．

入出力間のクロススペクトル，出力のオートスペクトルが求められたので，入出力間のコヒーレンシーを計算することができる．これは (14.33), (14.35) 式から

$$\gamma_{oi}(\omega) = \frac{H(\omega)}{|H(\omega)|} \tag{14.36}$$

となる．したがって

$$|\gamma_{oi}(\omega)| = 1$$

となって，入出力間のコヒーレンシーの絶対値は周波数によらず 1 になる．これはシュワルツの不等式 (14.28) で等号が成り立つのは，$x(t)$ と $y(t)$ が入出力の関係にある場合であることを意味している．

AR モデルによるパワースペクトルの推定　これまでパワースペクトルは自己相関関数のフーリエ変換 (14.13) として定義してきた．しかしこの式を用いてスペクトルを求めるためにはウィンドウを掛ける必要があり，そのためにスペクトルの歪，分解能の低下が避けられない．

下図のように，分散 σ_ξ^2 のホワイトノイズをフィルター $H(z)$ に通したときの出力を $x(t)$ とする．$x(t)$ のパワースペクトル $R_{xx}(\omega)$ は，(14.35)，(14.16) 式から

$$R_{xx}(\omega) = \sigma_\xi^2 |H(\omega)|^2$$

で与えられる．いま，与えられたデータ $x(t)$ に対して $H(z)$ と $\xi(t)$ が求められたとすれば，$x(t)$ のパワースペクトルが求められることになる．もちろん σ_ξ^2 と $H(z)$ は独立に決まるわけではないし，このことを考慮したとしても，解が一義的であるという保障はない．

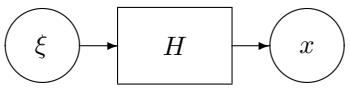


Fig. 14.2

この問題は次のように書きかえることができる．与えられたデータをあるフィルター $A(z)$ に通したとき，出力 $\xi(t)$ をホワイトノイズにすることができるか，という問題である．$A(z)$ と先の $H(z)$ が互いに逆の関係にあることは明らかである．

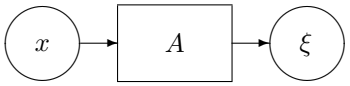


Fig. 14.3

$A(z)$ として因果的なフィルターを仮定すると，上図の入出力の関係は

$$\begin{aligned} &\sum_{k=0}^{M} a(k)x(t-k) = x(t) + a(1)x(t-1) + \cdots \\ &\quad + a(k)x(t-k) + a(M)x(t-M) = \xi(t) \\ &\qquad a(0) = 1 \end{aligned} \tag{14.37}$$

と書くことができる．$a(0) = 1$ としたのは，上式からは $a(k)$ と $\xi(t)$ の絶対値は独立には決まらないからである．上式を

$$x(t) = -a(1)x(t-1) - a(2)x(t-2) - \cdots + \xi(t)$$

と書いてみれば，この式は現在の $x(t)$ を過去のデータ $x(t-1), x(t-2), \cdots$ で線型予測したときの予測誤差が $\xi(t)$ であると読みかえることができる．統計学では，ある確率変数 y が別の確率変数 x によって規定されるような構造を，一般に回帰モデルと呼ぶ．典型的なのは線型回帰モデル

$$y = \alpha + \beta x + \varepsilon$$

である．(14.37) 式は自分自身で自分を近似しているので，自己回帰モデル (auto-regressive model)，簡単に AR モデルと呼ぶ．

こう考えると，$\xi(t)$ を最小二乗法的な意味でできるだけ小さくすれば，結果として $\xi(t)$ がホワイトノイズになることが期待される．最小二乗法は誤差の二乗和を最小にするのであるが，$\xi(t)$ の二乗和は発散してしまうので，二乗和の平均，つまりパワーを最小にすることにする．

パワーを計算するために，はじめに $\xi(t)$ の自己相関関数を計算しておく．これは (14.34) 式から

$$r_{\xi\xi}(\tau) = \sum_{k,l=0}^{M} a(k) r_{xx}(\tau + l - k)\overline{a(l)} \qquad (14.38)$$

である．パワーはラグが 0 のときの値であるから

$$r_{\xi\xi}(0) = \sum_{k,l=0}^{M} a(k) r_{xx}(l - k)\overline{a(l)}$$

である．これが最小になるように $a(k)$ を決める．そのために上式を $a(l)$ で微分して 0 と置くと

$$\sum_{k=0}^{M} a(k) r_{xx}(l-k) = 0 \quad l = 1, 2, 3, \cdots, M \quad (14.39)$$

が得られる．これがいわば正規方程式である．これは M 個の未知数 $a(k)$ についての M 本の式であるから解くことができる．これを解いてその解を $r_{\xi\xi}(0)$ に代入すると，$l = 0$ 以外の項は正規方程式によって 0 になるから

$$\sigma_\xi^2 = r_{\xi\xi}(0) = \sum_{k=0}^{M} a(k) r_{xx}(-k) \qquad (14.40)$$

から最小化されたパワーが求められることになる．一方，(14.38) 式から

$$R_{\xi\xi}(\omega) = |A(\omega)|^2 R_{xx}(\omega)$$

が成り立っている．$A(\omega)$ は $a(k)$ のフーリエ変換である．求められた $\xi(t)$ がホワイトノイズなら $R_{\xi\xi}(\omega) = \sigma_\xi^2$ であるから

$$R_{xx}(\omega) = \frac{\sigma_\xi^2}{|A(\omega)|^2} \qquad (14.41)$$

によって $x(t)$ のパワースペクトルが求められることになる．

以上をまとめると，$x(t)$ のパワースペクトル推定の手順は次のようになる．$x(t)$ から自己相関関数 $r_{xx}(\tau)$ を計算する．これを用いて連立方程式 (14.39) から $a(k)$ を求め，これを (14.40) 式に代入して σ_ξ^2 を求める．$a(k)$，σ_ξ^2 を (14.41) 式に代入すればパワースペクトルが求められる．

自己相関関数のフーリエ変換によってパワースペクトルを求めるときには，自己相関関数が十分減衰するまで，あるいは卓越周期の数倍以上のラグをとらなければならない．しかし後にも示すように，AR モデルによる方法では，非常に短い M で十分な精度のパワースペクトルを求めることができる．

レビンソンのアルゴリズム　正規方程式 (14.39) は対称性がよいので消去法よりは少ない計算量で解くことができる．はじめに，(14.39)，(14.40) 式はまとめて

$$\sum_{k=0}^{M} a(k) r_{xx}(l-k) = \sigma_\xi^2 \delta_l$$
$$l = 0, 1, 2, \cdots, M$$

と書けることに注意する．これは行列を用いると

$$[a(0)\ a(1)\ \cdots a(M)]\boldsymbol{R}_M = [\sigma_\xi^2\ 0\ \cdots\ 0] \quad (14.42)$$

と書くことができる．$\boldsymbol{R}_M$ は $(M+1)\times(M+1)$ の正方行列

$$\boldsymbol{R}_M = \begin{bmatrix} r(0) & r(1) & \cdots & r(M) \\ r(-1) & r(0) & \cdots & r(M-1) \\ \vdots & \vdots & \ddots & \vdots \\ r(-M) & r(-M+1) & \cdots & r(0) \end{bmatrix} \qquad (14.43)$$

である．ただし紙面の節約のために $r_{xx}(k)$ を $r(k)$ と書いてある．このように，副対角線上の要素がすべて等しい行列をテプリッツ型の行列と呼ぶ．さら

にこの行列は自己相関関数の対称性 $r(-k)=\overline{r(k)}$ により

$$\boldsymbol{R}_M^* \equiv \overline{\boldsymbol{R}}_M^T = \boldsymbol{R}_M \tag{14.44}$$

を満足している．すなわち $\boldsymbol{R}_M$ はエルミート対称である．

(14.42) 式を解くために，ある m に対して

$$\begin{aligned}[a_m(0)\ a_m(1)\ \cdots a_m(m)]\boldsymbol{R}_m &= [\alpha_m\ 0\ \cdots\ 0]\\ a_m(0) &= 1\end{aligned} \tag{14.45}$$

を満たす $a_m(k)$ と α_m が求められたとする．$\boldsymbol{R}_m$ は (14.43) 式で定義された $(m+1)\times(m+1)$ の行列である．実際，$m=0$ のときは

$$\alpha_0 = r(0)$$

が解である．したがって (14.45) 式の解から $m+1$ の解が導かれれば任意の m についての解が求められることになる．

そこで (14.45) 式の行ベクトルに 0 を一つ加えた関係

$$\begin{aligned}&[a_m(0)\ a_m(1)\ \cdots\ a_m(m)\ 0]\boldsymbol{R}_{m+1}\\ &\quad = [\alpha_m\ 0\ \cdots\ 0\ \gamma_m]\end{aligned} \tag{14.46}$$

を考えてみる．掛け算を実行してみればわかるように，上式の最初の $m+1$ 本の式は (14.45) 式と同じである．最後の式は

$$\gamma_m = \sum_{k=0}^{m} a_m(k) r(m-k+1) \tag{14.47}$$

であるが，$a_m(k)$ はすべて既知なのでこれは γ_m を決める式である．次に (14.46) 式の左辺の行ベクトルの順序を反対にした式を考えると

$$\begin{aligned}&[0\ \overline{a_m(m)}\ \cdots\ \overline{a_m(1)}\ \overline{a_m(0)}]\boldsymbol{R}_{m+1}\\ &\quad = [\overline{\gamma}_m\ 0\ \cdots\ 0\ \overline{\alpha}_m]\end{aligned} \tag{14.48}$$

が成り立つことがわかる．実際に掛け算を行って対称性 $r(-k)=\overline{r(k)}$ を利用すれば上式が (14.46) 式の複素共役になっていることがわかる．

(14.45) 式の右辺は第一成分以外は 0 である．そこで (14.48) 式に定数 u_m を掛けて (14.46) 式に加えて右辺の最後の成分が 0 になるようにする．そのためには

$$\gamma_m + u_m\overline{\alpha}_m = 0 \tag{14.49}$$

となるように u_m を選べばよい．こうすると加えた式は (14.45) 式の m を $m+1$ で置きかえた式になる．したがって

$$\begin{aligned}a_{m+1}(k) &= a_m(k) + u_m\overline{a_m(m-k+1)}\\ &\qquad\qquad k = 1\sim m\\ a_{m+1}(m+1) &= u_m\\ \alpha_{m+1} &= \alpha_m + u_m\overline{\gamma}_m = \alpha_m(1-|u_m|^2)\end{aligned} \tag{14.50}$$

が得られた．$m=0$ のときの解は求まっているから，上式を $m=0$ からはじめて $m=1,2,\cdots,M-1$ まで計算すれば (14.42) 式の解が得られる．なお，(14.42) 式と (14.45) 式の比較から，α_M が求めたい σ_ξ^2 であることがわかる．

上ではデータが複素数としてきたが，上式の最後の式からわかるようにそのときでも α_m は実数である．また α_m は $\xi(t)$ の分散という意味をもっているので，常に正である．これは連立方程式 (14.42) が正則であることを意味している．これらは $r(k)$ が正定値であることの帰結である．これを逆にいえば，自己相関関数が正しく計算されていないと，(14.42) 式は特異になってしまうかもしれない．

自己相関関数が正しく計算されていたとしても，α_m が非常に小さくなるようなデータでは，丸め誤差のために α_m が負になってしまって正しい解が得られないことがある．このようなことを避けるために係数行列 $\boldsymbol{R}_m$ の対角要素を $1+\delta$ 倍して，いいかえれば $r(0)$ だけを $1+\delta$ 倍して，方程式を安定化する方法がよく用いられる．δ としては 0.01 あるいはそれ以下で十分である．これはデータに分散が $\delta\times r(0)$ の人工的なホワイトノイズを加えることに相当している．

反復計算の 1 ステップは (14.47), (14.49), (14.50) 式からなる．これらの計算量は $2m+2$ であるから，これらを $m=0$ から $M-1$ まで反復するのに必要な計算量は $M(M+1)$ である．これは (14.42) 式を消去法で解くときに必要なオーダー M^3 よりオーダー M だけ小さくなっている．

AIC を用いた次数 M の決定　これまで自己回帰の次数 M は与えられたものとしてきた．容易に想像できるように，M を大きくすればするほど，残

差 $\xi(t)$ を小さくすることができ，またスペクトルの微細な構造を表すことができる．しかし M が大きければ大きいほどいいわけではない．極端な場合，M をデータの個数 N と同じにしてしまえば，$\xi(t)$ がすべての t で 0 になるように $a(k)$ を決めることができる．しかしこれではもちろん意味がない．

最適な次数 M を求める一つの方法は，赤池の情報量規準 AIC を用いることである．m 次のモデルに対する AIC は

$$\mathrm{AIC}_m = N(\ln 2\pi\alpha_m + 1) + 2(m+1) \qquad (14.51)$$

で定義される．N はデータ $x(t)$ の個数である．α_m は m とともに減少するから，上式右辺の第一項は m の減少関数，第二項は増加関数であるから，AIC_m はある m で極小値をとる．このときの m が最適な次数である．

AIC の絶対値にはあまり意味はない．なぜならデータの単位を変えれば第一項の大きさは変わってしまうからである．第二項は 1 ステップで 2 変わるだけであるから，極小点付近の AIC の変化は十進一位の桁以下の数値である．

データの数 N が大きいときには分散 α_m のわずかな減少が N によって拡大されてしまうので，AIC がなかなか極小に達しないことがある．また局所的な弱い極小のために本当の最小値を見逃してしまうことがある．

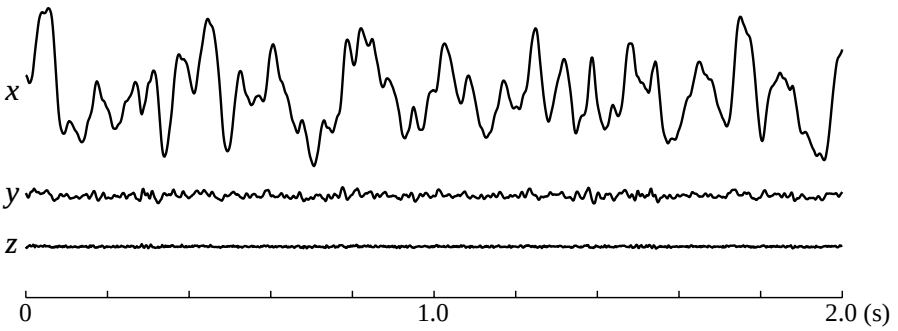


Fig. 4(a) 上段：サンプルデータ x_t．$\Delta t = 2$ms．中段：AR モデルの残差，$y_t = A(z)x_t$．下段：y_t にもう一度 AR モデルを当てはめたときの残差，$z_t = A(z)y_t$．スケールは 3 トレースとも同じ．

例題　微動のデータを用いて計算を行なってみた．Fig. 4(a) の上段が用いたデータで，サンプル間隔は $\Delta t = 2$ ms である．数 Hz から 20 〜30 Hz のさまざまな周期の波がみえる．ここには 2 秒間しか示していないが，実際には 2.4 秒間，1200 個のデータを用いて自己相関関数を計算した．これが Fig. 4(b) 上段の r_x と記されたカーブがそれである．ここではラグが 30 ポイント，すなわち 60 ms までしか示していないが，この後は緩やかに減少を続け，約 100 ms で最小値 $-0.33(r(0)$ を単位として) をとり，増加に転じて 140 ms で零線を切る．

この自己相関関数を用いて $M = 30$ として計算した AR モデルの係数 $a(k)$ が同じ上段に丸印で示してある．これは典型的な $a(k)$ の振舞を示している．すなわち，$a(1)$ が最小で $k \geq 1$ では減衰振動のように変化する．減衰は非常に速く，この例では $k > 10$ では振幅が非常に小さいので，$M = 30$ までとる必要はないのではないかと思われる．

$a(k)$ が求められると $\xi(t)$ がホワイトノイズだと仮定して，(14.41) 式からパワースペクトル $R_{xx}(\omega)$ が計算できる．Fig. 4(c) の実線がこれである．およそ 4 Hz にピークがあり，高周波に向かって急激に減少している．約 50 Hz 以上ではフラットになっている．

(14.41) 式は $\xi(t)$ がホワイトという仮定に基づいている．これが正しいかどうかをチェックするために係数 $a(k)$ を用いて (14.37) 式から $\xi(t)$ を計算したのが Fig. 4(a) の中段のトレースである．縦軸は $x(t)$ と同じスケールであるから，AR モデルがよく当てはまっていることがわかる．

しかしこれだけでは予測誤差がホワイトかどうかわからないので，その自己相関関数を計算したのが Fig. 14.4(b) の中段の実線 r_y である．本当にホワイトならこのグラフは原点以外では 0 になるはずで

あるが，ここでは約 10 ms 程度の幅をもっている．そこでこの自己相関関数から自己回帰係数 $a(k)$ を求めたのが，中段の a_y の丸印のグラフである．これも減衰振動の形をしているが，上段の a_x よりは暴れている．この係数を用いて $\xi(t)$ のパワースペクトルを計算したものが Fig. 14.4(c) の P_y で示した破線である．これは P_x よりも 3 桁小さく，約 50

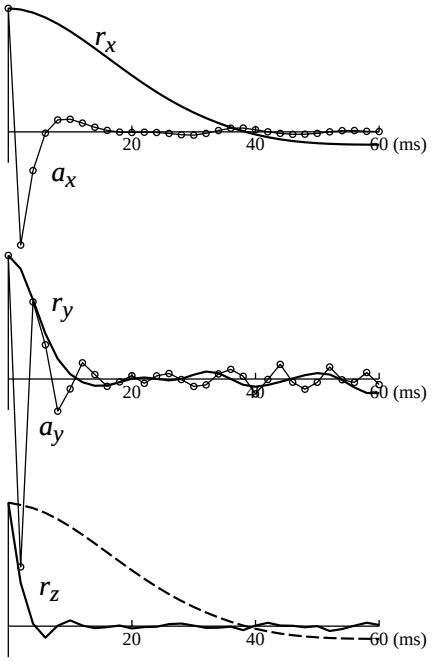


Fig. 4(b) 上段：$x(t)$ の自己相関関数 (r_x) と自己回帰係数 (a_x)．中段：$\xi(t)$ の自己相関関数 (r_y) と自己回帰係数 (a_y)．下段：$\xi(t)$ に AR モデルを当てはめたときの残差の自己相関関数．破線は a_x の逆の自己相関関数．

Hz までほとんど平坦である．したがってこの周波数までを考える限り，$\xi(t)$ はホワイトという仮定は成り立っていると思ってよいことがわかる．

Fig. 14.4(b) の中段の a_y を $\xi(t)$ の掛けて予測誤差を計算したものが Fig. 14.4(a) の下段，z である．これは $\xi(t)$ に自己回帰モデルを当てはめたときの予測誤差である．スケールは x と同じである．このトレースの自己相関関数が Fig. 14.4(b) の下段の r_z である．これはラグが 3 以上で 0 とみなしてよい．

これまでの計算はすべて $M = 30$ としており，この次数が AIC の判断基準を満たしているわけではない．そこで $x(t)$ に対して計算された α_m と AIC_m を下表に示す．ただしここに示してあるのは AIC_m

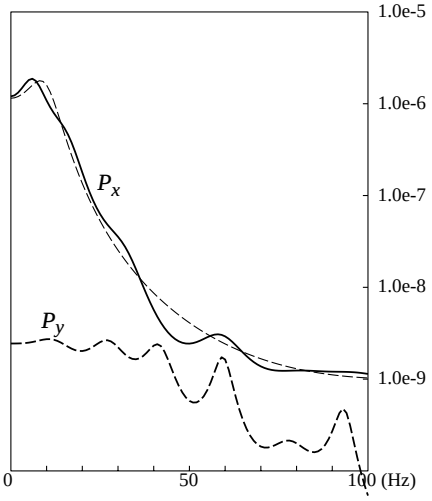


Fig. 14.4(c) P_x：実線は $M = 30$ のとき，破線は $M = 6$ のときのパワースペクトル．P_y：予測誤差 $\xi(t)$ のパワースペクトル．

