7　最小二乗法とデータの補間

実験結果の解析にあたって、実験値や実験曲線を数式に表す必要がある。例えば、多結晶は結晶粒径が小さいほど降伏応力が大きくなるという Hall–Petch（ホール　ペッチ）則　$\sigma_y = \sigma_0 + Kd^{1/2}$ は良く知られているが、粒径が様々に異なる試料を用意して降伏強度を測定したとして、上式の係数 σ_0 と K がどの様な値を取るか分らなければ解析はこれ以上進まない。この例では決定すべきパラメータは二つなので、異なる粒径を持つ二つの試料の強度を測定すればこれらのパラメータは決定できるが、通常は信頼性の高い実験値を得るために粒径の異なるもっと多くの試料に対して強度を測定する。二つのパラメータを決めるのにデータが三個以上あれば、全ての測定点を通る曲線を引くことはできない。ではどうすれば良いだろうか。ここでは、一連の実験データが与えられた時、最も確からしい実験式を作る方法として、最小２乗法を学ぶ。この方法は、数少ない観測データから、小惑星の軌道を見積もる方法として、ガウスによって考案されたものである。

7.1　最小二乗法の考え方

最小二乗法は、学生実験の「状態図の作成」の項目で、熱電対の温度特性を較正する際に用いた。復習するとこうである。即ち、K 熱電対の熱起電力は 500°C 以下の温度域ではほぼ線型の温度依存性を持つので、特性曲線を $E = aT + b$ と書くことが出来ると仮定する。温度定点 $T_1 = 0$°C (水の融点)、$T_2 = 100$°C (水の沸点)、$T_3 = 232$°C (スズの融点)、$T_3 = 419$°C (亜鉛の融点)、における熱起電力 E_1、E_2、E_3、E_4 を測定し、得られた四つのデータ点 (T_1, E_1)、(T_2, E_2)、(T_3, E_3)、(T_4, E_4) の「全てに最も近くなる」ように $E = aT + b$ のパラメータ a と b を選ぶ。ここで 4 点の全てに最も近いというのは残差と呼ばれる量の二乗和、

$$S = \sum_{i=1}^{4} [E_i - (aT_i + b)]^2$$

を最小にするように直線を選ぶということである。S が最小になる条件は、上式を a 及び b で微分して

$$\begin{aligned} \frac{\partial S}{\partial a} &= 2a\sum_{i=1}^{4} T_i^2 + 2b\sum_{i=1}^{4} T_i - 2\sum_{i=1}^{4} E_i T_i = 0 \\ \frac{\partial S}{\partial b} &= 2a\sum_{i=1}^{4} T_i + 2b\sum_{i=1}^{4} 1 - 2\sum_{i=1}^{4} E_i = 0 \end{aligned} \tag{1}$$

で与えられる。これは a、b を未知数とする連立二元一次方程式であり、これを解けば実験結果に「最も良く」フィットする特性直線が得られる。

上記の処方箋は容易に一般の関数形に拡張することができる。即ち、フィッティングをすべき理論式が $y = f(x; a_1, a_2, \ldots, a_n)$ と書けるのであれば、これへの最良のフィットは連立方程式

$$\frac{\partial S}{\partial a_i} = 2\sum_{j=1}^{N} [f(x_j; a_1, a_2, \ldots, a_n) - y_j] \left.\frac{\partial f(x; a_1, a_2, \ldots, a_n)}{\partial a_i}\right|_{x=x_j} = 0 \quad (i = 1, 2, \ldots, n)$$

を解くことで得られる。ただし (x_i, y_i)　$(i = 1, 2, \ldots, N)$ は実験データであり、N および n はそれぞれデータ点とパラメータ変数の数である。これも $a_i\,(i = 1, 2, \ldots, n)$ を未知数とする連立方程式ではあるが、一般には非線形な方程式、つまり一次方程式ではない連立方程式となり、解くのはやっかいである。しかし $f(x)$ が x の多項式 $y = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$ で表わされる時には、問題は簡単に

