最小二乗近似 2

最小二乗法の概要

実変数 x_k に対してデータの値が y_k $(k = 1, 1, 2, \cdots, N)$ であるとする.

この事象が，関数 $y = f(x)$ で表現できるとする（近似関数）.

データ (x_k, y_k) とその理論値（近似値） $y = f(x)$ が与えられるとき，

標本数（サンプル数）をNとして，その誤差の分散は

$$J = \sum_{j=1}^{N} |f(x_j) - y_j|^2$$

と定義される.

J/Nを二乗平均誤差という.

最小二乗法（最小誤差近似）は，二乗平均誤差（データの分散）を最小にする近似である.

近似関数を1次式とする場合（連立方程式を解く）

近似関数として1次式 $f(x) = ax + b$ を仮定すれば， $J = \sum_{j=1}^{N} |(ax_j + b) - y_j|^2$.

J を a, b でそれぞれ偏微分して0と置き $\left(\dfrac{\partial J}{\partial a} = 0, \quad \dfrac{\partial J}{\partial b} = 0\right)$

$$\begin{pmatrix} {}^t\mathbf{xx} & {}^t\mathbf{xu} \\ {}^t\mathbf{xu} & {}^t\mathbf{uu} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} {}^t\mathbf{xy} \\ {}^t\mathbf{uy} \end{pmatrix}$$

を得る．ただし $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ とする.

これから a, b が定まる.

具体的に書けば

$$\begin{aligned}
&{}^t\mathbf{ux} = {}^t\mathbf{xu} = \sum x_i \equiv N\overline{x}, \qquad {}^t\mathbf{uy} = {}^t\mathbf{yu} = \sum y_i \equiv N\overline{y} \\
&{}^t\mathbf{xx} = \sum x_i^2 \equiv N(V(x) + \overline{x}^2) = N(\sigma_{xx} + \overline{x}^2) \\
&{}^t\mathbf{xy} = \sum x_i y_i \equiv N(\sigma_{xy} + \overline{x}\ \overline{y}) = N(\sigma_{yx} + \overline{x}\ \overline{y}) \\
&{}^t\mathbf{uu} = N.
\end{aligned}$$

ただし $\sigma_{xx} = \dfrac{1}{N} \sum (x_i - \overline{x})^2$, $\sigma_{xy} = \dfrac{1}{N} \sum (x_i - \overline{x})(y_i - \overline{y})$ と定義する.

ここで ${}^t\mathbf{xx} \cdot {}^t\mathbf{uu} - ({}^t\mathbf{xu})^2 = N^2 \sigma_{xx}$ であるから，

$$\begin{pmatrix} a \\ b \end{pmatrix} = \frac{1}{N^2\sigma_{xx}} \begin{pmatrix} N & -N\overline{x} \\ -N\overline{x} & N(\sigma_{xx}+\overline{x}^2) \end{pmatrix} \begin{pmatrix} N(\sigma_{yx}+\overline{x}\ \overline{y}) \\ N\overline{y} \end{pmatrix} = \begin{pmatrix} \sigma_{yx}/\sigma_{xx} \\ \overline{y}-a\overline{x} \end{pmatrix}$$

を得る.

「$(x_i, y_i) \quad (i = 1, 2, \cdots, N)$）全てを通る直線を決めたい」

$x_1 a + b = y_1$

$\vdots$

$x_N a + b = y_N$

N個の条件で2つの定数a, b を決めるのであるから，一般に解は無い.

しかしひとまずこの式を行列形式で書けば

$M\mathbf{a} = \mathbf{y}$

となる．ただし $M = (\mathbf{x}\ \ \mathbf{u}) = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} a \\ b \end{pmatrix}$ と定義する.

行列形式の立場からいっても，M は正方行列ではないので，この式には一般に解は無い.

この式に左から ${}^tM$ をかければ，それぞれ

$${}^tMM = {}^t(\mathbf{x}\ \ \mathbf{u})(\mathbf{x}\ \ \mathbf{u}) = \begin{pmatrix} {}^t\mathbf{x} \\ {}^t\mathbf{u} \end{pmatrix} (\mathbf{x}\ \ \mathbf{u}) = \begin{pmatrix} {}^t\mathbf{x}\mathbf{x} & {}^t\mathbf{x}\mathbf{u} \\ {}^t\mathbf{u}\mathbf{x} & {}^t\mathbf{u}\mathbf{u} \end{pmatrix}$$

$${}^tM\mathbf{y} = {}^t(\mathbf{x}\ \ \mathbf{u})\mathbf{y} = \begin{pmatrix} {}^t\mathbf{x} \\ {}^t\mathbf{u} \end{pmatrix}\mathbf{y} = \begin{pmatrix} {}^t\mathbf{x}\mathbf{y} \\ {}^t\mathbf{u}\mathbf{y} \end{pmatrix}$$