そのものではなく，絶対値を小さくするために m に依存しない項を除いた

$$\mathrm{AIC}_m - \mathrm{AIC}_0 = N \ln(\alpha_m/\alpha_0) + 2m$$

である．残差の分散 α_m は，はじめ急速に減少するがすぐに減少の速度が鈍り第二項の効果がきいてきて AIC_m は増加に転じる．この例の場合 α_m は $m = 6$ 以降ほとんど変化せず，AIC は $m = 6$ で最小値をとる．しかしこの最小値の谷は非常に浅く，また $m = 9$ にも孤立した弱い極小値がある．

m	α_m	AIC_m	m	α_m	AIC_m
0	8.754e-8	0.0	10	2.574e-9	-4211.8
1	3.388e-9	-3900.3	11	2.573e-9	-4210.3
2	3.306e-9	-3927.6	12	2.573e-9	-4208.4
3	2.883e-9	-4089.9	13	2.573e-9	-4206.4
4	2.656e-9	-4186.1	14	2.573e-9	-4204.5
5	2.593e-9	-4213.0	15	2.572e-9	-4203.0
6	2.587e-9	-4213.7	16	2.569e-9	-4202.2
7	2.587e-9	-4212.0	17	2.567e-9	-4201.5
8	2.582e-9	-4212.2	18	2.566e-9	-4199.9
9	2.577e-9	-4212.5	19	2.566e-9	-4197.9

AIC_m は $m = 6$ で最小になっているので，$M = 6$ として $x(t)$ のパワースペクトルを推定したものが Fig. 14.4(c) の P_x の細い破線である．これは $M = 30$ のときのパワースペクトルをスムーズにし

たものになっている．自己相関関数のグラフ r_x ではラグが 6 ポイントでは十分に減衰していないから，フーリエスペクトルの常識から $M = 6$ は短すぎる．

これは次のように考えれば理解できる．(14.37) 式を z 変換を用いて表すと

$$x(t) = \frac{1}{A(z)}\xi(t)$$

となる．逆フィルター $1/A(z)$ が安定であることは，$a(k)$ が (14.39) 式の解であることで保障されている．§13 で示したように，逆フィルター $1/A(z)$ のインパルス応答 $h(k)$ は一般に半無限であり，$\xi(t)$ がホワイトのときには $x(t)$ の自己相関関数は $h(k)$ のそれ，すなわち

$$r_{xx}(\tau) = \sigma_\xi^2 \sum_{k=0}^{\infty} h(\tau + k)\overline{h(k)}$$

で表される．M が短くても右辺は $r_{xx}(\tau)$ のよい近似になっている．実際，$M = 30$ のときの $1/A(z)$ を計算し，上式の右辺をプロットしたものが Fig. 14.4(b) の下段の破線であり，このスケールではこれは上段の r_x と見分けがつかない．

多変量 AR モデルによるクロスパワースペクトルの推定　クロスパワースペクトルを求めるためには，多変量の AR モデルを用いなければならない．多変量という意味は，時系列が $x(t)$ だけでなく，$y(t)$, $z(t)$ など，たくさんの時系列を同時に考えるということである．

n チャンネルのデータ $x_1(t), x_2(t), \cdots, x_n(t)$ をひとまとめにして列ベクトル

$$\boldsymbol{x}(t) = [x_1(t)\ x_2(t)\ \cdots\ x_n(t)]^T \tag{14.52}$$

で表す．各チャンネル間の相互相関関数は，前と同様に

$$r_{ij}(\tau) = E_t[x_i(t+\tau)\overline{x_j(t)}]$$

で定義されるが，これを成分とする行列を

$$\begin{aligned}\boldsymbol{r}(\tau) &= E_t[\boldsymbol{x}(t+\tau)\boldsymbol{x}^*(t)] \\ &= \begin{bmatrix} r_{11}(\tau) & r_{12}(\tau) & \cdots & r_{1n}(\tau) \\ r_{21}(\tau) & r_{22}(\tau) & \cdots & r_{2n}(\tau) \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1}(\tau) & r_{n2}(\tau) & \cdots & r_{nn}(\tau) \end{bmatrix}\end{aligned} \tag{14.53}$$

と書く．* は複素共役をとって転置すること

$$\boldsymbol{x}^* = \overline{\boldsymbol{x}}^T$$

を意味している．$\boldsymbol{x}$ が列ベクトルのとき $\boldsymbol{x}^*$ は行ベクトルになる．$\boldsymbol{r}(\tau)$ の中身はチャンネル間の相互相関関数であるが，$\boldsymbol{x}(t)$ という多チャンネルデータから見ると自己相関関数になっている．相互相関関数の対称性 (14.21) 式を用いると $\boldsymbol{r}(\tau)$ は対称性

$$\boldsymbol{r}(-\tau) = \boldsymbol{r}^*(\tau) \tag{14.54}$$

を満足している．

多チャンネルの予測誤差フィルターの関係は (14.37) 式を多チャンネル化した

$$\boldsymbol{\xi}(t) = \sum_{k=0}^{M} \boldsymbol{a}(k)\boldsymbol{x}(t-k) \qquad \boldsymbol{a}(0) = \boldsymbol{I}_n \tag{14.55}$$

と書くことができる．入力 $\boldsymbol{x}(t)$ は n 元のベクトルであるから，フィルターの係数 $\boldsymbol{a}(k)$ は $n \times n$ の正方行列であり，予測誤差 $\boldsymbol{\xi}(t)$ は n 元のベクトルである．また $\boldsymbol{I}_n$ は $n \times n$ の単位行列である．

予測誤差の自己相関関数は

$$\begin{aligned}\boldsymbol{r}_\xi(\tau) &= E_t[\boldsymbol{\xi}(t+\tau)\boldsymbol{\xi}^*(t)] \\ &= E_t\Big[\sum_{k,l} \boldsymbol{a}(k)\boldsymbol{x}(t+\tau-k)\boldsymbol{x}^*(t-l)\boldsymbol{a}^*(l)\Big] \\ &= \sum_{k,l=0}^{M} \boldsymbol{a}(k)\boldsymbol{r}(\tau+l-k)\boldsymbol{a}^*(l)\end{aligned} \tag{14.56}$$

となる．$\boldsymbol{r}(\tau)$ は (14.53) 式で定義された自己相関関数である．この式は (14.38) 式と同じ形をしているが，行列の演算であるから積の順序を入れかえることはできない．

1 チャンネルのときには自己相関関数のラグが 0 のときの値は分散であったが，多チャンネルのときには $\boldsymbol{\xi}(t)$ のチャンネル間の共分散行列になる．

$$\begin{aligned}\boldsymbol{\sigma}_\xi^2 &= \boldsymbol{r}_\xi(0) \\ &= \sum_{k,l=0}^{M} \boldsymbol{a}(k)\boldsymbol{r}(l-k)\boldsymbol{a}^*(l)\end{aligned} \tag{14.57}$$

1 チャンネルのときには分散が最小になるように $a(k)$ を決めることができたが，多チャンネルの場合には $\boldsymbol{\sigma}_\xi^2$ が行列であるからこれが最小というのは意味がない．そこでここでは $\boldsymbol{\sigma}_\xi^2$ が $\boldsymbol{a}(k)$ に関して停留と

いう条件で $\boldsymbol{a}(k)$ を決めることにする．$\boldsymbol{\sigma}_\xi^2$ の対角成分は各チャンネルの分散であるから，これが停留であるということは各チャンネルの分散が極小であることを意味している．

(14.57) 式を $\boldsymbol{a}(m)$ について変分をとると

$$\delta\boldsymbol{\sigma}_\xi^2 = \delta\boldsymbol{a}(m)\sum_{l=0}^{M}\boldsymbol{r}(l-m)\boldsymbol{a}^*(l) + \sum_{k=0}^{M}\boldsymbol{a}(k)\boldsymbol{r}(m-k)\delta\boldsymbol{a}^*(m)$$

となるが，対称性 (14.54) 式によって第一項と第二項は互いに複素共役転置の関係になっている．したがって $\boldsymbol{\sigma}_\xi^2$ が停留になるための条件は

$$\sum_{k=0}^{M}\boldsymbol{a}(k)\boldsymbol{r}(l-k) = \boldsymbol{0} \quad l = 1, 2, \cdots, M \tag{14.58}$$

となる．右辺は $n \times n$ の零行列である．$\boldsymbol{a}(0) = \boldsymbol{I}_n$ であるから上式は $\boldsymbol{a}(1)$ から $\boldsymbol{a}(M)$ までの M 個の未知行列に関する連立方程式である．これは多チャンネルに拡張したレビンソンの方法で解くことができる．

$\boldsymbol{a}(k)$ が求められたとすると残差の共分散行列は (14.57) 式から

$$\boldsymbol{\sigma}_\xi^2 = \sum_{k=0}^{M}\boldsymbol{a}(k)\boldsymbol{r}(-k) \tag{14.59}$$

となる．これも (14.40) 式と同じ形をしている．

次にパワースペクトルの関係を導くために (14.56) 式の両辺をフーリエ変換する．

$$\begin{aligned}\boldsymbol{P}_\xi(\omega) &\equiv \sum_{\tau=-\infty}^{\infty}\boldsymbol{r}_\xi(\tau)e^{i\omega\tau} \\ &= \sum_{k,l,\tau}\boldsymbol{a}(k)\boldsymbol{r}(\tau+l-k)\boldsymbol{a}^*(l)e^{i\omega\tau} \\ &= \sum_{k}\boldsymbol{a}(k)e^{i\omega k}\sum_{\tau}\boldsymbol{r}(\tau+l-k)e^{i\omega(\tau+l-k)} \\ &\quad \times\Big[\sum_{l}\boldsymbol{a}(l)e^{i\omega l}\end{aligned}$$

$\boldsymbol{P}_\xi(\omega)$ は $n \times n$ の行列 $\boldsymbol{r}_\xi(\tau)$ の各成分をフーリエ変換して得られる $n \times n$ の行列である．いま $\boldsymbol{a}(k)$ のフーリエ変換を

$$\boldsymbol{A}(\omega) = \sum_{k=0}^{M}\boldsymbol{a}(k)e^{i\omega k} \tag{14.60}$$

と定義すれば先の式は

$$\boldsymbol{P}_\xi(\omega) = \boldsymbol{A}(\omega)\boldsymbol{P}(\omega)\boldsymbol{A}^*(\omega) \tag{14.61}$$

となる．$\boldsymbol{P}(\omega)$ の i, j 成分は相互相関関数 $r_{ij}(\tau)$ のフーリエ変換，すなわちクロスパワースペクトルで，これが求めたいものである．シングルチャンネルのときと同様に $\boldsymbol{\xi}(t)$ がホワイトになったと仮定すれば

$$\boldsymbol{r}_\xi(\tau) = 0 \quad \tau \neq 0$$

であるから

$$\boldsymbol{P}_\xi(\omega) = \boldsymbol{r}_\xi(0) = \boldsymbol{\sigma}_\xi^2$$

となる．したがって (14.61) 式を $\boldsymbol{P}(\omega)$ について解くと

$$\boldsymbol{P}(\omega) = \boldsymbol{A}^{-1}(\omega)\boldsymbol{\sigma}_\xi^2[\boldsymbol{A}^*(\omega)]^{-1} \tag{14.62}$$

が得られる．

以上をまとめると，相互相関関数 $r_{ij}(\tau)$ を計算して連立方程式 (14.58) 式から $\boldsymbol{a}(k)$ を求め，(14.59) 式から共分散行列 $\boldsymbol{\sigma}_\xi^2$ を計算する．行列 (14.60) 式の逆行列を計算すれば (14.62) 式からパワースペクトル $\boldsymbol{P}(\omega)$ が得られる．$\boldsymbol{P}(\omega)$ の対角成分はオートスペクトル，非対角成分はクロススペクトルである．

マルチチャンネルのレビンソン法　パワースペクトル $\boldsymbol{P}(\omega)$ を求めるのに必要なのは (14.58) 式の $\boldsymbol{a}(k)$ と，(14.59) 式の $\boldsymbol{\sigma}_\xi^2$ である．これらはひとまとめにして

$$[\boldsymbol{a}(0)\,\boldsymbol{a}(1)\,\cdots\,\boldsymbol{a}(M)]\boldsymbol{R}_M = [\boldsymbol{\sigma}_\xi^2\,\boldsymbol{0}\,\cdots\,\boldsymbol{0}] \tag{14.63}$$

と書くことができる．[ ] でくくったのは $M+1$ 成分の行ベクトルを表しているが，その成分は $n \times n$ の行列である．$\boldsymbol{0}$ は $n \times n$ の零行列，また $\boldsymbol{R}_M$ は (14.43) 式とまったく同様に定義されるが，要素がスカラー $r(\tau)$ ではなく，$n \times n$ の行列 $\boldsymbol{r}(\tau)$ である．この行列の行列も対称性 (14.44) 式を満たしている．

(14.63) 式を解くために，ある m に対して

$$\begin{aligned}&[\boldsymbol{a}_m(0)\,\boldsymbol{a}_m(1)\,\cdots\,\boldsymbol{a}_m(m)]\boldsymbol{R}_m = [\boldsymbol{\alpha}_m\,\boldsymbol{0}\,\cdots\,\boldsymbol{0}] \\ &[\boldsymbol{b}_m(m)\,\cdots\,\boldsymbol{b}_m(1)\,\boldsymbol{b}_m(0)]\boldsymbol{R}_m = [\boldsymbol{0}\,\cdots\,\boldsymbol{0}\,\boldsymbol{\beta}_m] \\ &\boldsymbol{a}_m(0) = \boldsymbol{b}_m(0) = \boldsymbol{I}_n\end{aligned} \tag{14.64}$$

を満たす解 $\boldsymbol{a}_m(k)$, $\boldsymbol{\alpha}_m$, $\boldsymbol{b}_m(k)$, $\boldsymbol{\beta}_m$ があったとする．実際，$m=0$ のときの解は

$$\boldsymbol{\alpha}_0 = \boldsymbol{r}(0) = \boldsymbol{\beta}_0$$

である．

$\boldsymbol{a}_m(k)$ は次数 m の予測誤差フィルター，$\boldsymbol{\alpha}_m$ は予測誤差の共分散行列である．一方 $\boldsymbol{b}_m(k)$ は時間の逆方向に予測するときの予測誤差フィルターで，$\boldsymbol{\beta}_m$ はそのときの予測誤差の共分散行列を意味している．1 変数のときには $b_m(k) = \bar{a}_m(k)$ と選ぶことができたが，マルチチャンネルのときには正逆フィルターを独立に決めなければならないので計算が複雑になる．これは行列の積が可換でないためである．

(14.64) 式の行ベクトルに零行列を一つ加えた式

$$\begin{aligned}
&[\boldsymbol{a}_m(0)\,\boldsymbol{a}_m(1)\,\cdots\,\boldsymbol{a}_m(m)\,\boldsymbol{0}]\boldsymbol{R}_{m+1} \\
&\qquad = [\boldsymbol{\alpha}_m\,\boldsymbol{0}\,\cdots\,\boldsymbol{0}\,\boldsymbol{\gamma}_m] \qquad (14.65)\\
&[\boldsymbol{0}\,\boldsymbol{b}_m(m)\,\cdots\,\boldsymbol{b}_m(1)\,\boldsymbol{b}_m(0)]\boldsymbol{R}_{m+1} \\
&\qquad = [\boldsymbol{\delta}_m\,\boldsymbol{0}\,\cdots\,\boldsymbol{0}\,\boldsymbol{\beta}_m] \qquad (14.66)
\end{aligned}$$

を考えてみる．ベクトルの要素がひとつ増えたので，行列としては行と列がひとつづつ増えた $\boldsymbol{R}_{m+1}$ を用いなければならない．掛け算を実行してみればわかるように，(14.65) 式の最初の $m+1$ 本の式，(14.66) 式の最後の $m+1$ 本の式は (14.64) 式と全く同じである．残りの式から

$$\begin{aligned}
\boldsymbol{\gamma}_m &= \sum_{k=0}^{m} \boldsymbol{a}_m(k)\boldsymbol{r}(m-k+1) \\
\boldsymbol{\delta}_m &= \sum_{k=0}^{m} \boldsymbol{b}_m(k)\boldsymbol{r}(k-m-1) \qquad (14.67)
\end{aligned}$$

が計算される．

(14.65) 式と (14.66) 式の適当な線型結合を作ってやれば (14.64) と同様な式を作ることができる．$\boldsymbol{a}_m(k)$ に対する式を作るには (14.66) 式に左からある行列 $\boldsymbol{u}_m$ を掛けて (14.65) 式に加えてやればよい．その結果右辺の最後の成分が $\boldsymbol{\gamma}_m + \boldsymbol{u}_m\boldsymbol{\beta}$ になるが，これを 0 にすれば (14.64) の第一式と同じ形になる．すなわち

$$\boldsymbol{\gamma}_m + \boldsymbol{u}_m\boldsymbol{\beta}_m = \boldsymbol{0} \qquad (14.68)$$

のように $\boldsymbol{u}_m$ を選べばよい．この式は $\boldsymbol{u}_m$ の成分に対する連立一次方程式で，n 元の連立方程式を n 回解くことによって $\boldsymbol{u}_m$ が得られる．この $\boldsymbol{u}_m$ を用いると加えた式は (14.64) の第一式の m を $m+1$ で置き換えたものになる．したがって m がひとつ増えた係数 $\boldsymbol{a}_{m+1}(k)$ は

$$\begin{aligned}
&\boldsymbol{a}_{m+1}(k) = \boldsymbol{a}_m(k) + \boldsymbol{u}_m\boldsymbol{b}_m(m-k+1) \\
&\qquad\qquad\qquad\qquad k = 1 \sim m \\
&\boldsymbol{a}_{m+1}(m+1) = \boldsymbol{u}_m \qquad (14.69)\\
&\boldsymbol{\alpha}_{m+1} = \boldsymbol{\alpha}_m + \boldsymbol{u}_m\boldsymbol{\delta}_m
\end{aligned}$$

によって求めることができる．これで $\boldsymbol{a}_{m+1}(k)$ が求まったがマルチチャンネルの場合には $\boldsymbol{b}_{m+1}(k)$ も求めておかなければならない．そのためには (14.65) 式に $n \times n$ の行列 $\boldsymbol{v}_m$ を掛けて (14.66) 式に加え，第一成分を零行列にしてやればよい．すなわち

$$\boldsymbol{\delta}_m + \boldsymbol{v}_m\boldsymbol{\alpha}_m = \boldsymbol{0} \qquad (14.70)$$

から $\boldsymbol{v}_m$ を決め，$\boldsymbol{b}_{m+1}(k)$ は

$$\begin{aligned}
&\boldsymbol{b}_{m+1}(k) = \boldsymbol{b}_m(k) + \boldsymbol{v}_m\boldsymbol{a}_m(m+1-k) \\
&\qquad\qquad\qquad\qquad k = 1 \sim m \\
&\boldsymbol{b}_{m+1}(m+1) = \boldsymbol{v}_m \qquad (14.71)\\
&\boldsymbol{\beta}_{m+1} = \boldsymbol{\beta}_m + \boldsymbol{v}_m\boldsymbol{\gamma}_m
\end{aligned}$$