なり、最適な関数を決める条件式は行列を用いて

$$\begin{pmatrix} \sum_{j=1}^{N} x_j^{2n} & \sum_{j=1}^{N} x_j^{2n-1} & \cdots & \sum_{j=1}^{N} x_j^{n+1} & \sum_{j=1}^{N} x_j^{n} \\ \sum_{j=1}^{N} x_j^{2n-1} & \sum_{j=1}^{N} x_j^{2n-2} & \cdots & \sum_{j=1}^{N} x_j^{n} & \sum_{j=1}^{N} x_j^{n-1} \\ \cdots & & & & \\ \sum_{j=1}^{N} x_j^{n+1} & \sum_{j=1}^{N} x_j^{n} & \cdots & \sum_{j=1}^{N} x_j^{2} & \sum_{j=1}^{N} x_j^{1} \\ \sum_{j=1}^{N} x_j^{n} & \sum_{j=1}^{N} x_j^{n-1} & \cdots & \sum_{j=1}^{N} x_j^{1} & \sum_{j=1}^{N} x_j^{0} \end{pmatrix} \begin{pmatrix} a_n \\ a_{n-1} \\ \cdots \\ a_1 \\ a_0 \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{N} y_j x_j^{n} \\ \sum_{j=1}^{N} y_j x_j^{n-1} \\ \cdots \\ \sum_{j=1}^{N} y_j x_j^{1} \\ \sum_{j=1}^{N} y_j x_j^{0} \end{pmatrix}$$

と書くことができる。この方程式を正規方程式と呼ぶ。正規方程式は Gauss の消去法など、連立一次方程式の解法を解説した時の手法で解くことができる。

練習問題 19: ある気体について、温度 T と定圧比熱 C_p との間に次のような 6 組の実験データが得られている。これらのデータに基づいて $C_p = a_0 + a_1 T + a_2 T^2$ の形の実験式を最小二乗法によって求めよ。

T (°C)	0	100	200	300	400	500
C_p (cal/g·°C)	0.07	0.12	0.16	0.19	0.21	0.22

練習問題 20: 次の表は、自動車の停止距離と時速との関係である ((財) 全日本交通安全協会編、警察庁交通局監修「交通の教則」より引用)。これを二次曲線で近似し、時速 45km の時の停止距離を求めよ。

速度 (km/h)	20	40	60	80	100
停止距離 (m)	9	22	44	76	112

7.2 spline (スプライン) 補間

強度の粒径依存や拡散係数の温度依存のように実験結果が従うべき原理や法則が知られているなら、前節の最小二乗法は有力な方法であるが、どのような式を用いて実験結果をフィッティングすれば良いか分らない場合も多い。そのような時にプロットを滑かに繋ぐ方法を一つ紹介する。

例えば表 1 のような測定データが得られたとしよう。何らかの法則に従ったデータには見えない。取り敢えずデータ点同士を直線で結んでみると、図 1(A) のようなグラフが得られる。当然のことながら折れ線になってしまい、かなり不恰好である。数値積分をする時に放物線を用いると結果が改善されたので、ここでも二次式を用いて点と点をつないだら格好良くなるだろうか。隣合う 3 点を通る放物線を順に求めて線を引くと図 1(B) のようなグラフが得られた。前より格好が良いとは言い難い。異なる

表 1: 何かの測定データ。実は乱数を用いて生成したデータなので、規則性はない。

x	0.000	0.125	0.277	0.349	0.513	0.607	0.702	0.773	0.860	1.000
y	3.502	5.551	5.258	4.898	5.094	4.312	4.075	4.577	5.197	4.738

