サンプリング探索法

１．概要

ソートされたデータの探索法としては、ほとんど二分探索法が用いられますが、他にないか自分で考えてみました。

昇順にソートされたデータを対象として、ｘ軸にデータの順番（ｘ＝１～ｎ）をとり、ｙ軸にデータの値をとった平面を考えます。このデータの集合（x_i、y_i）（ｉ＝１～ｎ）の中から適当にm個をサンプリングして、最小二乗法により３次近似式 $y = a x^3 + b x^2 + c x + d$ の係数を求めます。

ここで探索したいデータ y_0を３次近似式に代入して根を求めます。この根が求めるデータの順番になりデータが発見されたことになります。

ただし近似式なので、根も近似解となり根の周辺をさらに探索して真の解を求めます。データの分布状況により近似解の精度が変わってきます。

おおむねランダムや正規分布のデータでは精度は比較的良くなります。また、サンプリングの個数を増やせば精度は良くなります。

２．最小二乗法により３次近似式を求める

データの集合（x_i、y_i）（ｉ＝１～ｎ）が与えられたとき、この各点の近くを通る３次近似式 $f(x) = a x^3 + b x^2 + c x + d$ の係数を求めます。

データと３次近似式との残差の２乗和、

$$J = \sum_{i=1}^{n} \{ y_i - f(x_i) \}^2$$

においてＪが最小になるように係数を定めます。

$$\frac{\delta J}{\delta a_j} = 0 \qquad (j = 1 \sim 4)$$

これより未知数ａ、ｂ、ｃ、ｄに関する４元連立１次方程式（正規方程式）を解いて未知数を求めます。

$$\begin{bmatrix} \Sigma & \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 & \Sigma x_i^5 \\ \Sigma x_i^3 & \Sigma x_i^4 & \Sigma x_i^5 & \Sigma x_i^6 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} \Sigma f(x_i) \\ \Sigma f(x_i)\, x_i \\ \Sigma f(x_i)\, x_i^2 \\ \Sigma f(x_i)\, x_i^3 \end{bmatrix}$$

３．３次方程式の根を求める

３次方程式　$a\ x^3+b\ x^2+c\ x+d=0$　－①　において、

$x=t-\dfrac{b}{3a}$　　とおくと①式は

$$t^3+pt+q=0 \quad -②$$

$$p=\frac{1}{a}\left(c-\frac{b^2}{3a}\right) \qquad q=\frac{1}{a}\left(\frac{2b^3}{27a^2}-\frac{bc}{3a}+d\right)$$

となります。

②式を解いて（カルダノの解法)、３つの根を求めます。

$$t=-\sqrt[3]{r}-\sqrt[3]{s}$$
$$t=-\omega\sqrt[3]{r}-\omega^2\sqrt[3]{s}$$
$$t=-\omega^2\sqrt[3]{r}-\omega\sqrt[3]{s}$$