を得る．これで前式 $\begin{pmatrix} {}^t\mathbf{x}\mathbf{x} & {}^t\mathbf{x}\mathbf{u} \\ {}^t\mathbf{x}\mathbf{u} & {}^t\mathbf{u}\mathbf{u} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} {}^t\mathbf{x}\mathbf{y} \\ {}^t\mathbf{u}\mathbf{y} \end{pmatrix}$ は

$${}^tMM\mathbf{a} = {}^tM\mathbf{y}$$

であることが分かる．${}^tMM$ は正方行列であり，この方程式は必ず解をもつ.

簡単な例題

$x - y$ 平面の3点（$1, 1.5$）, (2,2.5), (3, 3) を結ぶ最良の直線の方程式を求めよう.

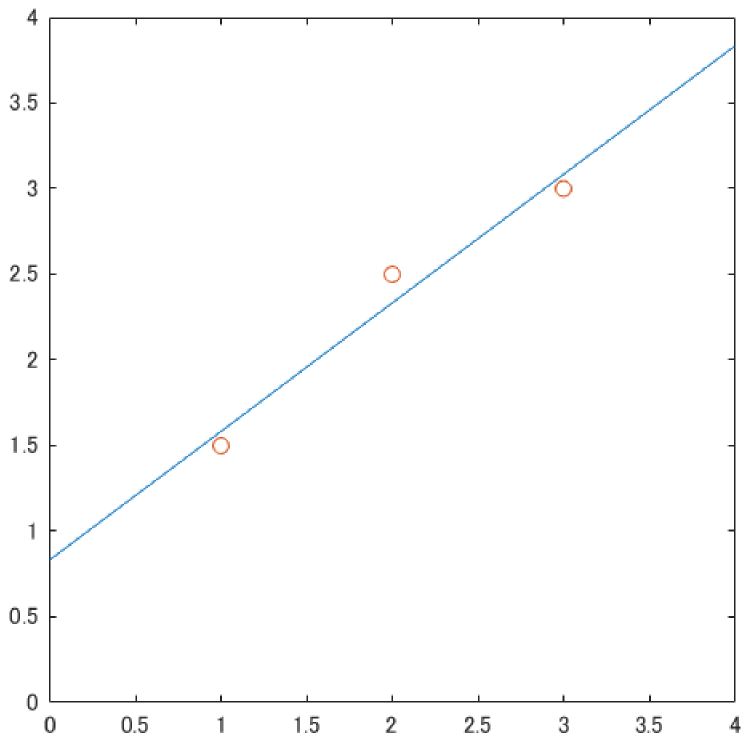
```
x=[1;2;3];
y=[1.5;2.5;3.0];
u=[1;1;1];
M=[x u];
```

```
a=(M'*M)\(M'*y)
```

```
a = 2×1
    0.7500
    0.8333
```

解に対して3点と，得られた直線を図示せよ.

```
f = @(x,y) y-a(1,1)*x-a(2,1);
fimplicit(f,[0 4 0 4])
hold on
px=[1 2 3];
py=[1.5 2.5 3];
plot(px,py,'o')
daspect([1 1 1])
hold off
```



一般の場合

与えられたデータ点$(x_1, x_2, \cdots, x_n)$ におけるデータ y_i $(i = 1, 2, \cdots, n)$ が与えられている.

このときデータを多項式 $y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m$ と表すことを考える.

評価関数

$$L = \sum_{i=1}^{n} \{y_i - (a_0 + a_1 x_i + \cdots + a_m x_i^m)\}^2$$

を最小にするように $a_0, a_1, \cdots$ を決めよう.

$$\mathbf{x}^{(k)} = \begin{pmatrix} (x_1)^k \\ (x_2)^k \\ \vdots \\ (x_n)^k \end{pmatrix}, \quad k = 0, 1, \cdots m, \quad M = [\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}]$$

$$\mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

とおけば,

$$M \cdot \mathbf{a} = \mathbf{x}^{(0)} a_0 + \mathbf{x}^{(1)} a_1 + \mathbf{x}^{(2)} a_2 + \cdots + \mathbf{x}^{(m)} a_m$$

であることに注意すれば,

$$L =^t (\mathbf{y} - M \cdot \mathbf{a}) \cdot (\mathbf{y} - M \cdot \mathbf{a})$$

を得る. これから $\mathbf{a}$ の各成分で微分して

$${}^t M M \mathbf{a} = {}^t M \mathbf{y}$$

を得る.