とすればよい．$m=0$ のときの解は求まっているから，上式の反復を $m = 1,\ 2, \cdots,\ M-1$ まで行えばよい．

$\boldsymbol{\alpha}_m$，$\boldsymbol{\beta}_m$ は予測誤差の共分散行列であるから，相互相関関数が正しく計算されていれば

$$\boldsymbol{\alpha}_m^* = \boldsymbol{\alpha}_m \qquad \boldsymbol{\beta}_m^* = \boldsymbol{\beta}_m \qquad (14.72)$$

を満たしているはずである．とくに実数データのときには $\boldsymbol{\alpha}_m$，$\boldsymbol{\beta}_m$ は対称行列になるはずである．上の公式ではこの性質を用いていない．また $\boldsymbol{\gamma}_m$ と $\boldsymbol{\delta}_m$ は実は独立ではない．(14.69) 式，(14.71) 式の最後の式から $\boldsymbol{u}_m$，$\boldsymbol{v}_m$ を消去すると

$$\begin{aligned}
\boldsymbol{\alpha}_{m+1} &= \boldsymbol{\alpha}_m - \boldsymbol{\gamma}_m\boldsymbol{\beta}_m^{-1}\boldsymbol{\delta}_m \\
\boldsymbol{\beta}_{m+1} &= \boldsymbol{\beta}_m - \boldsymbol{\delta}_m\boldsymbol{\alpha}_m^{-1}\boldsymbol{\gamma}_m
\end{aligned}$$

となるから，たとえば $\boldsymbol{\alpha}_{m+1}^*$ は

$$\boldsymbol{\alpha}_{m+1}^* = \boldsymbol{\alpha}_m^* - \boldsymbol{\delta}_m^*(\boldsymbol{\beta}_m^{-1})^*\boldsymbol{\gamma}_m^*$$

となる．したがって $\boldsymbol{\alpha}_m$ と $\boldsymbol{\beta}_m$ が対称であるときに $\boldsymbol{\alpha}_{m+1}$ が対称であるためには

$$\boldsymbol{\delta}_m = \boldsymbol{\gamma}_m^* \qquad (14.73)$$

でなければならない．これらふたつの対称性 (14.72), (14.73) 式を用いれば計算量をかなり節約することができる．

マルチチャンネルの場合の赤池の情報量規準 (AIC) は

$$\mathrm{AIC}_m = N(n\log 2\pi + \log|\boldsymbol{\alpha}_m| + n) + n(n+1) + 2n^2 m \tag{14.74}$$

で定義される．ここに N はデータ $\boldsymbol{x}(t)$ の長さ，n はチャンネル数，$|\boldsymbol{\alpha}_m|$ は共分散行列 $\boldsymbol{\alpha}_m$ の行列式の値で，1 チャンネルの場合の α_m に相当する量である．

15 常微分方程式の解法

1 階常微分方程式 本節では

$$\frac{dy}{dx} = f(x, y) \tag{15.1}$$

の形の常微分方程式を，初期条件

$$y(x_n) = y_n$$

で積分する，すなわち初期値問題を解く．以下では次のような記号を用いる．

$$\begin{aligned} x_{n+1} &= x_n + h \\ y_{n+1} &= y(x_{n+1}) \\ f_n &= f(x_n, y_n) \end{aligned}$$

連立常微分方程式 未知関数が一つはでなく，たとえば連立常微分方程式

$$\begin{aligned} \frac{dy_1}{dx} &= f_1(x, y_1, y_2) \\ \frac{dy_2}{dx} &= f_2(x, y_1, y_2) \end{aligned} \tag{15.2}$$

を解きたい場合には，ベクトル

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \qquad \boldsymbol{f} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \tag{15.3}$$

を定義すれば，連立常微分方程式 (15.2) は

$$\frac{d\boldsymbol{y}}{dx} = \boldsymbol{f}(x, \boldsymbol{y}) \tag{15.4}$$

と書くことができる．これは (15.1) 式と全く同じ形をしている．

高階常微分方程式 たとえば二階の常微分方程式

$$\frac{d^2y}{dx^2} = f(x, y, y')$$

を解きたいときには

$$y_1 = y \qquad y_2 = \frac{dy}{dx}$$

と置けば，もとの微分方程式はベクトルの常微分方程式

$$\frac{d}{dx}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \boldsymbol{f}(x, \boldsymbol{y}) = \begin{bmatrix} y_2 \\ f(x, y_1, y_2) \end{bmatrix}$$

に書きかえることができる．一般に n 階の常微分方程式は n 元のベクトルの 1 階の常微分方程式に書きかえることができる．

したがってベクトル型の一階常微分方程式 (15.4) が最も一般的であるが，記述を簡単にするためにここではスカラー型の常微分方程式 (15.1) だけをとりあげる．以下で述べるアルゴリズムをベクトル型に書きかえるのは簡単である．

テーラー展開法 $y(x_n + h)$ をテーラー展開すると

$$\begin{aligned} y(x_n + h) = y_n &+ h\left(\frac{dy}{dx}\right)_n + \frac{1}{2}h^2\left(\frac{d^2y}{dx^2}\right)_n \\ &+ \frac{1}{3!}h^3\left(\frac{d^3y}{dx^3}\right)_n + \cdots \end{aligned} \tag{15.5}$$

となる．微分の添字 n は x_n における値であることを示している．これらの微分は (15.1) 式を用いれば

$$\begin{aligned} \frac{dy}{dx} &= f(x, y) \\ \frac{d^2y}{dx^2} &= \frac{df}{dx} = f_x + f_y\frac{dy}{dx} = f_x + f_y f \\ \frac{d^3y}{dx^3} &= f_{xx} + 2f_{xy}f + f_x f_y + f_y^2 f + f_{yy}f^2 \end{aligned} \tag{15.6}$$

になる．f の添字 x, y は偏微分を意味している．これらの微分が計算できるなら，これを (15.5) 式に代入すれば x_{n+1} における y の値，y_{n+1} が求められることになる．しかし $f(x, y)$ がよほど簡単な関数でなければ，これらの微分を計算することは難しい．

最も簡単なのは (15.5) 式の展開を一次までとった公式

$$y_{n+1} = y_n + hf(x_n, y_n) \tag{15.7}$$

で，これはオイラーの公式と呼ばれているが，もちろん実用的ではない．

ルンゲ-クッタ型公式の原理 $y_n = y(x_n)$ が与えられているとき，x_{n+1} における $y(x)$ の値 y_{n+1} が

$$\begin{aligned} y_{n+1} &= y_n + \gamma_1 k_1 + \gamma_2 k_2 \\ k_1 &= hf(x_n, y_n) \\ k_2 &= hf(x_n + \alpha h, y_n + \beta k_1) \end{aligned} \tag{15.8}$$

で表されると仮定する．y_{n+1} はテーラー展開を二次までとると，(15.5)，(15.6) 式から

$$y_{n+1} = y_n + hf + \frac{1}{2}h^2(f_x + f_y f) + O(h^3)$$

になる．f などの引数は (x_n, y_n) であるが省略してある．一方，k_2 を展開すれば

$$k_2 = hf + h^2(\alpha f_x + \beta f_y f) + O(h^3)$$

となる．これを (15.8) 式に代入して先のテーラー展開の式の h^2 までのと係数を比較すれば

$$\begin{aligned} &\gamma_1 + \gamma_2 = 1 \\ &\gamma_2\alpha = \frac{1}{2} \quad \gamma_2\beta = \frac{1}{2} \end{aligned} \tag{15.9}$$

が得られる．四つの未知数 $\alpha, \beta, \gamma_1, \gamma_2$ に対して式が三つしかないから式の数が少なすぎる．しかし (15.6) 式などを用いて h^3 の係数を比較すると，こんどは式の数の方が多すぎて解が存在しなくなる．したがって (15.8) 式の形の解を仮定するかぎり，テーラー展開の 2 次までが一致する (15.9) 式が最良のものである．この公式では $f(x, y)$ を 2 回計算しており，h の 2 次までがテーラー展開の解に一致するので，2 段 2 次の公式と呼ぶ．

(15.9) 式では $\alpha = \beta$ であるから β を消去すると未知数は 3 個，式は二つである．いま α をパラメータに選ぶと，次のような積分公式が得られる．

$$\begin{aligned} y_{n+1} &= y_n + \Big(1 - \frac{1}{2\alpha}\Big)k_1 + \frac{1}{2\alpha}k_2 \\ k_1 &= hf(x_n, y_n) \\ k_2 &= hf(x_n + \alpha h, y_n + \alpha k_1) \end{aligned} \tag{15.10}$$

$\alpha = 1$ のときはホイン (Heun) の公式，$\alpha = 1/2$ のときは修正オイラー公式と呼ばれる．

上では 2 段公式を導いた．一般に p 段公式は

$$\begin{aligned} y_{n+1} &= y_n + \sum_{i=1}^{p} \gamma_i k_i \\ k_i &= hf\big(x_n + \alpha_i h, y_n + \sum_{j=1}^{i-1} \beta_{ij} k_j\big) \\ &\qquad i = 1, 2, \cdots, p \end{aligned} \tag{15.11}$$

と書くことができる．係数 $\alpha_i, \beta_{ij}, \gamma_i$ を決めるためには (15.6) 式のような展開を p 回微分まで計算しなければならない．これは大変な計算になり，ちょっとためしにやってみるというわけにはいかない．そこでここでは結果だけを示す．

ルンゲ-クッタ法　4 段 ($p = 4$) の公式はテーラー展開の 4 次まで合わせることができるが，2 段のときと同様に一義的な解は存在しない．古典的なルンゲ-クッタ法の公式はそのなかの一つで

$$\begin{aligned} y_{n+1} &= y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) + O(h^5) \\ k_1 &= hf(x_n, y_n) \\ k_2 &= hf\big(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1\big) \\ k_3 &= hf\big(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_2\big) \\ k_4 &= hf\big(x_n + h, y_n + k_3\big) \end{aligned} \tag{15.12}$$

で表される．係数が非常に単純で覚えやすい．$f(x, y)$ の引数 y の部分には直前の k だけしか含まれていないのもわかりやすい．また引数 x はステップの始点と終点 x_n, $x_n + h$ と，中点 $x_n + h/2$ しか含まれていない．

最後の点は実用上非常に便利である．たとえば弾性定数が x だけによって変化する媒質中を伝わる弾性波の問題を解くときに，$f(x, y)$ に含まれる弾性定数は表の形で与えられることが多い．この表の間隔が $h/2$ であれば上の公式で積分を行うことができる．もちろんすべての間隔が $h/2$ である必要はなく，$h_1/2, h_1/2, h_2/2, h_2/2, \cdots$ のように二つづつが同じであればよい．

微係数 $f(x, y)$ が y によらないとき，微分方程式 (15.1) 式は

$$\frac{dy}{dx} = f(x)$$

であるから，積分は

$$y(x) = y_n + \int_{x_n}^{x} f(x)dx$$

になる．第二項の積分にシンプソンの公式を用いると

$$y_{n+1} = y_n + \frac{h}{6}\big[f(x_n) + 4f(x_n + h/2) + f(x_{n+1})\big]$$

となるが，この場合 (15.12) 式で $k_2 = k_3$ であるから，ルンゲ-クッタの公式はシンプソンの公式に等しくなる．

この公式を連立常微分方程式 (15.4) に拡張することは容易である．y，f がベクトルになるのであるから，k_i もベクトルになる．

(15.12) 式のような式を導いた先人は本当に偉いと思う．いまになってみれば答えがあることがわかっているから，たとえば k_4 を h^4 まで展開しろと宿題に出されたら，がんばって何時間もかけて計算するかもしれない．しかし答えがあるかないかわからない段階で，この展開を計算するには，よほど計算が達者でなければできないことである．

ルンゲ-クッタ-ギル法　この方法は本質的にはルンゲ-クッタ法と同じであるが，丸め誤差を少なくするための工夫がこらされている．

$$
\begin{aligned}
k_1 &= hf(x_n, y_n)\\
z_1 &= y_n + \big(\frac{1}{2}k_1 - q_0\big)\\
q_1 &= q_0 + 3\big(\frac{1}{2}k_1 - q_0\big) - \frac{1}{2}k_1\\
k_2 &= h\big(x_n + \frac{1}{2}h, z_1\big)\\
z_2 &= z_1 + \gamma_2(k_2 - q_1)\\
q_2 &= q_1 + 3\gamma_2(k_2 - q_1) - \gamma_2 k_2\\
k_3 &= hf\big(x_n + \frac{1}{2}h, z_2\big)\\
z_3 &= z_2 + \gamma_3(k_3 - q_2)\\
q_3 &= q_2 + 3\gamma_3(k_3 - q_2) - \gamma_3 k_3\\
k_4 &= hf(x_n + h, z_3)\\
y_{n+1} &= z_3 + \frac{1}{3}\big(\frac{1}{2}k_4 - q_3\big)\\
q_4 &= q_3 + \big(\frac{1}{2}k_4 - q_3\big) - \frac{1}{2}k_4
\end{aligned}
\qquad (15.13)
$$

ここに

$$
\gamma_2 = 1 - \frac{1}{\sqrt{2}} = \frac{1}{2+\sqrt{2}} = 0.292893\cdots
$$

$$
\gamma_3 = 1 + \frac{1}{\sqrt{2}} = \frac{2+\sqrt{2}}{2} = 1.70710\cdots
$$

である．q_i は丸め誤差を補正するための変数で，一番初めに積分を開始するときに q_0 を 0 にする．q_4 は次のステップの q_0 として用いる．これらの式の括弧は重要で，この形のままで計算を行わなければならない．たとえば第三式を

$$
q_1 = k_1 - 2q_0
$$

と書きかえてしまうと丸め誤差の補正が行われない．

ルンゲ-クッタ法，ルンゲ-クッタ-ギル法は 4 次の公式であるから打切り誤差は 5 次から始まる．したがって刻み幅 h で 1 ステップだけ積分したときの値は

$$
y_1(x+h) = y(x+h) + ch^5 + O(h^6)
$$

と書けるであろう．右辺の $y(x+h)$ は真の値，c は定数である．いま x から $x+2h$ までを刻み幅 $2h$ で 1 ステップとして積分した値を $y_1(x+2h)$，刻み幅 h で 2 ステップで積分した値を $y_2(x+2h)$ とすると，以下の関係が成り立つ．

$$
\begin{aligned}
y_1(x+2h) &= y(x+2h) + c(2h)^5 + O(h^6)\\
y_2(x+2h) &= y(x+2h) + 2ch^5 + O(h^6)
\end{aligned}
$$

第二式に係数 2 が掛かっているのは積分を 2 ステップ行ったからである．上式から h^5 の項を消去すれば

$$
\begin{aligned}
y(x+2h) &= y_2(x+2h) + \frac{\delta}{15} + O(h^6)\\
\delta &= y_2(x+2h) - y_1(x+2h)
\end{aligned}
\qquad (15.14)
$$

が得られる．これはもともとの 4 段の公式よりも次数が一つ上がっている．しかし次数が上がったのが問題ではなく，δ を用いてステップ幅を調整できるできることの方が重要である．その方法は後で述べる．

フェールベルグ法　ルンゲ-クッタ法 (15.12)，ルンゲ-クッタ-ギル法 (15.13) はどちらも 4 段 ($p=4$)4 次の方法であった．これから述べるフェールベルグの公式は (15.11) 式で $p=6$ とした 6 段の公式

$$
\begin{aligned}
y_{n+1} &= y_n + \sum_{i=1}^{6} \gamma_i k_i + O(h^6)\\
k_i &= hf\big(x_n + \alpha_i h, y_n + \sum_{j=1}^{i-1} \beta_{ij} k_j\big)\\
i &= 1, 2, \cdots, 6
\end{aligned}
\qquad (15.15)
$$

ではあるが，実は 5 次の精度しかない．一般に p が 4 を越えると p 次の公式は存在せず，$p-1$ 次，最悪のときには $p-2$ 次の公式しか作れない．フェールベルグの方法のうまいところは，6 段のうちの 5 段を使って 5 次の公式を作ると同時に，4 段を用いて 4 次の公式も作れるように係数を選んであることである．4 次の公式による積分結果を

$$
y^*_{n+1} = y_n + \sum_{i=1}^{6} \gamma^*_i k_i + O(h^5) \qquad (15.16)
$$

と書くことにする．γ_i^* の中で 0 でないのは 4 項しかない．実際に必要なものは y_{n+1}^* そのものの値ではなく

$$\delta = y_{n+1} - y_{n+1}^* = \sum_i (\gamma_i - \gamma_i^*) k_i \qquad (15.17)$$

であるから，下表では γ_i^* の値ではなく，$\gamma_i - \gamma_i^*$ の値を示してある．

α_i	β_{ij}					γ_i	$\gamma_i - \gamma_i^*$
0						16/135	1/360
1/4	1/4					0	0
3/8	8/32	9/32				6656/12825	−128/4275
12/13	1932/2197	−7200/2197	7296/2197			28561/56430	−2197/75240
1	439/216	−8	3680/513	−845/4104		−9/50	1/50
1/2	−8/27	2	−3544/2365	1859/4104	−11/40	2/55	2/55

Table 15.1 Fehlberg 公式の係数

フェールベルグの公式では，表でも見られるように，$f(x, y)$ を計算する x の値が不規則な分布をしている．そのために $f(x, y)$ に含まれるパラメーターが表の形で与えられているときには，これを内挿する必要がある．単に不規則なだけではなく，Fig. 15.1 にみられるように偏っている．6 段の公式はフェールベルグ以来何人かの人たちによって求められているが，ここでは Cash-Karp の公式を示しておく．分点の分布はフェールベルグの公式よりも一様である．

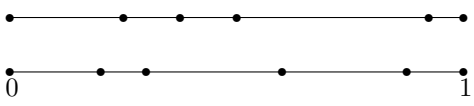


Fig. 15.1 分点 α_i の分布．上段：Fehlberg 法
下段：Cash-Karp 法

α_i	β_{ij}					γ_i	$\gamma_i - \gamma_i^*$
0						37/378	−277/64512
1/5	1/5					0	0
3/10	3/40	9/40				250/621	6925/370944
3/5	3/10	−9/10	6/5			125/594	−6925/202752
1	−11/54	5/2	−70/27	35/27		0	−277/14336
7/8	1631/55296	175/512	575/13824	44275/110592	253/4096	512/1771	277/7084

Table 15.2 Cash-Karp 公式の係数

ステップ幅の調節 (15.14)，(15.17) 式の δ はいずれも h^5 のオーダーの量である．これらは近似的に

$$\delta = ch^5 \qquad (15.18)$$

と書くことができる．

いま，y_{n+1} に許される誤差が δ_0 であったとする．もし

$$\delta \leq \delta_0$$

になったとすれば，y_{n+1} の誤差は許容範囲に入っているから，この y_{n+1} を積分の値として採用する．この場合，次の積分区間，x_{n+1} から x_{n+2} までのス

テップ幅は現在の幅 h よりも狭くとってもよいであろう．この幅を h_0 とする．次の積分の誤差がちょうど許容範囲 δ_0 になったとすれば

$$\delta_0 = ch_0^5$$

であるから，(15.18) 式を用いて c を消去すれば，次の積分ステップの幅 h_0 としては

$$h_0 = h\Big(\frac{\delta_0}{\delta}\Big)^{1/5} \qquad \delta \le \delta_0 \tag{15.19}$$

を選べばよいことがわかる．

問題は $\delta > \delta_0$ になってしまったときである．このときには y_{n+1} を積分値として採用することができない．改めて x_n から狭いステップ幅で積分をし直さなければならない．この幅を h_1 とすると，$y(x_n + h_1)$ まで積分したときの誤差は

$$\delta_1 = ch_1^5$$

となる．しかしこれを誤差の上限 δ_0 と等しいと置いて h_1 を求めることはできない．なぜなら，δ_0 は x_{n+1} における誤差の上限だからである．ステップ幅 h_1 で x_n から $x_n + h$ まで積分するには h/h_1 回の積分を要する．したがって $x_n + h$ における誤差は $\delta_1 \times (h/h_1)$ 程度である．これを δ_0 に等しいと置いたものが h_1 の上限を決める式となる．

$$\frac{h}{h_1}\delta_1 = chh_1^5 = \delta_0$$

c を消去すれば

$$h_1 = h\Big(\frac{\delta_0}{\delta}\Big)^{1/4} \qquad \delta > \delta_0 \tag{15.20}$$