となる。ただし、

$$r=\frac{q+\sqrt{q^2+(4/27)p^3}}{2}$$

$$s=\frac{q-\sqrt{q^2+(4/27)p^3}}{2}$$

$$\omega=\frac{-1+i\sqrt{3}}{2}$$

です。

ここで判別式　$D=-(4p^3+27q^3)$　において、

$D<0$のとき　１つの実根と２つの虚根となります。

実根は　$t=-\sqrt[3]{r}-\sqrt[3]{s}$

$D>0$のとき　３つの実根となります。

$$t=2\sqrt[6]{a^2+b^2}\cos\frac{\theta}{3}$$

$$t=-2\sqrt[6]{a^2+b^2}\cos\left(\frac{\pi}{3}-\frac{\theta}{3}\right)$$

$$t=-2\sqrt[6]{a^2+b^2}\cos\left(\frac{\pi}{3}+\frac{\theta}{3}\right)$$

$$\theta=\tan^{-1}\left(\frac{b}{a}\right)$$

$D=0$のとき　１つの３重根となります。

$$t=-2\sqrt[3]{(q/2)}$$

です。

４．探索のアルゴリズム

与えられたデータ（ｉ、y_i）（ｉ＝１～ｎ）に最小二乗法を適用することによって３次近似式が求められました。

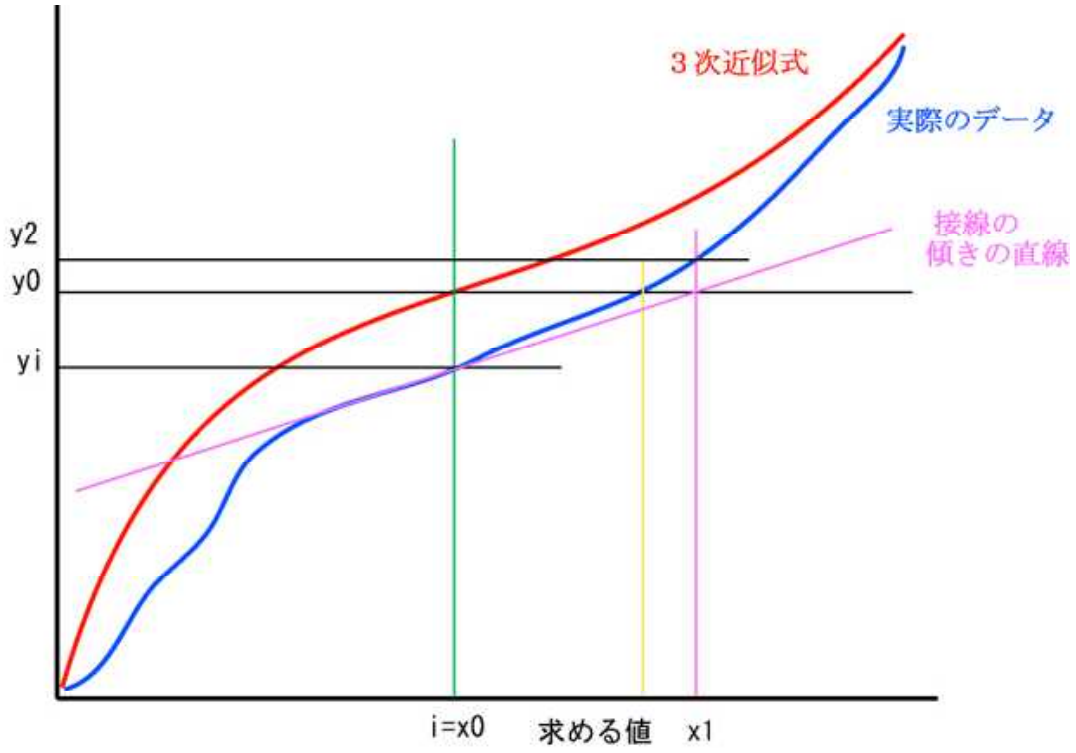
アルゴリズムＡは、判別式Ｄ＜＝０の場合（１実根）に用います。

（１）調べたいデータy_0を３次近似式に代入してその根を求めます。この根の値x_0がy_0の番号ｉになります。

$y_i = y_0$ ならば、データを発見。

（２）それ以外ならば、x_0において３次近似式の接線の傾きk_1を求めて、点（x_0、y_i）を通る傾きk_1の直線$y = k_1 x + m_1$を求めます。この直線の式にy_0を代入してx_1を求めます。

ｉ＝ x_1 のときの y_i を求めてy_2とします。



（３）$y_0 < y_2$ ならば、

$$x_0 = \frac{x_0 + x_1}{2}$$ として（１）へ行きます。

$y_0 > y_2$ ならば、x_1において３次近似式の接線の傾きk_2を求めて、点（x_1、y_2）を通る傾きk_2の直線$y = k_2 x + m_2$を求めます。この直線の式にy_0を代入してx_2を求めます。