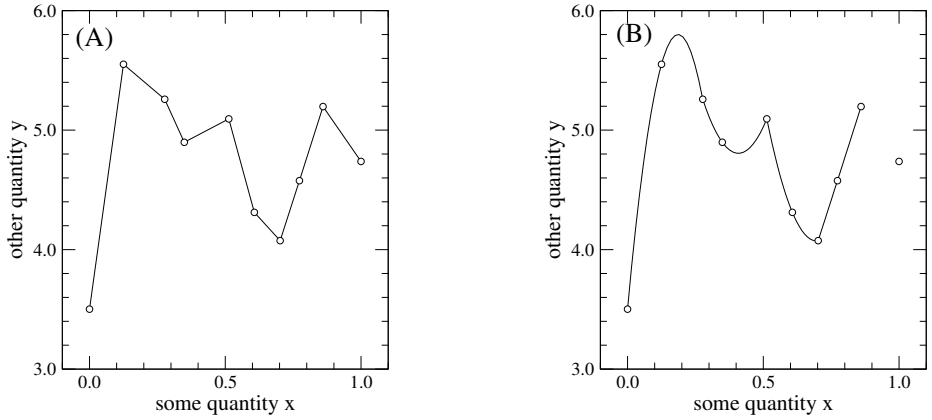

図 1: 測定データの直線 (A) と放物線 (B) による補間。「繋ぎ目」のところで折れ曲がっていて不自然な曲線になってしまう。

区間でフィッティングした線同士が出会う場所で傾きが違うと、曲線は折れ曲って見えてしまうのである。そこで、隣り合う点同士を多項式で表わされる曲線で結んで、傾きが不連続にならないように線を決定する方法を考えよう。

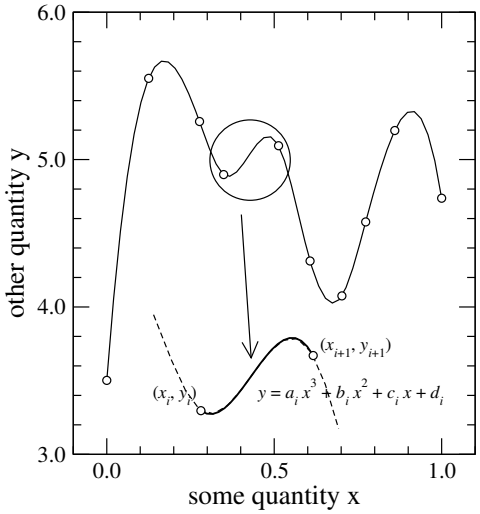

図 2: 3次spline関数によるデータ点の補間。

滑かに繋ぎたいデータを (x_1, y_1)、(x_2, y_2)、…、(x_n, y_n) の n 個とする。今、点 (x_i, y_i) と点 (x_{i+1}, y_{i+1}) の間の区間を考えよう。求める曲線はこの区間の両端の点を通り、かつ両端で隣りの曲線と等しい傾きを持たねばならぬので、四つの条件を満さなくてはならない。四つの条件を自由に決められる最低次の多項式は三次式なので、i 番目の点と $i+1$ 番目の点の間の区間を式 $y = a_i x^3 + b_i x^2 + c_i x + d_i \quad (i = 1, 2, \ldots, n-1)$ で補間することにする。この区間の両端で与えられたデータ点 (x_i, y_i)、(x_{i+1}, y_{i+1}) を通る条件は

$$\begin{aligned} y_i &= a_i x_i^3 + b_i x_i^2 + c_i x_i + d_i \\ y_{i+1} &= a_i x_{i+1}^3 + b_i x_{i+1}^2 + c_i x_{i+1} + d_i \end{aligned}$$

と書ける。$i = 1, 2, \ldots, n-1$ に対してこの形の式が成り立つので、条件式の数は $2n-2$ 個ある。一方、点 (x_i, y_i) において傾きが一致する条件は

$$3a_{i-1}x_i^2 + 2b_{i-1}x_i + c_{i-1} = 3a_i x_i^2 + 2b_i x_i + c_i$$

と書くことができる。この条件式は $i = 2, 3, \ldots, n-1$ に対して成立するので、式の数は $n-2$ 個である。未知数は $a_1, b_1, c_1, d_1, a_2, b_2, c_2, d_2, \ldots, a_{n-1}, b_{n-1}, c_{n-1}, d_{n-1}$ の $4n-4$ 個なので、ここまでの条件式だけでは全てを決めることはできない。更に各点において、曲線の二階微分まで一致することを要請すると

$$6a_{i-1}x_i + 2b_{i-1} = 6a_i x_i + 2b_i$$

の形の $n-2$ 個の式が得られる。更に、全区間の両端 $x = x_1$ と x_n で二階微分が 0 になる条件、

$$6a_1 x_1 + 2b_1 = 6a_{n-1}x_n + 2b_{n-1} = 0$$

を課すと全ての係数を決定することができて、点を滑かに繋ぐことができる。このように両端で曲線の二回微分が 0 になると言う条件を課して曲線を決定したものを自然スプライン曲線と言う。上に出て来た連立方程式を行列の形式で書くと、