が得られた．

(15.19) 式の h_0 は積分が成功して次のステップに進むときの幅，(15.20) 式の h_1 は積分が失敗してやり直すときの幅である．どちらも $y(x_n + h)$ の誤差が許される上限として見積もったのであるから，安全係数 0.9 位を掛けた方が安全である．

予測子-修正子法　これまでの公式は $y(x_n + h)$ を求めるのに $y(x_n)$ の値だけが必要であった．このような公式を自己出発型 (self-starting) という．この場合にはステップ幅 h はステップによって変化させてもよい．この柔軟性によって誤差の大きなところではステップ幅を小さくとって積分する，適応的積分が可能であった．

自己出発型の公式で積分を進めていくと，y_n が求まった段階で過去の値，$y_{n-1}, y_{n-2}, \cdots$ およびそれらに対応した $f(x, y)$ の値，$f_{n-1}, f_{n-2}, \cdots$ が得られている．したがってこれらの値を次のステップに利用しないという手はない．

微分方程式 (15.1) を形式的に積分すると

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} f(x, y)dx$$

となる．$f(x, y)$ をこれまでステップ幅 h で等間隔に計算された f_n, $f_{n-1}, \cdots$ などを用いてニュートン-コーツ型の公式によって内 (外) 挿して積分をおこなうと y_{n+1} 表すさまざまな公式が導かれる．4 次のアダムス-バシュフォースの公式は

$$y_{n+1} = y_n + \frac{h}{24}\big(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}\big) + O(h^5) \tag{15.21}$$

である．これは過去に計算された値だけを用いて y_{n+1} を予測するもので，予測子と呼ばれている．一方，アダムス-ムルトンの公式は

$$y_{n+1} = y_n + \frac{h}{24}\big(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}\big) + O(h^5) \tag{15.22}$$

で表される．ここで注意しなければならないのは，右辺の f_{n+1} は $f(x_{n+1}, y_{n+1})$ であるから，この式は y_{n+1} を未知数とする陰的な (implicit) 方程式になっていることである．これを解くためにはニュートン法などを用いなければならないが，通常は逐次代入法を用いた，予測子-修正子法で解く．

$$\mathrm{P}:\ y^*_{n+1} = y_n + \frac{h}{24}\big(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}\big) \tag{15.23}$$

$$\mathrm{E}:\ f^*_{n+1} = f(x_{n+1}, y^*_{n+1}) \tag{15.24}$$

$$\mathrm{C}:\ y_{n+1} = y_n + \frac{h}{24}\big(9f^*_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}\big) \tag{15.25}$$

$$\mathrm{E}:\ f_{n+1} = f(x_{n+1}, y^*_{n+1}) \tag{15.26}$$

P は**予測子** (predictor)，C は**修正子** (corrector)，E は微分の評価 (evaluation) を意味している．第一式の右辺はアダムス-バシュフォースの公式そのもので，これによって y_{n+1} の値を予測する．予測値 y^*_{n+1} を用いて f_{n+1} の暫定値 f^*_{n+1} を計算し，これをアダ

ムス-ムルトンの式に用いて y_{n+1} の値を修正する．最後に，次のステップに用いるために f_{n+1} を計算しておく．

陰方程式を反復代入法で解くときには (15.24), (15.25) 式の計算を一回だけでなく反復するまで何回も繰り返すので，アルゴリズムは記号的に P(EC)(EC)···E と表される．しかし予測子-修正子法では (15.23) 式から (15.26) 式までを一回づつ計算する，すなわち PECE で十分である．反復を繰り返すとかえって不安定になることがある．

中点差分法　ここでは $y(x)$ を点 x から点 $x+H$ まで積分することを考える．そのための公式に次のようなものがある．

$$
\begin{aligned}
&z_0 = y(x) \qquad h = \frac{H}{n} \\
&z_1 = z_0 + hf(x, z_0) \\
&z_{m+1} = z_{m-1} + 2hf(x+mh, z_m) \qquad (15.27) \\
&\qquad m = 1, 2, \cdots, n-1 \\
&y(x+H) \doteq y_n = \frac{1}{2}[z_n + z_{n-1} + hf(x+H, z_n)]
\end{aligned}
$$

ここでの h はこれまでの h と意味が違う．求めたいのは $y(x+H)$ であるから H がこれまでの h に相当する．上の第二式は微分方程式 (15.1) を中点公式

$$\frac{dy(x)}{dx} = \frac{y(x+h) - y(x-h)}{2h} + O(h^2)$$

を用いて差分化したものになっている．

中点公式は 2 次の近似であるが，H を固定したときの近似の誤差が

$$y_n - y(x+H) = \sum_{i=1}^{\infty} c_i h^{2i} \qquad (15.28)$$

のように h の偶数次の冪だけで表されることがわかっている．これは台形公式を用いた定積分のときと全く同じである．したがってロンバーグの方法を用いて精度を上げることができる．たとえば n が偶数のときに $y_{n/2}$ と y_n から

$$\frac{4y_n - y_{n/2}}{3}$$

を作ればこの誤差は h^4 から始まる．n を 2 倍ずつに増やしていけばロンバーグ法と全く同じ形式で正確な解が求められることになる．この方法によって大きな H に対しても積分が求められることになる．

ロンバーグ積分の場合には h を半分にしたとき前に計算した関数値がそのまま利用できるが，微分方程式のときには $f(x,y)$ の y も変化するから改めて計算し直さなければならない．したがって h を半分にしていくのは効率が悪い．それよりも n を

$$n = 2, 4, 6, 8, \cdots$$

の順に変えていくほうが効率がよい．これらの n に対する積分結果を h^2 の関数と考えて内挿公式を作り，$h^2 = 0$ の値を計算する．これはリチャードソンの補外法にほかならない．そのためには §5 で述べたエイトキンのアルゴリズムや，これと同等なネヴィルのアルゴリズムが便利である．

```
      Implicit Real*8 (a-h,o-z)
      Dimension x(50),y(50)
      x0 = 0
      y0 = 1
      yp0 = f(x0,y0)
      Do nn=1,10
        n = nn*2
        h = (xn-x0)/n
        x(nn) = h*h
        h2 = h*2
        z1 = y0
        z2 = z1+h*yp0
        Do m=1, n-1
          z0 = z1
          z1 = z2
          z2 = z0+h2*f(x0+m*h,z1)
        Enddo
        y(nn) = (z2+z1+h*f(xn,z2))/2
        If( nn.gt.1 ) then
          Do i=nn-1,1,-1
            y(i) = y(i+1)+(y(i+1)-y(i))
     *           *(x(nn)/(x(i)-x(nn)))
          Enddo
        Endif
        Write(6,6) n,(y(i),i=1,nn)
    6   Format(i5,1p5e18.10/(5x,1p5e18.10))
      Enddo
      Stop
      End
      Function f(x,y)
```

```
Real*8 f,x,y
f=-x*y
Return
End
```

中点差分法とリチャードソン補外のプログラム例

上のプログラムは初期値問題

$$\frac{dy}{dx} = -xy \quad y(0) = 1$$

を $H = 1$ として 1 ステップだけ積分するもので，`Do i=nn-1,1,-1` のループがネヴィルのアルゴリズムを用いたリチャードソン補外の部分である．y(1) に $h^2 = 0$ の補外値が入っている．この初期値問題の厳密解は

$$y(x) = e^{-x^2/2}$$

であるから

$$y(1) = e^{-1/2} = 0.606530659712\cdots$$

となるはずである．上のプログラムの出力は次のようになる．

n	$y(1)$	y_n	RKG
2	0.6250	0.6250	0.606494
4	0.6035156	0.60888	0.606531
6	0.6065243	0.60738	0.606531
8	0.6065285	0.606975	0.6065308
10	0.606530639	0.606803	0.60653072
12	0.60653065892	0.606716	0.60653069
14	0.60653065970	0.6066649	0.606530678
16	0.60653065971	0.6066325	0.606530670

$y(1)$ の欄は補外法で求めた値である．y_n は (15.27) 式で定義された値である．参考のために同じステップ幅 h でルンゲ-クッタ- ヂル法で $x = 1$ まで積分した値 (RKG) も示してある．中点差分法では $n = 14$ で既に有効数字 10 桁まで正しく求められているので，これ以上反復を繰り返してもあまり意味がない．y_n の収束が遅いにもかかわらず $y(1)$ の収束が速いのは，補外法の威力である．

ステルマーの方法　2 階以上の常微分方程式は 1 階の連立常微分方程式に変換して解くのが常道であるが

$$\begin{aligned} &\frac{d^2y}{dx^2} = f(x,y) \\ &y(x_0) = y_0 \quad y'(x_0) = y'_0 \end{aligned} \tag{15.29}$$

のように右辺に $y'(x) = dy(x)/dy$ が含まれない場合には前項と同様に差分近似で解くことができる．2 次微係数の差分近似

$$\frac{d^2y(x)}{dx^2} = \frac{y(x+h) - 2y(x) + y(x-h)}{h^2} + O(h^2)$$

を利用すれば次式が導かれる．

$$\begin{aligned} &z_0 = y_0 \quad z_1 = y_0 + h[y'_0 + \frac{1}{2}hf(x_0,y_0)] \\ &z_{m+1} - 2z_m + z_{m-1} = h^2 f(x_0 + mh, z_m) \\ &\qquad m = 1,2,\cdots,n-1 \\ &y_n = z_n \\ &y'_n = \frac{1}{h}(z_m - z_{m-1}) + \frac{1}{2}hf(x_0 + H, z_n) \end{aligned} \tag{15.30}$$

y_n が $y(x_0+H)$ の値，y'_n が x_0+H における 1 次微分の値である．これは 2 階の差分方程式であるが，差分 $\Delta_m = z_{m+1} - z_m$ を用いると 1 階の差分方程式で表すことができる．この方が丸め誤差が少ない．

$$\begin{aligned} &\Delta_0 = h[y'_0 + \frac{1}{2}hf(x_0,y_0)] \quad z_1 = y_0 + \Delta_0 \\ &\Delta_m = \Delta_{m-1} + h^2 f(x_0 + mh, z_m) \\ &z_{m+1} = z_m + \Delta_m \\ &\qquad m = 1,2,\cdots,n-1 \end{aligned} \tag{15.31}$$

この方法による y_n の誤差も (15.28) 式の形に書くことができ，前項と全く同様に補外法を用いて解の精度を上げることができる．

16 偏微分方程式の解法

偏微分方程式の分類 偏微分方程式は数学的性質によって，放物型，双曲型，楕円型の 3 タイプに分けることができる．これらは二次曲線の 3 タイプ，$y = x^2 + c$ (放物線)，$x^2 - y^2 = c$ (双曲線)，$x^2 + y^2 = c$ (楕円) に対応している．

放物型の代表は 1 次元の拡散方程式

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} \tag{16.1}$$

である．D は拡散係数を表す．この方程式を解くためには，時刻 t_0 における初期条件 $u(x, t_0)$ と，領域の端における境界条件が必要である．これらが与えられたとき，その後の $u(x, t)$ の時間発展を解くことができる．すなわち (16.1) 式は初期値問題として解くことができる．

双曲型の代表は波動方程式

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2} \tag{16.2}$$

である．c は波の伝播速度を表す．この方程式も t_0 における $u(x, t_0)$ およびその微分 $\partial u/\partial t$ と境界条件を与えて時間発展を追うことができる．すなわち (16.2) 式も初期値問題である．

楕円型の方程式の一つ，ポアソンの方程式

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \rho(x, y) \tag{16.3}$$

は上二つとは異なり，境界値問題である．すなわち，(x, y) 平面上の与えられた領域で上式を満足し，領域の境界で境界条件を満足する解を求める，という問題である．境界条件としては，境界で $u(x, y)$ の値そのものが与えられている場合（ディリクレ問題）と，境界に垂直方向の微係数が与えられている場合（ノイマン問題）が代表的である．

拡散方程式 以下では x 方向には間隔 Δx で格子点を取り，時間方向には間隔 Δt で格子点を取る．格子点上の $u(x, t)$ の値を

$$u_j^n = u(j\Delta x, n\Delta t) \quad 0 \le j \le J$$

と表す．簡単のためにここでは境界値 u_0^n, u_J^n が与えられているものとする．拡散方程式 (16.1) を時間，空間で差分化するために，時間微分については前進差分

$$\left(\frac{\partial u}{\partial t}\right)_j^n = \frac{u_j^{n+1} - u_j^n}{\Delta t} + O(\Delta t) \tag{16.4}$$

空間微分については中心差分

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_j^n = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2} + O(\Delta x^2) \tag{16.5}$$

によって近似する．そうすると (16.1) 式は

$$u_j^{n+1} = u_j^n + \beta(u_{j+1}^n - 2u_j^n + u_{j-1}^n) \qquad 0 < j < J \tag{16.6}$$

$$\beta = \frac{D\Delta t}{(\Delta x)^2} \tag{16.7}$$

となる．右辺は時刻 n における値だけで書かれているので，この式から次の時刻 $n+1$ における値を陽に求めることができる．

拡散方程式の解は $t \to \infty$ で発散することはあり得ない．(16.6) 式から計算される解が $t \to \infty$ で安定であるかどうかを見るために，境界条件を無視して

$$u_j^n = \lambda^n e^{ikj\Delta x} \tag{16.8}$$

の形の解を仮定する．k は x 方向の波数である．これを (16.6) 式に代入すると

$$\lambda = 1 - 4\beta\sin^2\frac{k\Delta x}{2} \tag{16.9}$$

が得られる．λ は 1 ステップ当たりの増幅倍率であるから，解が安定であるためには

$$-1 \le \lambda \le 1$$

でなければならない．(16.9) 式からこの条件がすべての波数 k に対して成り立つためには

$$0 \le \beta = \frac{D\Delta t}{(\Delta x)^2} \le \frac{1}{2} \tag{16.10}$$

でなければならない．この関係は安定な解が求められるためには Δt と Δx とを独立に選ぶことができないことを意味している．空間方向の微分 (16.5) 式の精度を上げようとして，たとえば Δx を半分にしたとすれば，(16.10) 式を満足させるためには Δt を 4 分の 1 にしなければならない．また左側の不等式で等号のときは $\Delta t = 0$ であるから意味がないが，$\Delta t > 0$ でなければならないということは，拡散方程式 (16.1) を時間の逆方向に解くことができないことを意味している．

条件 (16.10) は実は非常に厳しいものである．拡散が距離 l だけ進行する時間，すなわち拡散時間は

$$\tau = \frac{l^2}{D}$$

で表される．この時間内の積分ステップの数は

$$\frac{\tau}{\Delta t} = \frac{l^2}{D\Delta t} = \frac{(\Delta x)^2}{2D\Delta t}\cdot\frac{2l^2}{(\Delta x)^2}$$

であるから，条件 (16.10) を用いると

$$\frac{\tau}{\Delta t} \geq 2\Big(\frac{l}{\Delta x}\Big)^2$$

となる．われわれが空間スケール l の現象をシミュレーションしようとしているなら，格子間隔 Δx は l よりも十分小さくとらなければならないから

$$l \gg \Delta x$$

が成り立っているはずである．したがって $\tau/\Delta t$ は非常に大きくなってしまう．そこで Δt が大きくても安定な方法が必要になる．

陰解法　(16.1) 式を差分化するのに (16.6) 式が唯一つの方法ではない．ここでは (16.1) 式の右辺を時刻 n で評価しているが，時間の一階微分 (16.4) 式は，時刻 n と $n+1$ の中間 $n+1/2$ における微分と考えた方がよい．これは

$$\left(\frac{\partial u}{\partial t}\right)_j^{n+1/2} = \frac{u_j^{n+1}-u_j^n}{\Delta t} + O(\Delta t^2) \qquad (16.11)$$

であるから，(16.4) 式よりも精度が高い．そこで空間微分も時刻 n と $n+1$ の加重平均と考えて

$$\frac{\partial^2 u}{\partial x^2} \doteq \frac{1}{(\Delta x)^2}\big[\theta(u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}) + (1-\theta)(u_{j+1}^n-2u_j^n+u_{j-1}^n)\big] \qquad (16.12)$$

とする．θ は重みを表すパラメーターで

$$0 \leq \theta \leq 1$$

である．$\theta = 0$ が先の (16.6) 式である．この近似を用いて拡散方程式を差分化すると

$$u_j^{n+1} = u_j^n + \beta\big[\theta\big(u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}\big) + (1-\theta)\big(u_{j+1}^n-2u_j^n+u_{j-1}^n\big)\big] \qquad (16.13)$$

が得られる．求めたいのは u_j^{n+1} であるが，右辺にも時刻 $n+1$ における量が現れているので，この式は u_j^{n+1} を陽に決める式ではなく，陰に決める式になっている．この形ではわかりにくいので未知数を左辺に移項すると

$$-\beta\theta u_{j+1}^{n+1} + (1+2\beta\theta)u_j^{n+1} - \beta\theta u_{j-1}^{n+1} = \beta(1-\theta)\big(u_{j+1}^n - 2u_j^n + u_{j-1}^n\big) \qquad (16.14)$$

となる．右辺は既知である．左辺には求めようとする時刻 $n+1$ における三つの格子点 j+1, j, j–1 における値が現れている．したがって上式は u_j^{n+1}, $j = 1, 2, \cdots, J$–1 に関する連立一次方程式の形になっている．しかしこれは対称で，しかも優対角の三重対角方程式であるから §4 で述べた方法で簡単に解くことができる．

解の安定性を見るために，(16.8) 式のように仮定してこれを (16.13) 式に代入すると

$$\lambda = 1 - 4\beta[\lambda\theta + (1-\theta)]\sin^2\frac{k\Delta x}{2}$$

すなわち

$$\lambda = \frac{1-4\beta(1-\theta)\sin^2 k\Delta x/2}{1+4\beta\theta\sin^2 k\Delta x/2} \qquad (16.15)$$

が得られる．$|\lambda| \leq 1$ となるためには

$$0 \leq \beta\sin^2\frac{k\Delta x}{2}$$
$$2\beta(1-2\theta)\sin^2\frac{k\Delta x}{2} \leq 1 \qquad (16.16)$$

でなければならない．第一式は $\beta \geq 0$ であるから (16.10) 式の左辺と変わらない．第二式は θ の値によって二つの場合にわかれる．

まず $\theta < 1/2$ のときには

$$\beta \leq \frac{1}{2(1-2\theta)} \qquad 0 \leq \theta < \frac{1}{2} \qquad (16.17)$$

でなければならない．これは β に上限があるという意味では (16.10) 式と同じである．この上限は θ が増えるにつれて増加し，$\theta = 1/2$ で無限大になる．これは (16.16) 式の第二式が $\theta \geq 1/2$ では常に成立することに対応している．すなわち，$\theta \geq 1/2$ のときには陰解法 (16.14) 式の解は無条件で安定である．特に $\theta = 1/2$ のときを**クランク-ニコルソンの方法**という．このような陰解法を用いれば Δt を大きくとっても安定な解を求めることができる．

移流項（風上差分） 流体力学における一次元のナヴィエ-ストークスの方程式は

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = D\frac{\partial^2 v}{\partial x^2} \tag{16.18}$$

で表される．左辺の第二項がなければこれは拡散方程式である．この項は移流項と呼ばれ，非線型効果を表している．移流項は波数の小さな成分から波数の大きな成分へエネルギーを運ぶ働きをするので，流体力学的な不安定を引き起こすと同時に，数値的な不安定を引き起こす要因ともなる．