ｉ＝x_2 のときの y_i を求めてy_2として、（３）へ戻ります。

$y_0 = y_2$ ならば、データを発見。

以上がアルゴリズムであるが、判別式Ｄ＞０になった場合には３つの実根が存在するので、アルゴリズムＢを用います。

（4）３次近似式の２つの極値を求めます。

３次近似式を微分した　$y' = 3ax^2 + 2bx + c = 0$　の２根x_1、x_2（$x_1 < x_2$）が極値となります。

$$x = \frac{-b \pm \sqrt{b^2 - 3ac}}{3a}$$

$i = x_1$ のときの y_i を求めてy_1 とする。$i = x_2$ のときの y_i を求めてy_2 とします。

（5）$y_0 > y_2$ または$y_0 < y_1$ ならば、アルゴリズムAを用います。

（6）$y_1 < y_0 < y_2$ ならば、区間$x_1 \sim x_2$ を直線の式で近似します。

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) \quad -③$$

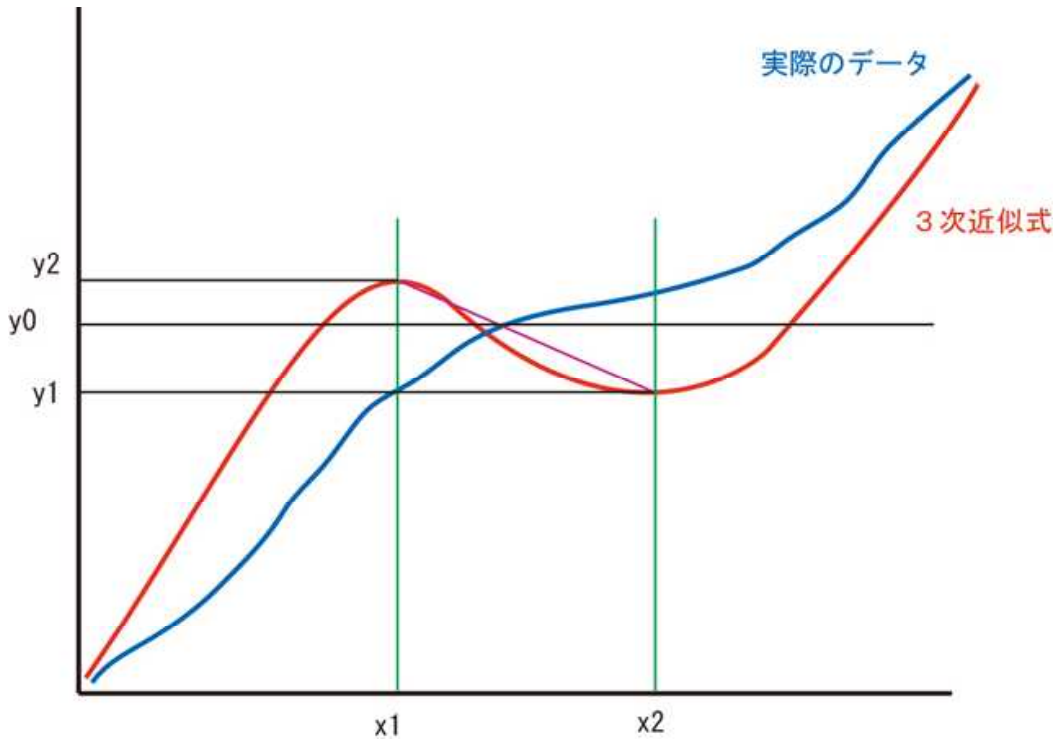
（7）

$x_0 = \frac{x_1 + x_2}{2}$　として、③式に代入して y_3 を求めます。

$y_0 = y_3$ ならば、データを発見。

$i = x_0$ のときの y_i を求めてy_4 とします。

（8）$y_0 > y_4$ ならば $x_1 = x_0$ として、$y_0 < y_4$ ならば$x_2 = x_0$ として（7）に戻ります。



５．その他

サンプリング探索法では、データの分布が偏ると近似が悪くなり探索により多くの回数がかかるようになります。

データの値が大きくなるとオーバーフローする可能性があり、最小二乗法を使うときに行列を用いるのでメモリオーバーになるかもしれません。

また文字列の探索には、あまり向いていないように思われます。

近似式として３次式を用いたのは、必ず実根が１つはあり、ほぼ（３実根のときは一部で減少）単調増加であるという理由によります。