$$
\begin{pmatrix} y_1 \\ y_2 \\ y_2 \\ y_3 \\ y_3 \\ y_4 \\ y_4 \\ \cdots \\ y_{n-2} \\ y_{n-2} \\ y_{n-1} \\ y_{n-1} \\ y_n \\ 0 \\ 0 \\ \cdots \\ 0 \\ 0 \end{pmatrix}
=
\begin{pmatrix}
x_1^3 & x_1^2 & x_1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
x_2^3 & x_2^2 & x_2 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & x_2^3 & x_2^2 & x_2 & 1 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & x_3^3 & x_3^2 & x_3 & 1 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
 & & & & & & & \cdots & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & x_{n-1}^3 & x_{n-1}^2 & x_{n-1} & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & x_n^3 & x_n^2 & x_n & 1 \\
 & & & & & & & \cdots & & & & & \\
3x_2^2 & 2x_2 & 1 & 0 & -3x_2^2 & -2x_2 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 3x_3^2 & 2x_3 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\
 & & & & & & & \cdots & & & & & \\
6x_1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
6x_2 & 2 & 0 & 0 & -6x_2 & -2 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 6x_3 & 2 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
 & & & & & & & \cdots & & & & & \\
0 & 0 & 0 & 0 & & & & \cdots & & 6x_{n-1} & 2 & 0 & 0 \\
0 & 0 & 0 & 0 & & & & \cdots & & 6x_n & 2 & 0 & 0
\end{pmatrix}
\begin{pmatrix} a_1 \\ b_1 \\ c_1 \\ d_1 \\ a_2 \\ b_2 \\ c_2 \\ d_2 \\ a_3 \\ b_3 \\ c_3 \\ d_3 \\ a_4 \\ \cdots \\ a_{n-1} \\ b_{n-1} \\ c_{n-1} \\ d_{n-1} \end{pmatrix}
$$

が得られる。Gauss の消去法などを用いてこの $4(n-1)$ 元連立一次方程式を解くと, 曲線の係数が決まりグラフを描くことができる。[1]

練習問題 21: 下の表のようなデータを 3 次 spline 関数で補間せよ。補間して得られた 3 次多項式の値を、0 から 1 まで 0.01 刻みで計算して出力できれば良い。

x	0.00	0.25	0.50	0.75	1.00
y	0.0000	0.9179	0.9933	0.9994	1.0000

実は上の問題のデータは関数 $y = 1 - \exp(-10x)$ 上の点を 5 個取ったものである。指数関数は多項式に比べて急激に変化するので、指数関数的に変化する量を spline 補間すると曲線は波を打つような形になることがある (図 3)。パソコンソフトを用いて点をなめらかにつないだ結果出現した波打ちを見て、「収束する前に小さな極大値を持つ」などと結論してはいけない。時間と共に指数関数的な減少をする量というものは世の中に沢山あるので要注意である。

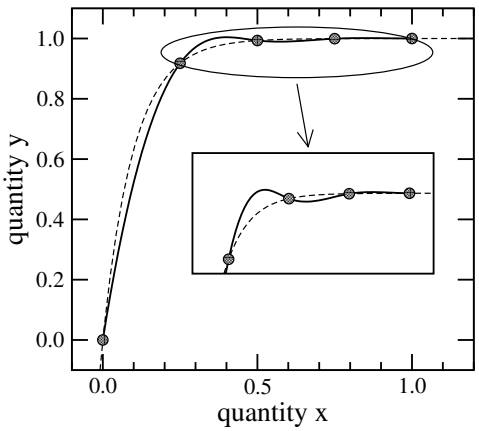

図 3: 3 次 spline 関数による指数関数の補間。データが収束しかける所で「波打ち」が見られる。

[1] 難しくはないのだけれど、面倒な結果になってしまった。アイディアを単純に留めたいがために、工夫が足りないことが原因である。少し工夫をすれば、ずっと綺麗な形で同じ結果を得ることができる。数値計算の教科書やネット等で調べて欲しい。