移流項の影響だけを見るために，拡散項は十分に小さいとしてこれを無視した方程式

$$\frac{\partial u}{\partial t} = -v\frac{\partial u}{\partial x} \tag{16.19}$$

を考えてみる．v はここでは定数である．陽解法をとることにし，左辺は前進差分 (16.4) 式，右辺は時刻 n で評価することにする．もっとも単純なのは右辺を中央差分で置きかえる方法

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = -v\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} \tag{16.20}$$

である．しかし (16.8) 式を上式に代入して増幅倍率を求めると

$$\lambda = 1 - i\frac{v\Delta t}{\Delta x}\sin k\Delta x$$

であるから，$\Delta t = 0$ でない限り $|\lambda| > 1$ となって (16.20) 式は無条件に不安定である．

陽解法でも安定な方法はいくつかあるが，その中でも広く用いられているのが風上差分 (upwind differencing) という方法である．移流項は風上の流体が風下の流体に影響を及ぼすことによって生じるのであって，その逆ではない．そこで v の正負によって移流項の差分のとり方を変えて (16.20) 式を

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \begin{cases} -v\dfrac{u_j^n - u_{j-1}}{\Delta x} & v > 0 \\ -v\dfrac{u_{j+1}^n - u_j^n}{\Delta x} & v < 0 \end{cases} \tag{16.21}$$

で近似する．このときの増幅倍率は，少し面倒な計算になるが

$$\lambda = 1 - \left|\frac{2v\Delta t}{\Delta x}\right|\sin^2\frac{k\Delta x}{2} - i\frac{2v\Delta t}{\Delta x}\sin\frac{k\Delta x}{2}\cos\frac{k\Delta x}{2}$$

したがって

$$|\lambda|^2 = 1 - 4\left|\frac{v\Delta t}{\Delta x}\right|\left(1 - \left|\frac{v\Delta t}{\Delta x}\right|\right)\sin^2\frac{k\Delta x}{2}$$

となる．よって (16.21) 式が安定であるためには

$$\left|\frac{v\Delta t}{\Delta x}\right| < 1 \tag{16.22}$$

でなければならない．これを**クーランの条件**という．

それでは拡散項を含んだ方程式 (16.18) はどうやって解くのか．それについては次項で述べる．

二次元の拡散方程式 (ADI 法) 二次元の拡散方程式は

$$\frac{\partial u}{\partial t} = D\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \tag{16.23}$$

である．$u(x, y, t)$ を x 方向，y 方向ともに Δx で離散化して

$$u_{j,l}^n = u(j\Delta x, l\Delta x, n\Delta t)$$

と書くことにする．x 方向，y 方向の二次微分に (16.5) 式を用いて陽解法の式を作るのは容易であるからここには書かない．安定条件は (16.10) 式にかわって

$$0 \le \frac{D\Delta t}{(\Delta x)^2} \le \frac{1}{4} \tag{16.24}$$

になる．

クランク-ニコルソンの方法を二次元に拡張するのも形式的には容易である．表記を簡単にするために二階差分を定義する．

$$\delta_x^2 u_{j,l}^n = u_{j+1,l}^n - 2u_{j,l}^n + u_{j-1,l}^n$$
$$\delta_y^2 u_{j,l}^n = u_{j,l+1}^n - 2u_{j,l}^n + u_{j,l-1}^n$$

これを用いると二次元のクランク-ニコルソンの方法は

$$u_{j,l}^{n+1} = u_{j,l}^n + \frac{1}{2}\beta\big[\delta_x^2 u_{j,l}^{n+1} + \delta_x^2 u_{j,l}^n + \delta_y^2 u_{j,l}^{n+1} + \delta_y^2 u_{j,l}^n\big] \tag{16.25}$$

と表すことができる．この式は一次元のときと異なり，x 方向と y 方向の未知数がカップルしているので，簡単に解くことはできない．しかし連立方程式の係数行列は非常に疎であるから，この性質を利用することはできる．

(16.25) 式を解くときに最もよく用いられる方法は ADI 法 (Alternating- Direction Implicit method,

交互方向法) である．この方法は時間ステップ Δt を二つにわけて一方向ごとに時間を進めるものである．すなわち

$$
\begin{aligned}
u_{j,l}^{n+1/2} &= u_{j,l}^{n} + \frac{1}{2}\beta\big(\delta_x^2 u_{j,l}^{n+1/2} + \delta_y^2 u_{j,l}^{n}\big) \\
u_{j,l}^{n+1} &= u_{j,l}^{n+1/2} + \frac{1}{2}\beta\big(\delta_x^2 u_{j,l}^{n+1/2} + \delta_y^2 u_{j,l}^{n+1}\big)
\end{aligned}
\tag{16.26}
$$

である．第一式では $u_{j,l}^{n+1/2}$ が未知数であり，ある l を固定すればこの式は $u_{j+1,l}^{n+1/2}$, $u_{j,l}^{n+1/2}$, $u_{j-1,l}^{n+1/2}$ を未知数とする三重対角方程式である．これを解いて第二式に代入すれば，これは $u_{j,l+1}^{n+1}$, $u_{j,l}^{n+1}$, $u_{j,l-1}^{n+1}$ に関する三重対角方程式になる．

この方法は演算子を分割して，それぞれに対して別々に時間発展を行わせるという方法の変形である．たとえばナヴィエ-ストークスの方程式 (16.18) 式を移流項と拡散項に分割して，前者は陽解法，後者はクランク-ニコルソンの陰解法で解くことができる．

$$
\begin{aligned}
u_j^{n+1/2} &= u_j^n - \frac{v\Delta t}{\Delta x}\begin{cases} (u_j^n - u_{j-1}^n) & v > 0 \\ (u_{j+1}^n - u_j^n) & v < 0 \end{cases} \\
u_j^{n+1} &= u_j^{n+1} + \frac{1}{2}\beta\big(\delta^2 u_j^{n+1} + \delta^2 u_j^{n+1/2}\big) \\
\delta^2 u_j^n &= u_{j+1}^n - 2u_j^n + u_{j-1}^n
\end{aligned}
\tag{16.27}
$$

第一式の n から $n+1/2$ のステップ幅は Δt，第二式のステップ幅も Δt である．このステップ幅はクーランの条件 (16.22) 式を満たしていなければならない．上式では時間が $2\Delta t$ 進んでしまうようにみえるが，そうではない．移流項，拡散項それぞれが Δt だけ進むのであるから，全体としてはやはり Δt しか時間は進まないからである．

これに対して ADI 法 (16.26) 式では第一式では時間は $\Delta t/2$ 進み，第二式でも $\Delta t/2$ 進む．ここでは両式ともに拡散項が x 成分 y 成分ともに含まれているから，ここは $\Delta t/2$ でなければならないのである．

波動方程式　(16.2) 式を時間，空間について (16.5) 式によって差分化すれば

$$
\begin{aligned}
&u_j^{n+1} - 2u_j^n + u_j^{n-1} \\
&\quad = \alpha^2\big(u_{j+1}^n - 2u_j^n + u_{j-1}^n\big)
\end{aligned}
\tag{16.28}
$$

$$
\alpha = \frac{c\Delta t}{\Delta x} \tag{16.29}
$$

が得られる．時刻 n までの解が得られていたとすれば上式は u_j^{n+1} を決める式である．

上式の安定性を調べるために，u_j^n を

$$
u_j^n = e^{i(kj\Delta x - \omega n\Delta t)} \tag{16.30}
$$

と置く．ω は波の角周波数，k は波数である．これを (16.28) 式に代入すると

$$
\sin^2\frac{\omega\Delta t}{2} = \alpha^2 \sin^2\frac{k\Delta x}{2}
$$

が得られるが

$$
\sin\frac{\omega\Delta t}{2} = \alpha \sin\frac{k\Delta x}{2} \tag{16.31}
$$

だけを考えれば十分である．

(16.31) 式を波数 k を与えて角振動数 ω を求める式と考えてみる．いま

$$
\alpha \geq 1
$$

なら，(16.31) 式から，ある波数以上で

$$
\sin\frac{\omega\Delta t}{2} > 1
$$

となる．このときには (16.31) 式には

$$
|e^{-i\omega\Delta t}| > 1
$$

となる根があり，したがって (16.28) 式の解は不安定になる．すなわち，解が安定であるためには

$$
\alpha = \frac{c\Delta t}{\Delta x} \leq 1 \tag{16.32}
$$

でなければならない．

この条件の意味は単純である．(16.28) 式からは数値解では波は 1 ステップごとに 1 格子点以上進むことはできない．一方，物理的には 1 ステップで波は $c\Delta t$ だけ進む．したがって $c\Delta t > \Delta x$ なら，数値解の進行する速さが物理的な波の速さに追従することができなくなって，エネルギーが数値解の波面の後側に集積して解が発散するのである．

数値分散　波動方程式 (16.2) の解の波面は一定速度 c で伝播する．一方，差分方程式 (16.28) の解の波面の伝播速度を $\hat{c}$ とすると，(16.31) 式の解を用いて

$$
\hat{c} = \frac{\omega}{k} = \frac{2}{k\Delta t}\sin^{-1}\Big(\alpha \sin\frac{k\Delta x}{2}\Big)
$$

であるから

$$\frac{\hat{c}}{c} = \Big(\frac{k\Delta x}{2}\alpha\Big)^{-1} \sin^{-1}\Big(\alpha \sin\frac{k\Delta x}{2}\Big) \quad (16.33)$$

で表される．この式は差分解の伝播速度 $\hat{c}$ が波数 k によって変化してしまうことを表している．

波動方程式 (16.2) 式の解の伝播速度は本来，波の周波数，あるいは波長によって変化しない，したがって分散しないはずである．しかしこの方程式を差分化した (16.28) 式の解は上式に見られるように分散してしまう．これを数値分散という．Fig. 16.1 に (16.33) 式で計算した数値分散の分散曲線を示してある．横軸は波の波長 λ と空間格子間隔 Δx の比 $\Delta x/\lambda$，縦軸は $\hat{c}/c$，数字は α の値である．波長の短い波ほど数値分散の量が大きくなる．数値解の解釈にはこの点に注意しなければならない．

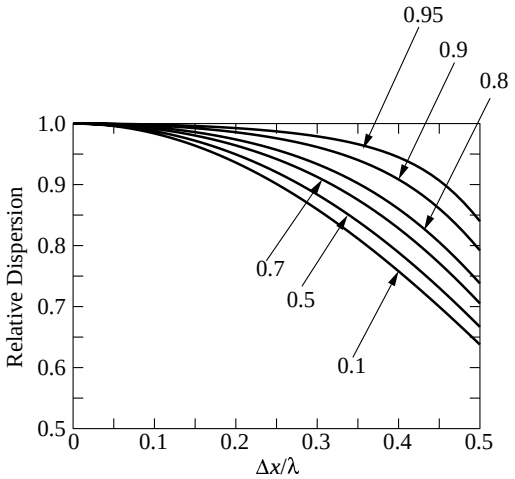


Fig. 16.1 数値分散．λ は波長，縦軸は $\hat{c}/c$，数字は $\alpha = c\Delta t/\Delta x$.

境界値問題　境界値問題の例として，ポアソンの方程式 (16.3) を矩形領域

$$0 \le x \le J\Delta x \quad 0 \le y \le L\Delta x$$

で解くことを考える．境界条件としては最も簡単なディリクレの条件，すなわち境界で関数値 $u(x, y)$ が与えられているとする．

二次微分を (16.5) 式で近似すると，(16.3) 式は

$$\begin{aligned} &u_{j+1,l} + u_{j-1,l} + u_{j,l+1} + u_{j,l-1} \\ &\quad -4u_{j,l} = (\Delta x)^2 \rho_{j,l} \\ &\quad 0 < j < J \quad 0 < l < L \end{aligned} \quad (16.34)$$

で近似される．この式は (j, l) が境界の内側の内点に対して成り立つ．一方，境界においては $u_{j,l}$ の値が与えられているから，未知数も内点の $u_{j,l}$ である．したがって先の式は未知数の数と同じだけあり，解は一義的に決まるはずである．この連立方程式の係数行列は一行に 5 個の成分だけが非零で残りの成分はすべて 0 という，非常に疎な行列である．

例として $J = L = 4$ の場合を考えてみる．下図のように内部の格子点に 1 から 9 までの番号を付ける．

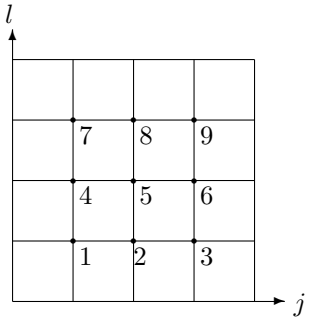


Fig. 16.2 格子点の番号付け

この格子点における $u_{j,l}$ の値からなるベクトルを $\boldsymbol{x}$ とするとき，(16.34) 式に対応する連立方程式は

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \quad (16.35)$$

と書くことができる．$\boldsymbol{b}$ は境界値やソース項 $\rho_{j,l}$ などからなる既知のベクトルである．このときの係数行列 $\boldsymbol{A}$ は次のような形になる．

$$\boldsymbol{A} = \begin{bmatrix} -4 & 1 & 0 & 1 & & & & & \\ 1 & -4 & 1 & 0 & 1 & & & & \\ 0 & 1 & -4 & 0 & 0 & 1 & & & \\ 1 & 0 & 0 & -4 & 1 & 0 & 1 & & \\ & 1 & 0 & 1 & -4 & 1 & 0 & 1 & \\ & & 1 & 0 & 1 & -4 & 0 & 0 & 1 \\ & & & 1 & 0 & 0 & -4 & 1 & 0 \\ & & & & 1 & 0 & 1 & -4 & 1 \\ & & & & & 1 & 0 & 1 & -4 \end{bmatrix} \quad (16.36)$$

数字の入っていないところは 0 を意味する．この行列は優対角であるから正則であり，したがって (16.34) 式の解は存在する．(16.34) 式の形から明らかなように，係数行列の対角要素は -4，その上下の副対角要素は 0 または 1 であり，この部分だけをみれば

対称な三重対角行列になっている．この例では領域のサイズが小さいため，三重対角から一つおいた副対角に 1 が入っている．領域が大きくなったときの行列の様子はこの例から容易に想像できる．

反復解法　この例のように疎な行列を係数とする連立方程式を解くときには反復解法が用いられる．まず行列 $\boldsymbol{A}$ を

$$\boldsymbol{A} = \boldsymbol{E} + \boldsymbol{F} \tag{16.37}$$

のように分解する．ここに $\boldsymbol{E}$ は方程式 $\boldsymbol{Ex} = \boldsymbol{c}$ が簡単に解ける部分，$\boldsymbol{F}$ は残りである．上の例では $\boldsymbol{E}$ として三重対角の部分を選べばよい．また，後で示すように，$\boldsymbol{E}$ として $\boldsymbol{A}$ の対角成分を選べば計算は非常に簡単になる．こうすると元の方程式 (16.35) 式は

$$\boldsymbol{Ex} = -\boldsymbol{Fx} + \boldsymbol{b}$$

となる．そこで初期ベクトル $\boldsymbol{x}^{(0)}$ を適当に選んで

$$\boldsymbol{Ex}^{(k)} = -\boldsymbol{Fx}^{(k-1)} + \boldsymbol{b} \quad k = 1, 2, \cdots \tag{16.38}$$

から次の近似解 $\boldsymbol{x}^{(k)}$ を求める，という反復法が導かれる．

k ステップ目の近似解と真の解との差を

$$\boldsymbol{e}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x} \tag{16.39}$$

とすれば，(16.38) 式から

$$\boldsymbol{Ee}^{(k)} = -\boldsymbol{Fe}^{(k-1)}$$

となるから，反復行列を

$$\boldsymbol{B} = -\boldsymbol{E}^{-1}\boldsymbol{F}$$

と置けば

$$\boldsymbol{e}^{(k)} = \boldsymbol{Be}^{(k-1)} \tag{16.40}$$

したがって

$$\boldsymbol{e}^{(k)} = \boldsymbol{B}^k \boldsymbol{e}^{(0)}$$

が成り立つ．$\boldsymbol{e}^{(k)}$ が $k \to \infty$ で零ベクトルに収束するためには，反復行列 $\boldsymbol{B}$ の固有値を λ_i とすると，すべての固有値が

$$|\lambda_i| < 1$$

を満たさなければならない (§6)．また収束のスピードは絶対値最大の固有値に左右される．一般に，ベクトル $\boldsymbol{B}$ の固有値 λ_i の絶対値の最大

$$\rho(\boldsymbol{B}) = \max_i |\lambda_i| \tag{16.41}$$

を $\boldsymbol{B}$ のスペクトル半径という．(16.38) 式では一回の反復で $\boldsymbol{e}^{(k)}$ のノルムはおよそ $\rho(\boldsymbol{B})$ 倍になる．したがってスペクトルノルムは収束の速さを表す指標となる．

緩和法　ポアソンの方程式 (16.3) は，ソース項を含む拡散方程式

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - \rho \tag{16.42}$$

の，十分時間がたった後の定常解

$$\frac{\partial}{\partial t} u(x, y, t) = 0 \quad (\text{すべての } x, y \text{ に対して})$$

と考えることもできる．たとえば最も単純な陽解法をとることにすれば

$$u_{j,l}^{n+1} = u_{j,l}^n + \frac{\Delta t}{(\Delta x)^2}\big[u_{j+1,l}^n + u_{j-1,l}^n + u_{j,l+1}^n + u_{j,l-1}^n - 4u_{j,l}^n - (\Delta x)^2 \rho_{j,l}\big]$$

である．$\Delta t/(\Delta x)^2$ は安定条件 (16.24) 式を満たしていなければならない．しかし定常状態だけが問題であるから，Δt として許される最大値を用いると $\Delta t/(\Delta x)^2 = 1/4$ であるから，反復の式は

$$u_{j,l}^{n+1} = \frac{1}{4}\Big[u_{j+1,l}^n + u_{j-1,l}^n + u_{j,l+1}^n + u_{j,l-1}^n - (\Delta x)^2 \rho_{j,l}\Big] \tag{16.43}$$

となる．このような方法を緩和法 (relaxation method) という．

先にも注意しておいたように，(16.42) 式が完全に緩和するためには非常に長い時間がかかる．そこで時間刻み Δt を大きく取るために ADI 法を用いることも考えられる．

緩和法は別の観点からも導くことができる．行列 $\boldsymbol{A}$ を

$$\boldsymbol{A} = \boldsymbol{D} + \boldsymbol{L} + \boldsymbol{U} \tag{16.44}$$

と分解する．ここに $\boldsymbol{D}$ は対角行列，$\boldsymbol{U}$ は上三角行列，$\boldsymbol{L}$ は下三角行列である．ここで $\boldsymbol{E} = \boldsymbol{D}$ と選べば，反復 (16.38) 式は

$$\boldsymbol{Dx}^{(k)} = -(\boldsymbol{L} + \boldsymbol{U})\boldsymbol{x}^{(k-1)} + \boldsymbol{b} \tag{16.45}$$

となる．$\boldsymbol{D}$ は対角行列であるから，この計算は単純である．これを**ヤコビの方法**という．

行列で書くと複雑に見えるが，成分に分けて書けば簡単である．(16.34) 式の場合には

$$u_{j,l}^{(k)} = \frac{1}{4}\Big[u_{j+1,l}^{(k-1)} + u_{j-1,l}^{(k-1)} + u_{j,l+1}^{(k-1)} + u_{j,l-1}^{(k-1)} - (\Delta x)^2\rho_{j,l}\Big] \tag{16.46}$$

にほかならない．いいかえれば，周囲の 4 点の平均値をもって新しい $u_{j,l}^{(k)}$ の値とするものである．これは拡散方程式から導いた (16.43) 式と全く同じである．

上式ではすべての (j,l) についての $u_{j,l}^{(k)}$ が求められてから次のステップに進むという方法をとっている．しかしある格子点について新しい $u_{j,l}^{(k)}$ が求められたら，これを別の格子点の $u^{(k)}$ を計算するのに用いても差し支えないだろう．いいかえれば，$u_{j,l}^{(k)}$ を $u_{j,l}^{(k-1)}$ の上に上書きしてしまうのである．Fig. 16.2 のように行方向に計算を行なうとき，これは上式のかわりに

$$u_{j,l}^{(k)} = \frac{1}{4}\Big[u_{j+1,l}^{(k-1)} + u_{j-1,l}^{(k)} + u_{j,l+1}^{(k-1)} + u_{j,l-1}^{(k)} - (\Delta x)^2\rho_{j,l}\Big] \tag{16.47}$$

を用いることに相当する．この方法を**ガウス-ザイデルの方法**という．これは行列で書くと

$$(\boldsymbol{D}+\boldsymbol{L})\boldsymbol{x}^{(k)} = -\boldsymbol{U}\boldsymbol{x}^{(k-1)} + \boldsymbol{b} \tag{16.48}$$

である．

ヤコビ法は収束が非常に遅い．ヤコビ法の反復行列は (16.45) 式から

$$\boldsymbol{B} = -\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})$$

である．この行列のスペクトル半径をヤコビ法のスペクトル半径という．矩形領域でディリクレ型の境界条件のとき，ヤコビ法のスペクトル半径 ρ_{J} は

$$\rho_{\mathrm{J}} = \frac{1}{2}\Big(\cos\frac{\pi}{J} + \cos\frac{\pi}{L}\Big) \tag{16.49}$$

で与えられる．ガウス-ザイデル法のスペクトル半径 ρ_{GS} はヤコビ法のスペクトル半径の二乗

$$\rho_{\mathrm{GS}} = \rho_{\mathrm{J}}^2$$

である．

先にあげた $J = L = 4$ に対する 9×9 の行列 $\boldsymbol{A}$ に対するヤコビ法のスペクトル半径は (16.49) 式から

$$\rho_{\mathrm{J}} = \frac{1}{\sqrt{2}}$$

であるから，ヤコビ法では反復 2 回で誤差はやっと半分にしかならない．J が非常に大きい正方格子の場合 (16.49) 式から

$$\rho_{\mathrm{J}} = \cos\frac{\pi}{J} \simeq 1 - \frac{\pi^2}{2J^2}$$

であるから，ρ_{J} は 1 に非常に近くなり，収束は遅くなる．誤差が $1/e = 0.367\cdots$ になるのに必要な反復回数を N_e とすれば

$$\rho_{\mathrm{J}}^{N_e} = e^{-1} \qquad N_e = -\frac{1}{\ln\rho_{\mathrm{J}}}$$

である．J が非常に大きいとき

$$\ln\rho_{\mathrm{J}} \simeq \ln\big(1 - \frac{\pi^2}{2J^2}\big) \simeq -\frac{\pi^2}{2J^2}$$

であるから

$$N_e \simeq \frac{2}{\pi^2}J^2$$

となる．これはヤコビ法の反復回数が $O(J^2)$ であることを意味している．ガウス-ザイデル法はヤコビ法に比べて速いといっても二倍速いだけである．したがって J が非常に小さな場合以外は両方法ともあまり実用的ではない．

SOR 法　ガウス-ザイデル法 (16.48) 式の右辺を書きかえると

$$\begin{aligned}(\boldsymbol{D}+\boldsymbol{L})\boldsymbol{x}^{(k)} &= (\boldsymbol{D}+\boldsymbol{L})\boldsymbol{x}^{(k-1)} \\ &\quad -(\boldsymbol{D}+\boldsymbol{L}+\boldsymbol{U})\boldsymbol{x}^{(k-1)} + \boldsymbol{b} \\ &= (\boldsymbol{D}+\boldsymbol{L})\boldsymbol{x}^{(k-1)} - \boldsymbol{A}\boldsymbol{x}^{(k-1)} + \boldsymbol{b}\end{aligned}$$

となる．近似解 $\boldsymbol{x}^{(k)}$ の残差を

$$\boldsymbol{r}^{(k)} = \boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b} \tag{16.50}$$

で定義すると，先の式は

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} - (\boldsymbol{D}+\boldsymbol{L})^{-1}\boldsymbol{r}^{(k-1)} \tag{16.51}$$

となる．第二項が補正項という意味をもっていることがわかる．

そこで補正項を ω 倍した

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} - \omega(\boldsymbol{D}+\boldsymbol{L})^{-1}\boldsymbol{r}^{(k-1)} \tag{16.52}$$

を考える．ω は overrelaxation parameter と呼ばれ，上の方法は SOR(successive overrelaxation) 法と呼ばれる．この反復は一般的に $0 < \omega < 2$ のときに収束することが知られており，(16.34) 式のような場合には $1 < \omega < 2$，すなわち過補正のときに収束が最も速くなることが知られている．

行列の表現 (16.51) 式はわかりにくいので，再び (16.34) 式の場合の計算式を書き下してみる．常に上書きすることにすれば，反復の添字はかえって邪魔になる．最新の値を用いて残差を

$$r_{j,l} = u_{j+1,l} + u_{j-1,l} + u_{j,l+1} + u_{j,l-1} - 4u_{j,l} - (\Delta x)^2 \rho_{j,l} \tag{16.53}$$

と定義すると，(16.51) 式は簡単に

$$u_{j,l}^{\text{new}} = u_{j,l} + \frac{\omega}{4} r_{j,l} \tag{16.54}$$

となる．$\omega = 1$ のときには (16.47) 式に一致する．

SOR 法では最適な ω を用いることが収束にとって非常に重要である．ヤコビ法のスペクトル半径 ρ_J がわかっているとき，最適な ω は

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \rho_{\text{J}}^2}} \tag{16.55}$$

であり，この ω を用いたときの SOR 法のスペクトル半径は

$$\rho_{\text{SOR}} = \left(\frac{\rho_{\text{J}}}{1 + \sqrt{1 - \rho_{\text{J}}^2}} \right)^2 \tag{16.56}$$

で与えられる．$J \times J$ の格子の場合，先の近似を用いれば

$$\omega_{\text{opt}} \simeq \frac{2}{1 + \pi/J}$$
$$\rho_{\text{SOR}} \simeq 1 - \frac{2\pi}{J}$$

で与えられる．したがって SOR の N_e は

$$N_e \simeq \frac{1}{2\pi} J$$

である．これは SOR 法の収束の速さがガウス-ザイデル法に比べて J 倍速いことを意味している．

17 共役勾配法

理学や工学で現れる大規模な連立一次方程式は係数行列が疎であることが多い．このようなときに消去法のような標準的な方法を用いるのは，メモリーの点からも計算時間の点からも現実的ではない．係数行列が三重対角のように対称性が高い場合には，その性質を最大限に利用して計算を行なわなければならない．本節で述べる共役勾配法 (Conjugate Gradient method, CG 法) は，収束はあまり速くはないが，計算が行列の列あるいは行とベクトルの内積だけで済むので，係数の格納法に工夫をすれば，対称性のあまりよくない一般的な疎行列に対して適用することができる．CG 法にはさまざまなアルゴリズムがあるが，ここでは代表的なものを二，三あげるにとどめる．

対称行列 $n \times n$ 行列 $\boldsymbol{S}$ は正値対称行列であるとする．すなわち $\boldsymbol{S}$ は

$$\boldsymbol{S}^T = \boldsymbol{S}$$

を満足し，任意の零でない n 次元ベクトル $\boldsymbol{a}$ に対して

$$\boldsymbol{a}^T \boldsymbol{S} \boldsymbol{a} > 0 \tag{17.1}$$

を満たしているとする．$\boldsymbol{S}$ を係数行列とする連立方程式

$$\boldsymbol{S}\boldsymbol{x} = \boldsymbol{c} \tag{17.2}$$

を解くときに，この式そのものを解くのではなく，二次形式

$$F(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T \boldsymbol{S}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{x}^T\boldsymbol{c} + \boldsymbol{c}^T\boldsymbol{x}) \tag{17.3}$$

が極小値をとるときの $\boldsymbol{x}$ として解を捜す．第二，三項はスカラーであるから，これらはまとめて $\boldsymbol{x}^T\boldsymbol{c}$ と書くこともできるが，ここでは対称性のよい書き方をしておく．$F(\boldsymbol{x})$ の第一項は (17.1) 式により正であるから，$\|\boldsymbol{x}\|$ を大きくすれば $F(\boldsymbol{x})$ いくらでも大きくなる．よって $F(\boldsymbol{x})$ が極小値をもつことは明らかである．なお，ここではユークリッド (L_2) ノルムを用いる．

そこで $F(\boldsymbol{x})$ の変分を取ると

$$\delta F(\boldsymbol{x}) = F(\boldsymbol{x}+\delta\boldsymbol{x}) - F(\boldsymbol{x}) = \frac{1}{2}\delta\boldsymbol{x}^T\boldsymbol{S}\delta\boldsymbol{x} + \frac{1}{2}\Big(\delta\boldsymbol{x}^T(\boldsymbol{S}\boldsymbol{x}-\boldsymbol{c}) + \big[\delta\boldsymbol{x}^T(\boldsymbol{S}\boldsymbol{x}-\boldsymbol{c})\big]^T\Big)$$

となる．第三項は $\boldsymbol{S}$ が対称であることを用いている．よって $F(\boldsymbol{x})$ が極小値を取るのは

$$\boldsymbol{S}\boldsymbol{x} = \boldsymbol{c}$$

のときである．つまり，(17.2) 式の解を求めるにはスカラー値関数 $F(\boldsymbol{x})$ の極小点を捜せばよいことになる．

極小点を捜すためには，適当な初期値 $\boldsymbol{x}_0$ から出発して近似ベクトル $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots$ を作る．近似解 $\boldsymbol{x}_k$ からベクトル $\boldsymbol{p}_k$ の方向へつぎの近似解 $\boldsymbol{x}_{k+1}$ を捜すことにして

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k \tag{17.4}$$

とする．係数 α_k は $F(\boldsymbol{x}_{k+1})$ がこの方向で極小になるように決める．$F(\boldsymbol{x}_k + \alpha_k\boldsymbol{p}_k)$ を α_k で微分して 0 と置くと

$$\frac{\partial F(\boldsymbol{x}_k+\alpha_k\boldsymbol{p}_k)}{\partial \alpha_k} = \alpha_k \boldsymbol{p}_k^T\boldsymbol{S}\boldsymbol{p}_k + \frac{1}{2}\Big(\boldsymbol{p}_k^T(\boldsymbol{S}\boldsymbol{x}_k-\boldsymbol{c}) + \big[\boldsymbol{p}_k^T(\boldsymbol{S}\boldsymbol{x}_k-\boldsymbol{c})\big]^T\Big) = 0$$

となる．第二，三項に現れた

$$\boldsymbol{s}_k = \boldsymbol{c} - \boldsymbol{S}\boldsymbol{x}_k \tag{17.5}$$

は方程式 (17.2) の近似解 $\boldsymbol{x}_k$ における残差ベクトルである．これを用いると α_k は

$$\alpha_k = \frac{\boldsymbol{p}_k^T\boldsymbol{s}_k}{\boldsymbol{p}_k^T\boldsymbol{S}\boldsymbol{p}_k} \tag{17.6}$$

と求められる．α_k が求められると $\boldsymbol{s}_{k+1}$ は (17.5) 式を用いなくても

$$\boldsymbol{s}_{k+1} = \boldsymbol{s}_k - \alpha_k\boldsymbol{S}\boldsymbol{p}_k \tag{17.7}$$

からも計算できる．この式から，(17.6) 式の α_k は

$$\boldsymbol{p}_k^T\boldsymbol{s}_{k+1} = 0 \tag{17.8}$$

すなわち，つぎの点 $\boldsymbol{x}_{k+1}$ における残差ベクトルが探索方向に直交するように決まっていることがわかる．

問題は方向 $\boldsymbol{p}_k$ をどのように選べばよいかということである．常識的には $F(\boldsymbol{x}_{k+1})$ が $F(\boldsymbol{x}_k)$ よりも

小さくなる方向，つまり $F(\boldsymbol{x})$ の傾斜の下り方向に選ぶべきであろう．$\boldsymbol{x}_k$ における $F(\boldsymbol{x})$ の下り方向は

$$-\nabla F(\boldsymbol{x}_k) = \boldsymbol{c} - \boldsymbol{S}\boldsymbol{x}_k = \boldsymbol{s}_k \tag{17.9}$$

であるから，残差ベクトル $\boldsymbol{s}_k$ を $\boldsymbol{p}_k$ として用いることは可能である．これを用いた方法を最大勾配法という．この方法では常に $F(\boldsymbol{x})$ が最も減る方向に探索しているので効率が高いように見えるが，実はそうではなく，収束も保証されない．

そこで $\boldsymbol{p}_k$ を $\boldsymbol{s}_k$ だけでなく，その前のステップの $\boldsymbol{p}_{k-1}$ を用いて

$$\boldsymbol{p}_k = \boldsymbol{s}_k + \beta_{k-1}\boldsymbol{p}_{k-1} \tag{17.10}$$

と表すことにする．$\boldsymbol{p}_k$ と $\boldsymbol{p}_{k-1}$ は互いに独立になるように，これらが $\boldsymbol{S}$ に関して共役 (conjugate) になるという条件

$$\boldsymbol{p}_k^T\boldsymbol{S}\boldsymbol{p}_{k-1} = \boldsymbol{p}_{k-1}^T\boldsymbol{S}\boldsymbol{p}_k = 0 \tag{17.11}$$

をつける．(17.10) 式をこの式に代入すると

$$\boldsymbol{p}_{k-1}^T\boldsymbol{S}(\boldsymbol{s}_k + \beta_{k-1}\boldsymbol{p}_{k-1}) = 0$$

よって

$$\beta_k = -\frac{\boldsymbol{p}_k^T\boldsymbol{S}\boldsymbol{s}_{k+1}}{\boldsymbol{p}_k^T\boldsymbol{S}\boldsymbol{p}_k} \tag{17.12}$$

が得られた．

これで必要な式は全部そろった．アルゴリズムをまとめると以下のようになる．対称行列に対する CG 法という意味で，CGS という名前をつける．

アルゴリズム CGS

(a) $\boldsymbol{s}_0 = \boldsymbol{c} - \boldsymbol{S}\boldsymbol{x}_0 \quad \boldsymbol{p}_0 = \boldsymbol{s}_0$

(b) $\alpha_k = \dfrac{\|\boldsymbol{s}_k\|^2}{\boldsymbol{p}_k^T\boldsymbol{S}\boldsymbol{p}_k}$

(c) $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{p}_k$

(d) $\boldsymbol{s}_{k+1} = \boldsymbol{s}_k - \alpha_k\boldsymbol{S}\boldsymbol{p}_k$ (17.13)

$\|\boldsymbol{s}_{k+1}\| \ll \|\boldsymbol{c}\|$ なら終了

(e) $\beta_k = \dfrac{\|\boldsymbol{s}_{k+1}\|^2}{\|\boldsymbol{s}_k\|^2}$

(f) $\boldsymbol{p}_{k+1} = \boldsymbol{s}_{k+1} + \beta_k\boldsymbol{p}_k$ (b) へ

α_k と β_k の式は前に導いたものとは異なるが，これについては後で説明する．

(17.11) 式が成り立つように β_k を決めると

$$\boldsymbol{p}_j^T\boldsymbol{s}_k = 0 \qquad j < k \tag{17.14a}$$

$$\boldsymbol{s}_j^T\boldsymbol{s}_k = 0 \qquad j \neq k \tag{17.14b}$$

$$\boldsymbol{p}_j^T\boldsymbol{S}\boldsymbol{p}_k = 0 \qquad j \neq k \tag{17.14c}$$

が成り立つ．これらを証明するには数学的帰納法を用いる．はじめにある k にまでに対して上の関係が $j < k$ に対して成立しているとする．$k = 1$ に対してこれが成り立っていることは実際に $\boldsymbol{p}_1$, $\boldsymbol{s}_1$ を計算して確かめることができる．$\boldsymbol{s}_{k+1}$ を計算するために，(17.13)(c) 式を用いて $\boldsymbol{x}_{k+1}$ を $\boldsymbol{p}_i$ によって展開する．

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_{j+1} + \sum_{i=j+1}^{k}\alpha_i\boldsymbol{p}_i$$

これを用いると $\boldsymbol{s}_{k+1}$ は

$$\boldsymbol{s}_{k+1} = \boldsymbol{c} - \boldsymbol{S}\boldsymbol{x}_{k+1} = \boldsymbol{s}_{j+1} - \sum_{i=j+1}^{k}\alpha_i\boldsymbol{S}\boldsymbol{p}_i$$

と書ける．したがって

$$\boldsymbol{p}_j^T\boldsymbol{s}_{k+1} = \boldsymbol{p}_j^T\boldsymbol{s}_{j+1} - \sum_{i=j+1}^{k}\alpha_i\boldsymbol{p}_j^T\boldsymbol{S}\boldsymbol{p}_i$$

となるが，第一項は (17.8) 式により 0，第二項は $j < i \leq k$ であるから仮定により 0 であるから

$$\boldsymbol{p}_j^T\boldsymbol{s}_{k+1} = 0 \quad j < k$$

が成り立つ．$j = k$ のときは (17.8) 式そのものである．よって (17.14a) 式が証明された．

つぎにいま導いた関係と (17.13)(f) 式から

$$0 = \boldsymbol{p}_j^T\boldsymbol{s}_{k+1} = (\boldsymbol{s}_j + \beta_{j-1}\boldsymbol{p}_{j-1})^T\boldsymbol{s}_{k+1}$$

となるが，右辺第二項はすぐ上で導いた関係によって 0 である．よって

$$\boldsymbol{s}_j^T\boldsymbol{s}_{k+1} = 0 \quad j \leq k$$

が成り立つ．よって (17.4b) 式が証明された．

最後に (17.13)(f) 式を用いると

$$\boldsymbol{p}_j^T\boldsymbol{S}\boldsymbol{p}_{k+1} = \boldsymbol{p}_j^T\boldsymbol{S}(\boldsymbol{s}_{k+1} + \beta_k\boldsymbol{p}_k)$$

第二項は仮定により 0，また $\boldsymbol{p}_j$ を (17.13)(c) 式を用いて書きかえると

$$\boldsymbol{p}_j^T\boldsymbol{S}\boldsymbol{p}_{k+1} = \frac{1}{\alpha_j}(\boldsymbol{x}_{j+1} - \boldsymbol{x}_j)^T\boldsymbol{S}\boldsymbol{s}_{k+1}$$

となる．ところで $\boldsymbol{s}_j = \boldsymbol{c} - \boldsymbol{S}\boldsymbol{x}_j$ であるから

$$\big(\boldsymbol{x}_{j+1} - \boldsymbol{x}_j\big)^T \boldsymbol{S} = -\big(\boldsymbol{s}_{j+1} - \boldsymbol{s}_j\big)^T$$

である．よって

$$\boldsymbol{p}_j^T \boldsymbol{S}\boldsymbol{p}_{k+1} = -\frac{1}{\alpha_j}\big(\boldsymbol{s}_{j+1} - \boldsymbol{s}_j\big)^T \boldsymbol{s}_{k+1} = 0$$

となる．最後の関係は先に導いた $\boldsymbol{s}_j$ の直交性を用いている．よって

$$\boldsymbol{p}_j^T \boldsymbol{S}\boldsymbol{p}_{k+1} = 0 \quad j \le k$$

が導かれた．これが (17.14c) 式である．上では $\alpha_j \neq 0$ と仮定したが，α_j が 0 のときには (17.6) 式より $\boldsymbol{p}_j^T \boldsymbol{s}_j = 0$ であるから，(17.13)(f) 式を用いれば

$$0 = \boldsymbol{p}_j^T \boldsymbol{s}_j = \big(\boldsymbol{s}_j + \beta_{j-1}\boldsymbol{p}_{j-1}\big)^T \boldsymbol{s}_j = \|\boldsymbol{s}_j\|^2$$

が成り立つ．よって $\alpha_j = 0$ のときには $\boldsymbol{s}_j = 0$ になる．$\boldsymbol{s}_j = 0$ であることは $\boldsymbol{x}_j$ が (17.2) 式の解になっていることを意味するから，反復をこれ以上続ける必要はない．

(17.14c) 式を満足する $\boldsymbol{p}_j$ は互いに一次独立である．もし一次独立でないなら

$$\sum_j c_j \boldsymbol{p}_j = 0$$

となるような定数 c_j が存在するはずである．上式に $\boldsymbol{p}_k^T \boldsymbol{S}$ を左から掛けると $j = k$ 以外の項はすべて 0 になって

$$0 = \sum_j c_j \boldsymbol{p}_k^T \boldsymbol{S}\boldsymbol{p}_j = c_k \boldsymbol{p}_k^T \boldsymbol{S}\boldsymbol{p}_k$$

よってすべての c_k が 0 になるから $\boldsymbol{p}_k$ は互いに独立である．

n ステップ目の近似解は

$$\boldsymbol{x}_n = \boldsymbol{x}_0 + \alpha_0\boldsymbol{p}_0 + \alpha_1\boldsymbol{p}_1 + \cdots + \alpha_{n-1}\boldsymbol{p}_{n-1}$$

と表される．n 次元空間には一次独立なベクトルは n 個しかないから，上式の α_k がうまく選ばれていればこれは (17.2) 式の解になっているはずである．実際にそうなっていることは，残差ベクトル $\boldsymbol{s}_k$ が互いに直交していることから保証される．n 次元空間では互いに直交する方向は n 個しかないから，丸め誤差がないならば $\boldsymbol{s}_n = 0$ になるはずである．これは $\boldsymbol{x}_n$ が (17.2) 式の解であることを意味している．このときには (17.13)(e)，(f) 式から $\beta_{n-1} = 0$, $\boldsymbol{p}_n = 0$ になる．

最後に (17.13) 式に現れた α_k と β_k を証明しておく．(17.10)，(17.8) 式から

$$\boldsymbol{p}_k^T \boldsymbol{s}_k = (\boldsymbol{s}_k + \beta_{k-1}\boldsymbol{p}_{k-1})^T \boldsymbol{s}_k = \boldsymbol{s}_k^T \boldsymbol{s}_k \qquad (17.15)$$

が成り立つから，(17.6) 式は (17.13)(b) 式の α_k に等しくなる．つぎに (17.4) 式を用いると

$$\boldsymbol{S}\boldsymbol{p}_k = \frac{1}{\alpha_k}\boldsymbol{S}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) = \frac{1}{\alpha_k}(\boldsymbol{s}_k - \boldsymbol{s}_{k+1})$$

となるから

$$\begin{aligned}\boldsymbol{s}_{k+1}^T \boldsymbol{S}\boldsymbol{p}_k &= \frac{1}{\alpha_k}\boldsymbol{s}_{k+1}^T(\boldsymbol{s}_k - \boldsymbol{s}_{k+1}) \\ &= -\frac{1}{\alpha_k}\|\boldsymbol{s}_{k+1}\|^2 = -\boldsymbol{p}_k^T \boldsymbol{S}\boldsymbol{p}_k \frac{\|\boldsymbol{s}_{k+1}\|^2}{\|\boldsymbol{s}_k\|^2}\end{aligned}$$

ここの関係を (17.12) 式に代入すれば (17.13)(e) 式の β_k が得られる．

収束の判定には $\|\boldsymbol{s}_{k+1}\|$ を用いるのがこの場合唯一の選択肢である．これを用いるとつぎのステップで β_{k+1} の計算で分母が 0 になることを避けることができる．もう一つの割り算は α_k に現れる．$\boldsymbol{S}$ が正定値ならこの分母は $\boldsymbol{p}_k = 0$ のとき以外は 0 にならない．しかし $\boldsymbol{p}_k = 0$ のときには (17.15) 式から $\|\boldsymbol{s}_k\| = 0$ となる．したがって収束判定には $\boldsymbol{s}_{k+1}$ の監視だけで十分である．

非対称行列 (CGNE)　つぎに，必ずしも対称とは限らない正則行列 $\boldsymbol{A}$ を係数行列とする連立方程式

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \qquad (17.16)$$

を CG 法で解く．$\boldsymbol{A}$ が正則なら

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b} \qquad (17.17)$$

の解は (17.16) 式の解に一致する．したがって上式 (17.17) 式を解けば十分である．ところでこの式は (17.2) 式において

$$\boldsymbol{S} = \boldsymbol{A}^T\boldsymbol{A} \quad \boldsymbol{c} = \boldsymbol{A}^T\boldsymbol{b} \qquad (17.18)$$

としたものであるから，前項の方法をそのまま適用することができる．このときの二次形式 (17.3) 式は

$$F(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 - \frac{1}{2}\|\boldsymbol{b}\|^2 \qquad (17.19)$$

となるから，非対称行列 $\boldsymbol{A}$ に対する共役勾配法では

$$\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 = \min \qquad (17.20)$$

となる解を捜すことになる．

そこで (17.13) 式に相当するアルゴリズムを示す．注意しなければならないのは (17.18) 式の関係をそのまま (17.13) 式に用いてはならないということである．$\boldsymbol{A}^T\boldsymbol{A}$ の条件数は $\boldsymbol{A}$ の条件数の二乗であるから，こうすると収束が非常に遅くなるからである．そこでたとえば

$$\boldsymbol{p}_k^T\boldsymbol{S}\boldsymbol{p}_k = \boldsymbol{p}_k^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{p}_k = \|\boldsymbol{A}\boldsymbol{p}_k\|^2$$

などの関係を用いて $\boldsymbol{A}^T\boldsymbol{A}$ を生で使わないように書きかえなければならない．

アルゴリズム CGNE

(a) $\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0 \quad \boldsymbol{p}_0 = \boldsymbol{s}_0 = \boldsymbol{A}^T\boldsymbol{r}_0$

(b) $\boldsymbol{q}_k = \boldsymbol{A}\boldsymbol{p}_k$

(c) $\alpha_k = \dfrac{\|\boldsymbol{s}_k\|^2}{\|\boldsymbol{q}_k\|^2}$

(d) $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{p}_k$ (17.21)

(e) $\boldsymbol{r}_{k+1} = \boldsymbol{r}_k - \alpha_k\boldsymbol{q}_k$

(f) $\boldsymbol{s}_{k+1} = \boldsymbol{A}^T\boldsymbol{r}_{k+1}$ 収束の判定

(g) $\beta_k = \dfrac{\|\boldsymbol{s}_{k+1}\|^2}{\|\boldsymbol{s}_k\|^2}$

(h) $\boldsymbol{p}_{k+1} = \boldsymbol{s}_{k+1} + \beta_k\boldsymbol{p}_k$ (b) へ

ここで $\boldsymbol{r}_k$ は方程式 (17.16) の残差

$$\boldsymbol{r}_k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k \tag{17.22}$$

を表している．興味深いことに，α_k は残差 $\boldsymbol{r}_{k+1}$ のノルムが減少するように選ばれている．(17.21)(e) 式から

$$\|\boldsymbol{r}_{k+1}\|^2 = \|\boldsymbol{r}_k\|^2 - 2\alpha_k\boldsymbol{q}_k^T\boldsymbol{r}_k + \alpha_k^2\|\boldsymbol{q}_k\|^2$$

となるから，$\|\boldsymbol{r}_{k+1}\|^2$ が極小になるように α_k を決めることにすれば，この式を α_k で微分して 0 と置くと

$$\alpha_k = \frac{\boldsymbol{q}_k^T\boldsymbol{r}_k}{\|\boldsymbol{q}_k\|^2}$$

が得られる．ところが

$$\boldsymbol{q}_k^T\boldsymbol{r}_k = (\boldsymbol{A}\boldsymbol{p}_k)^T\boldsymbol{r}_k = \boldsymbol{p}_k^T\boldsymbol{A}^T\boldsymbol{r}_k = \boldsymbol{p}_k^T\boldsymbol{s}_k$$

であるから，このように決めた α_k は (17.6) 式の α_k と等しくなる．したがってアルゴリズム CGNE の残差ベクトル $\boldsymbol{r}_{k+1}$ は

$$\|\boldsymbol{r}_{k+1}\|^2 = \|\boldsymbol{r}_k\|^2 - \alpha_k^2\|\boldsymbol{q}_k\|^2 \tag{17.23}$$

を満たしており，残差のノルムは 1 ステップごとに必ず減少する．

$\boldsymbol{p}_k$ と $\boldsymbol{s}_k$ は (17.14a)，(17.14b) 式を満たしている．いまの場合 (17.14a) 式は

$$\boldsymbol{p}_j^T\boldsymbol{s}_k = \boldsymbol{q}_j^T\boldsymbol{r}_k = 0 \quad j < k$$

と書くこともできる．また (17.14c) 式は

$$\boldsymbol{p}_j^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{p}_k = \boldsymbol{q}_j^T\boldsymbol{q}_k = 0 \quad j \neq k$$

であるから，$\boldsymbol{s}_k$ だけでなく $\boldsymbol{q}_k$ も互いに直交している．以上をまとめると CGNE では

$$\boldsymbol{q}_j^T\boldsymbol{r}_k = 0 \quad j < k \tag{17.24a}$$
$$\boldsymbol{s}_j^T\boldsymbol{s}_k = 0 \quad j \neq k \tag{17.24b}$$
$$\boldsymbol{q}_j^T\boldsymbol{q}_k = 0 \quad j \neq k \tag{17.24c}$$

が成り立っている．

アルゴリズム CGNE の収束判定は対称行列のときと同様に $\|\boldsymbol{s}_{k+1}\|$ を用いればよい．α_k の分母に $\|\boldsymbol{q}_k\|^2$ が現れているが

$$0 = \boldsymbol{q}_k^T\boldsymbol{r}_k = \boldsymbol{p}_k^T\boldsymbol{s}_k = (\boldsymbol{s}_k + \beta_{k-1}\boldsymbol{p}_{k-1})^T\boldsymbol{s}_k$$

の第二項は (17.14a) 式により 0 になる．よって $\boldsymbol{q}_k$ が 0 になるときには $\boldsymbol{s}_k$ も 0 になる．したがって $\|\boldsymbol{s}_{k+1}\| = 0$ でループを脱出すれば 0 で割り算をすることはない．

最小二乗解　アルゴリズム CGNE では $\boldsymbol{A}$ は正方行列である必要はない．$\boldsymbol{A}$ を $n \times m$ 行列とすると，方程式 (17.16) 式は一般には解をもたない．しかし共役勾配法では (17.20) 式を満たす解 $\boldsymbol{x}$ を捜すので，求められる解はある種の最小二乗解である．ここではもっともよくある場合，$n > m$ の場合を主に考える．

$n \neq m$ のときに注意しなければならないのは，ここに現れるベクトルの次元がつぎのようになっていることである．

$$\boldsymbol{x}_k,\ \boldsymbol{p}_k,\ \boldsymbol{s}_k \ :\ m \text{ 次元}$$
$$\boldsymbol{r}_k,\ \boldsymbol{q}_k \ :\ n \text{ 次元}$$

$\boldsymbol{s}_k$, $\boldsymbol{q}_k$ はそれぞれが直交系をなしているが, $n > m$ のときには $\boldsymbol{s}_m$ の方が先に 0 になる. このときには $\boldsymbol{s}_k$ の定義から

$$\boldsymbol{s}_m = \boldsymbol{A}^T\boldsymbol{r}_m = \boldsymbol{A}^T(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_m) = 0$$

が成り立っている. したがって $\boldsymbol{x}_m$ は正規方程式の解にほかならない. もし $\boldsymbol{A}$ がフルランクなら, この最小二乗解は一義的である. いいかえればどのような初期ベクトル $\boldsymbol{x}_0$ から出発しても一義的な解が得られる (丸め誤差がないとして). $\boldsymbol{A}$ が正方行列でなくてもアルゴリズム (17.21) 式は最小二乗解を与えるという意味で, normal equation をとって CGNE と名付けている.

$\boldsymbol{A}$ がフルランクでないときには話はややこしくなる. $\boldsymbol{A}$ にランク落ちがあるときにもアルゴリズム CGNE の解は正規方程式の解に収束する. しかしランク落ちがあるときには解は一義的ではなく, 初期値 $\boldsymbol{x}_0$ に依存する. どのような解に収束するかが問題である.

はじめに, $\boldsymbol{p}_k$ が $\boldsymbol{A}^T$ の値域に属する, すなわち

$$\boldsymbol{p}_k \in R(\boldsymbol{A}^T)$$

を満たしていることを証明する. $\boldsymbol{p}_0 = \boldsymbol{A}^T\boldsymbol{r}_0$ であるから, $\boldsymbol{p}_0$ は上式を満たしている. $\boldsymbol{p}_1$ は, (17.21) 式を用いると

$$\begin{aligned}
\boldsymbol{q}_0 &= \boldsymbol{A}\boldsymbol{p}_0 \\
\boldsymbol{r}_1 &= \boldsymbol{r}_0 - \alpha_0\boldsymbol{A}\boldsymbol{p}_0 \\
\boldsymbol{s}_1 &= \boldsymbol{p}_0 - \alpha_0\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{p}_0 \\
\boldsymbol{p}_1 &= (1+\beta_0)\boldsymbol{p}_0 - \alpha_0\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{p}_0
\end{aligned}$$

と計算されるから $\boldsymbol{p}_1 \in R(\boldsymbol{A}^T)$ である. 一般に $\boldsymbol{p}_k$ は (17.21) 式から漸化式

$$\boldsymbol{p}_{k+1} = (1+\beta_k)\boldsymbol{p}_k - \alpha_k\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{p}_k - \beta_{k-1}\boldsymbol{p}_{k-1}$$

を満たすから, $\boldsymbol{p}_0$, $\boldsymbol{p}_1$ を初期値として計算すれば $\boldsymbol{p}_k$ はつぎのような空間内にあることがわかる.

$$\begin{aligned}
&\boldsymbol{p}_k = \{\boldsymbol{p}_0, \boldsymbol{B}\boldsymbol{p}_0, \boldsymbol{B}^2\boldsymbol{p}_0, \cdots, \boldsymbol{B}^k\boldsymbol{p}_0\} \in R(\boldsymbol{A}^T) \\
&\boldsymbol{B} = \boldsymbol{A}^T\boldsymbol{A}
\end{aligned}$$

前にも用いた $\boldsymbol{x}_k$ の展開

$$\boldsymbol{x}_k = \boldsymbol{x}_0 + \sum_{j=0}^{k-1}\alpha_j\boldsymbol{p}_j$$

を用いれば, CGNE の場合

$$\boldsymbol{x}_0 \in R(\boldsymbol{A}^T) \quad \text{なら} \quad \boldsymbol{x}_k \in R(\boldsymbol{A}^T)$$

が証明された. ここで付録の補題 3 を用いるとアルゴリズム CGNE の解は

$$\boldsymbol{x}_0 \in R(\boldsymbol{A}^T) \quad \text{なら} \quad \boldsymbol{x} = \boldsymbol{A}^\dagger\boldsymbol{b} \tag{17.25}$$

に収束することが証明された. ここに $\boldsymbol{A}^\dagger$ は $\boldsymbol{A}$ の一般逆行列である. この解は正規方程式 $\boldsymbol{A}^T(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}) = 0$ を満足する解の中のノルム最小の解である.

ここで条件 $\boldsymbol{x}_0 \in R(\boldsymbol{A}^T)$ は重要である. この条件がなければ $\boldsymbol{x}_k$ がどこに収束するかわからない. この条件を満足させるためには任意の n 次元ベクトル $\boldsymbol{y}_0$ を用いて

$$\boldsymbol{x}_0 = \boldsymbol{A}^T\boldsymbol{y}_0$$

とすればよい. この式は $\boldsymbol{x}_0$ が $\boldsymbol{A}$ の行ベクトルの線型結合であることを示している. 最も簡単には $\boldsymbol{x}_0$ として 0 ベクトルを選べばよい. このときには $\boldsymbol{x}_k$ は必ず $R(\boldsymbol{A}^T)$ に属している.

方程式 (17.16) の右辺 $\boldsymbol{b}$ や初期値 $\boldsymbol{x}_0$ は一般的につぎのように書くことができる.

$$\begin{aligned}
&\boldsymbol{b} = \boldsymbol{b}' + \boldsymbol{b}'' \quad \boldsymbol{b}' \in N(\boldsymbol{A}^T) \quad \boldsymbol{b}'' \in R(\boldsymbol{A}) \\
&\boldsymbol{x}_0 = \boldsymbol{x}_0' + \boldsymbol{x}_0'' \\
&\quad \boldsymbol{x}_0' \in N(\boldsymbol{A}) \quad \boldsymbol{x}_0'' \in R(\boldsymbol{A}^T)
\end{aligned} \tag{17.26}$$

このとき, 定義により

$$\boldsymbol{A}^T\boldsymbol{b}' = 0 \quad \boldsymbol{A}\boldsymbol{x}_0' = 0$$

である. $\boldsymbol{x}_{k+1}$ を展開すると

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_0' + \Big[\boldsymbol{x}_0'' + \sum_{j=0}^{k}\alpha_j\boldsymbol{p}_j\Big]$$

と書ける. 第一項は $N(\boldsymbol{A})$ に属し, 第二項以下は $R(\boldsymbol{A}^T)$ に属している. この二つの空間は互いに直交しているので, 初期値に含まれている $N(\boldsymbol{A})$ 成分は反復によって変化しないことがわかる. また, 初期値によって解が異なることもこれで明らかである.

$\boldsymbol{x}_{k+1}$ に対する残差ベクトルは

$$\boldsymbol{r}_{k+1} = \boldsymbol{b}' + \Big[(\boldsymbol{b}'' - \boldsymbol{A}\boldsymbol{x}_0'') - \sum_{j=0}^{k}\alpha_j\boldsymbol{q}_j\Big]$$

と書くことができる. 第一項は $N(\boldsymbol{A}^T)$ に属し, $\boldsymbol{q}_j \in R(\boldsymbol{A})$ であるから, 第二項以下は $R(\boldsymbol{A})$ に属している. したがって残差ベクトルのノルムは $\|\boldsymbol{b}'\|$ よりも小さくはなり得ないことがわかる.

非対称行列 (CGMN) (17.16) 式のかわりに今度は

$$\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{z} = \boldsymbol{b} \quad \boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{z} \tag{17.27}$$

から $\boldsymbol{x}$ を求める．これは未知数よりも条件の数が少ないときの最小 2 乗法に現れる方程式である (§10, (10.52) 式参照)．CGS 法 (17.13) 式において $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^T$ とするとつぎのようなアルゴリズムが導かれる．ただし z_k は x_k で書き直してある．

アルゴリズム CGMN

$$\begin{aligned}
&\text{(a)} \quad \boldsymbol{p}_0 = \boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0 \\
&\text{(b)} \quad \boldsymbol{q}_k = \boldsymbol{A}^T\boldsymbol{p}_k \\
&\text{(c)} \quad \alpha_k = \frac{\|\boldsymbol{r}_k\|^2}{\|\boldsymbol{q}_k\|^2} \\
&\text{(d)} \quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{q}_k \\
&\text{(e)} \quad \boldsymbol{r}_{k+1} = \boldsymbol{r}_k - \alpha_k\boldsymbol{A}\boldsymbol{q}_k \\
&\text{(f)} \quad \beta_k = \frac{\|\boldsymbol{r}_{k+1}\|^2}{\|\boldsymbol{r}_k\|^2} \\
&\text{(g)} \quad \boldsymbol{p}_{k+1} = \boldsymbol{r}_{k+1} + \beta_k\boldsymbol{p}_k \quad \text{(b) へ}
\end{aligned} \tag{17.28}$$

$\boldsymbol{r}_k$ は (17.22) 式で定義された残差であるが，$\boldsymbol{p}_k$ や $\boldsymbol{q}_k$ の意味は前とは同じではない．

(17.16) 式の右辺 $\boldsymbol{b}$ が適合条件を満たしているとすると，すなわち $\boldsymbol{b} \in R(\boldsymbol{A})$ のとき，この方程式は解をもつ (§11)．この解は一義的ではないかもしれないが，一つの解を $\boldsymbol{h}$ とする．すなわち

$$\boldsymbol{A}\boldsymbol{h} = \boldsymbol{b} \tag{17.29}$$

が成り立っているとする．この解を用いて近似解 $\boldsymbol{x}_k$ の誤差を

$$\boldsymbol{e}_k = \boldsymbol{h} - \boldsymbol{x}_k \tag{17.30}$$

と置く．このとき (17.28)(d) 式から

$$\boldsymbol{e}_{k+1} = \boldsymbol{e}_k - \alpha_k\boldsymbol{q}_k \tag{17.31}$$

が成り立つ．$\boldsymbol{e}_{k+1}$ のノルムは

$$\|\boldsymbol{e}_{k+1}\|^2 = \|\boldsymbol{e}_k\|^2 - 2\alpha_k\boldsymbol{q}_k^T\boldsymbol{e}_k + \alpha_k^2\|\boldsymbol{q}_k\|^2$$

となるが，$\boldsymbol{q}_k$ の定義などから

$$\boldsymbol{q}_k^T\boldsymbol{e}_k = \boldsymbol{p}_k^T\boldsymbol{A}\boldsymbol{e}_k = \boldsymbol{p}_k^T(\boldsymbol{A}\boldsymbol{h} - \boldsymbol{A}\boldsymbol{x}_k) = \boldsymbol{p}_k^T\boldsymbol{r}_k$$

が成り立つから

$$\|\boldsymbol{e}_{k+1}\|^2 = \|\boldsymbol{e}_k\|^2 - 2\alpha_k\boldsymbol{p}_k^T\boldsymbol{r}_k + \alpha_k^2\|\boldsymbol{q}_k\|^2$$

である．したがって誤差のノルムを極小にするためには

$$\alpha_k = \frac{\boldsymbol{p}_k^T\boldsymbol{r}_k}{\|\boldsymbol{q}_k\|^2} \tag{17.32}$$

と選べばよい．α_k をこのように選ぶと

$$\|\boldsymbol{e}_{k+1}\|^2 = \|\boldsymbol{e}_k\|^2 - \alpha_k^2\|\boldsymbol{q}_k\|^2 \tag{17.33}$$

が成り立つ．すなわち，このアルゴリズムでは近似解の誤差ベクトルのノルムは 1 ステップごとに減少する．ただし，これは $\boldsymbol{b}$ が $\boldsymbol{A}$ の値域にあるときに成り立つもので，任意の $\boldsymbol{b}$ に対しては成り立たない．

係数 β_k は直交関係

$$\boldsymbol{q}_k^T\boldsymbol{q}_{k+1} = 0 \tag{17.34}$$

が成り立つように決められている．これは $\boldsymbol{x}_k$ に対する補正量が互いに直交するように選んでいることに相当している．(17.28) 式から

$$\boldsymbol{q}_{k+1} = \boldsymbol{A}^T\boldsymbol{r}_{k+1} + \beta_k\boldsymbol{q}_k$$

であるから，直交条件は

$$0 = \boldsymbol{q}_{k+1}^T\boldsymbol{q}_k = \boldsymbol{r}_{k+1}^T\boldsymbol{A}\boldsymbol{q}_k + \beta_k\|\boldsymbol{q}_k\|^2$$

となる．したがって

$$\beta_k = -\frac{\boldsymbol{r}_{k+1}^T\boldsymbol{A}\boldsymbol{q}_k}{\|\boldsymbol{q}_k\|^2} \tag{17.35}$$

が得られる．

$\boldsymbol{p}_j$, $\boldsymbol{q}_j$, $\boldsymbol{r}_j$ などが CGNE と似たような直交関係

$$\boldsymbol{p}_j^T\boldsymbol{r}_k = 0 \quad j < k \tag{17.36a}$$
$$\boldsymbol{r}_j^T\boldsymbol{r}_k = 0 \quad j \neq k \tag{17.36b}$$
$$\boldsymbol{q}_j^T\boldsymbol{q}_k = 0 \quad j \neq k \tag{17.36c}$$

を満たしていることは数学的帰納法によって証明することができる．この関係を用いると，(17.32) 式の α_k，(17.35) 式の β_k を (17.28) 式の形に書きかえることができる．まず (17.28)(g) 式を用いて (17.28)(a) 式に注意すれば

$$\boldsymbol{p}_k^T\boldsymbol{r}_k = \|\boldsymbol{r}_k\|^2 + \beta_{k-1}\boldsymbol{p}_{k-1}^T\boldsymbol{r}_k = \|\boldsymbol{r}_k\|^2$$

であるから，(17.32) 式は (13.28)(c) 式に一致する．つぎに (17.28) (e) 式と (17.28)(b) 式から

$$\|\boldsymbol{r}_{k+1}\|^2 = \boldsymbol{r}_{k+1}^T(\boldsymbol{r}_k - \alpha_k\boldsymbol{A}\boldsymbol{q}_k) = -\alpha_k\boldsymbol{r}_{k+1}^T\boldsymbol{A}\boldsymbol{q}_k$$

となる．この式に α_k を代入して (17.35) 式から β_k を求めれば (17.28)(f) 式が得られる．

アルゴリズム CGMN が停止するのは $\boldsymbol{r}_k = 0$ または $\boldsymbol{q}_k = 0$ のときである．前者 $\boldsymbol{r}_k = 0$ は方程式 (17.16) が厳密に成立するときである．$\boldsymbol{r}_j$ は互いに直交する n 次元ベクトルであるから，少なくとも n ステップ以内に 0 になる．しかしアルゴリズム CGNE と異なり $\|\boldsymbol{r}_j\|$ は単調減少ではないので，$\boldsymbol{r}_k$ が 0 になるのを数値的に判定するのは簡単ではない．$\|\boldsymbol{r}_j\|$ が増加しながら最後のステップで 0 になるかもしれないからである．

後者 $\boldsymbol{q}_k = 0$ は CGNE とは異なり，一般には必ずしも $\boldsymbol{r}_k = 0$ を意味しない．しかし $\boldsymbol{b} \in R(\boldsymbol{A})$ のときには $\boldsymbol{q}_k = 0$ が $\boldsymbol{r}_k = 0$ を意味することは，つぎのようにして示すことができる．

まず，(17.31) 式より

$$\boldsymbol{q}_k^T \boldsymbol{e}_{k+1} = \boldsymbol{q}_k^T \boldsymbol{e}_k - \alpha_k \|\boldsymbol{q}_k\|^2$$

であるが，(17.28)(b) 式から

$$\boldsymbol{q}_k^T \boldsymbol{e}_k = \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{e}_k = \boldsymbol{p}_k^T \boldsymbol{r}_k$$

となる．ここで (17.30) 式から $\boldsymbol{A}\boldsymbol{e}_k = \boldsymbol{r}_k$ が成り立っていることを用いている．したがって

$$\boldsymbol{q}_k^T \boldsymbol{e}_{k+1} = \boldsymbol{p}_k^T \boldsymbol{r}_k - \alpha_k \|\boldsymbol{q}_k\|^2 = 0$$

である．ここで (17.32) 式を用いている．したがって

$$\begin{aligned} \boldsymbol{e}_k^T \boldsymbol{q}_k &= \boldsymbol{e}_k^T (\boldsymbol{A}^T \boldsymbol{r}_k + \beta_{k-1} \boldsymbol{q}_{k-1}) \\ &= \|\boldsymbol{r}_k\|^2 + \beta_{k-1} \boldsymbol{e}_k^T \boldsymbol{q}_{k-1} = \|\boldsymbol{r}_k\|^2 \end{aligned}$$

が成り立つ．よって $\boldsymbol{q}_k = 0$ は $\boldsymbol{r}_k = 0$ を意味する．

(17.28)(b) より $\boldsymbol{q}_k \in R(\boldsymbol{A}^T)$ であるから，$\boldsymbol{x}_0 \in R(\boldsymbol{A}^T)$ なら $\boldsymbol{x}_k \in R(\boldsymbol{A}^T)$ である．したがって

$$\boldsymbol{b} \in R(\boldsymbol{A}) \quad \boldsymbol{x}_0 \in R(\boldsymbol{A}^T)$$

のとき，補題 4 によってアルゴリズム CGMN の反復は解

$$\boldsymbol{x} = \boldsymbol{A}^\dagger \boldsymbol{b} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

に収束する．この解は直接観測の最小二乗法におけるノルム最小の解に対応しているので (§10 参照)，minimum norm をとって CGMN と名づけている．ただしノルム最小になるのは変数 $\boldsymbol{z}$ であって，$\boldsymbol{b}$ が $R(\boldsymbol{A})$ に属さないときには $\boldsymbol{x}$ はノルム最小にはならないことに注意しなければならない．

$\boldsymbol{b} \in R(\boldsymbol{A})$ のときには (17.32) 式の解は

$$\boldsymbol{h} = \boldsymbol{A}^\dagger \boldsymbol{b} + \boldsymbol{x}_0' \quad \boldsymbol{A}\boldsymbol{x}_0' = 0$$

と書くことができる．近似解の誤差 $\boldsymbol{e}_k$ のノルムは，$\boldsymbol{x}_k \in R(\boldsymbol{A}^T)$ を考慮すれば

$$\begin{aligned} \|\boldsymbol{e}_k\|^2 &= \|\boldsymbol{A}^\dagger \boldsymbol{b} + \boldsymbol{x}_0' - \boldsymbol{x}_k\|^2 \\ &= \|\boldsymbol{A}^\dagger \boldsymbol{b} - \boldsymbol{x}_k\|^2 + \|\boldsymbol{x}_0'\|^2 \end{aligned}$$

である．(17.33) 式によってこれは 1 ステップごとに減少する．これは $\|\boldsymbol{A}^\dagger \boldsymbol{b} - \boldsymbol{x}_k\|^2$ が 1 ステップごとに減少することを意味している．

前処理　共役勾配法の収束は係数行列の条件数が大きいほど遅くなる．そこでもとの方程式 (17.2) 式や (17.16) 式を線型変換して，係数行列の条件数を小さくしてやることが考えられる．

たとえば正値対称行列 $\boldsymbol{S}$ を係数とする方程式 (17.2) 式を，正則な行列 $\boldsymbol{C}$ を用いて

$$\begin{aligned} &\boldsymbol{S}\boldsymbol{C}^{-1}\overline{\boldsymbol{x}} = \boldsymbol{b} \quad \overline{\boldsymbol{x}} = \boldsymbol{C}\boldsymbol{x} \\ &\boldsymbol{C}^{-1}\boldsymbol{S}\boldsymbol{x} = \boldsymbol{C}^{-1}\boldsymbol{b} \end{aligned}$$

などと変換したとする．行列 $\boldsymbol{C}$ をうまく選んで，条件数 $\kappa(\boldsymbol{S})$ よりも $\kappa(\boldsymbol{S}\boldsymbol{C}^{-1})$ や $\kappa(\boldsymbol{C}^{-1}\boldsymbol{S})$ の方を小さくすることができれば，もとの方程式を解くよりも変換された方程式を解いた方が収束が速くなる．極端な場合，$\boldsymbol{C} = \boldsymbol{S}$ とすれば (17.2) 式は解けてしまうが，これでは意味がない．数値計算上望ましいのは，$\boldsymbol{C}$ が $\boldsymbol{S}$ と同様な疎な構造をしており，また $\boldsymbol{C}$ を係数行列とする連立一次方程式が簡単に解けることである．簡単な例としては，$\boldsymbol{C}$ として $\boldsymbol{S}$ の対角成分を選ぶことである．これによっても条件数が減少することが証明されている．

ここでは (17.17) 式と同等な式

$$\begin{aligned} &\boldsymbol{S}\boldsymbol{z} = \boldsymbol{c} \quad \boldsymbol{z} = \boldsymbol{C}\boldsymbol{x} \\ &\boldsymbol{S} = \boldsymbol{C}^{-T}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{C}^{-1} \quad \boldsymbol{c} = \boldsymbol{C}^{-T}\boldsymbol{A}^T\boldsymbol{b} \end{aligned} \tag{17.37}$$

を解くことにする．ここに $\boldsymbol{C}^{-T}$ は $(\boldsymbol{C}^{-1})^T$ を意味している．$\boldsymbol{C}$ は $\boldsymbol{A}^T\boldsymbol{A}$ にできるだけ近くなるように，すなわち

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{C}^{-1} \simeq \boldsymbol{I}$$

であると同時に，$\boldsymbol{C}$ を係数行列とする連立方程式

$$\boldsymbol{Ct} = \boldsymbol{p} \tag{17.38}$$

が簡単に解けるものでなければならない．$\boldsymbol{A}$ が $n\times m$ 行列のとき $\boldsymbol{C}$ は $m \times m$ 行列である．

(17.37) 式で定義された係数行列 $\boldsymbol{S}$ は対称であるから，アルゴリズム CGS(17.13) 式を適用して z を求めることができる．これは $\boldsymbol{x}$ を用いて以下のように書きかえることができる．

アルゴリズム GCPCNE

$$\begin{aligned}
&\text{(a)} \quad \boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0 \quad \boldsymbol{p}_0 = \boldsymbol{s}_0 = \boldsymbol{C}^{-T}\boldsymbol{A}^T\boldsymbol{r}_0 \\
&\text{(b)} \quad \boldsymbol{q}_k = \boldsymbol{A}\boldsymbol{C}^{-1}\boldsymbol{p}_k \\
&\text{(c)} \quad \alpha_k = \frac{\|\boldsymbol{s}_k\|^2}{\|\boldsymbol{q}_k\|^2} \\
&\text{(d)} \quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{C}^{-1}\boldsymbol{p}_k \\
&\text{(e)} \quad \boldsymbol{r}_{k+1} = \boldsymbol{r}_k - \alpha_k\boldsymbol{q}_k \\
&\text{(f)} \quad \boldsymbol{s}_{k+1} = \boldsymbol{C}^{-T}\boldsymbol{A}^T\boldsymbol{r}_{k+1} \\
&\text{(g)} \quad \beta_k = \frac{\|\boldsymbol{s}_{k+1}\|^2}{\|\boldsymbol{s}_k\|^2} \\
&\text{(h)} \quad \boldsymbol{p}_{k+1} = \boldsymbol{s}_k + \beta_k\boldsymbol{p}_k
\end{aligned} \tag{17.39}$$

PC は前処理付き (preconditioned) の意味である．前処理とはいっても，実際は前処理をするのではなく，この処理は反復の中に組み込まれている．このアルゴリズムには逆行列 $\boldsymbol{C}^{-1}$ が現れているが，逆行列そのものは必要なく，たとえば $\boldsymbol{C}^{-1}\boldsymbol{p}$ は連立方程式 (17.38) を解いて計算する．

このアルゴリズムでは残差のノルム最小の解が求められるが，解のノルムが最小であるとは限らない．また，CGMN に対する前処理も考えられるがここでは省略する．

参考文献

Björck, Å. and T. Elfving (1979) : Accelerated projection methods for computing pseudo-inverse solutions of systems of linear equations, BIT, **19**, 145-163.

付　録

定義　$R(\boldsymbol{A})$ とは，$\boldsymbol{x}$ が存在して $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ と表されるような $\boldsymbol{y}$ の集合である．$R(\boldsymbol{A})$ を $\boldsymbol{A}$ の値域 (range) という．$N(\boldsymbol{A})$ とは，$\boldsymbol{A}\boldsymbol{z} = 0$ を満たす $\boldsymbol{z}$ の集合である．$N(\boldsymbol{A})$ を $\boldsymbol{A}$ の零空間 (null space) という．

$R(\boldsymbol{A}^T)$ と $N(\boldsymbol{A})$ は直交している．$\boldsymbol{A}$ を $n\times m$ の行列とするとき，m 次元のベクトルは $R(\boldsymbol{A}^T)$ に属する成分と，$N(\boldsymbol{A})$ に属する成分の和に表すことができる．これを証明するには $\boldsymbol{A}$ の特異値分解

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$$

を利用する．$\boldsymbol{V}$ は $m\times m$ の行列である．任意の m 次元ベクトル $\boldsymbol{x}$ は $\boldsymbol{V}$ の列ベクトル $\boldsymbol{v}_i$ によって展開することができる．このうち特異値が 0 に属するベクトル $\boldsymbol{v}_i$ は $\boldsymbol{A}\boldsymbol{v}_i = 0$ であるから，この成分は $N(\boldsymbol{A})$ に属している．特異値が 0 でない成分は，$\boldsymbol{U}$ の列を $\boldsymbol{u}_i$，特異値を λ_i とすれば

$$\boldsymbol{A}^T\boldsymbol{u}_i = \lambda_i\boldsymbol{v}_i \qquad \lambda_i > 0$$

が成り立っているから，$R(\boldsymbol{A}^T)$ に属している．

補題 1　$\boldsymbol{x} = \boldsymbol{A}^\dagger\boldsymbol{b}$ は $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|$ を極小にする $\boldsymbol{x}$ の中で $\|\boldsymbol{x}\|$ が極小となる解である．

これは §11 の特異値分解と一般逆行列の定義から明らかである．

補題 2

$$\text{(i) } \boldsymbol{x} \in R(\boldsymbol{A}^T) \quad \text{かつ} \quad \text{(ii) } (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) \perp R(\boldsymbol{A})$$

同じことであるが

$$\text{(i) } \boldsymbol{x} \perp N(\boldsymbol{A}) \quad \text{かつ} \quad \text{(ii) } (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) \in N(\boldsymbol{A}^T)$$

を満たす唯一のベクトルは $\boldsymbol{x} = \boldsymbol{A}^\dagger\boldsymbol{b}$ である．

条件 (ii) は $\boldsymbol{x}$ が正規方程式 $\boldsymbol{A}^T(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) = 0$ を満たしていることを意味している．ある $\boldsymbol{x}$ が正規方程式を満たしているとき，$\boldsymbol{x}$ に $\boldsymbol{A}\boldsymbol{x}' = 0$ を満たす $\boldsymbol{x}'$ を加えても正規方程式の解になっている．条件 (i) は $\boldsymbol{x}$ に $N(\boldsymbol{A})$ 成分が含まれていないことを意味しているので，定理が成立する．

補題 3　方程式

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b}$$

の解が $\boldsymbol{x} \in R(\boldsymbol{A}^T)$ なら，実は

$$\boldsymbol{x} = \boldsymbol{A}^\dagger\boldsymbol{b}$$

である．

補題 4　方程式

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \quad \text{ただし} \quad \boldsymbol{b} \in R(\boldsymbol{A})$$

の解が $\boldsymbol{x} \in R(\boldsymbol{A}^T)$ なら実は，

$$\boldsymbol{x} = \boldsymbol{A}^\dagger\boldsymbol{b}$$

である．