$${}^t M M = \begin{pmatrix} \mathbf{x}^{(0)} \\ \mathbf{x}^{(1)} \\ \vdots \\ \mathbf{x}^{(m)} \end{pmatrix} (\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)}) = \begin{pmatrix} {}^t\mathbf{x}^{(0)}\mathbf{x}^{(0)} & {}^t\mathbf{x}^{(0)}\mathbf{x}^{(1)} & \cdots & {}^t\mathbf{x}^{(0)}\mathbf{x}^{(m)} \\ {}^t\mathbf{x}^{(1)}\mathbf{x}^{(0)} & {}^t\mathbf{x}^{(1)}\mathbf{x}^{(1)} & \cdots & {}^t\mathbf{x}^{(1)}\mathbf{x}^{(m)} \\ \vdots & & & \\ {}^t\mathbf{x}^{(m)}\mathbf{x}^{(0)} & {}^t\mathbf{x}^{(m)}\mathbf{x}^{(1)} & \cdots & {}^t\mathbf{x}^{(m)}\mathbf{x}^{(m)} \end{pmatrix}$$

ただし

$${}^t M \mathbf{y} = \begin{pmatrix} \mathbf{x}^{(0)} \\ \mathbf{x}^{(1)} \\ \vdots \\ \mathbf{x}^{(m)} \end{pmatrix} \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix} = \begin{pmatrix} \mathbf{x}^{(0)} y_1 & \mathbf{x}^{(0)} y_2 & \cdots & \mathbf{x}^{(0)} y_n \end{pmatrix}$$

に注意.

これにより

$$\mathbf{a} = ({}^t M M)^{-1} \; {}^t M \mathbf{y}$$

が得られる. ${}^t M M$ をグラム行列という.

上のような方法を, 勿論直接用いても良いが, MATLABでは polyfit というルーチンが

用意されている：

```
p = polyfit(x,y,n)
```

は、y のデータに対して（最小二乗的に）最適な近似となる n 次の多項式 p(x) の係数を返す.

p の係数は降べきの順で、p の長さは n+1 になる.

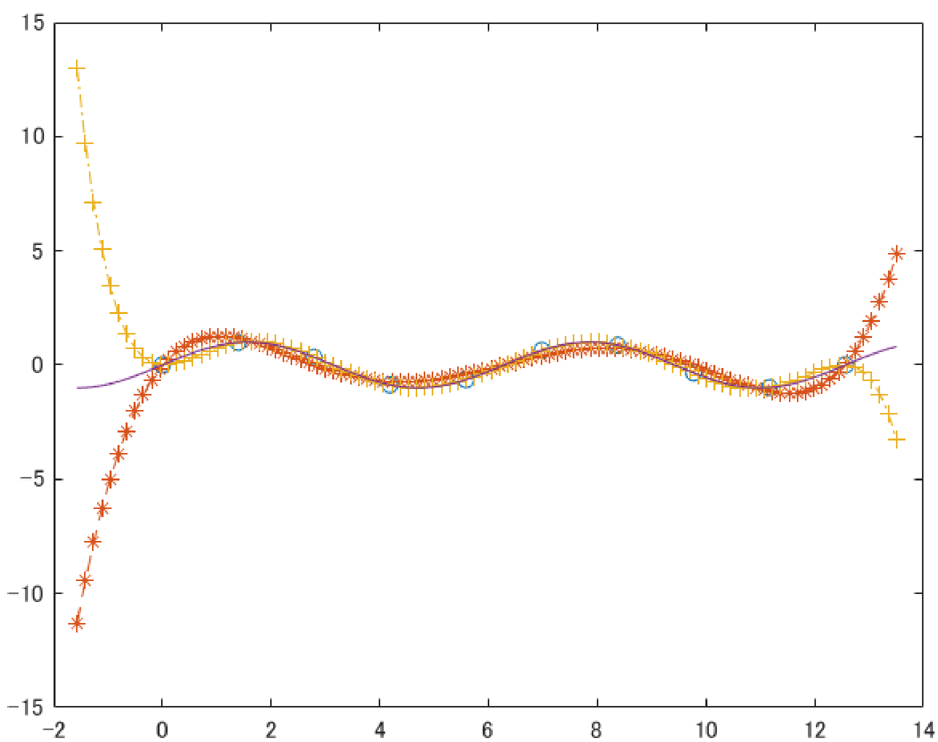
$$p(x) = p_1x^n + p_2x^{n-1} + \cdots + p_nx + p_{n+1}$$

上で与えた a_k のノーテーションと異なるので注意を要する.

例 1

等間隔10点で与えられているsin関数を5次および7次多項式で近似する：

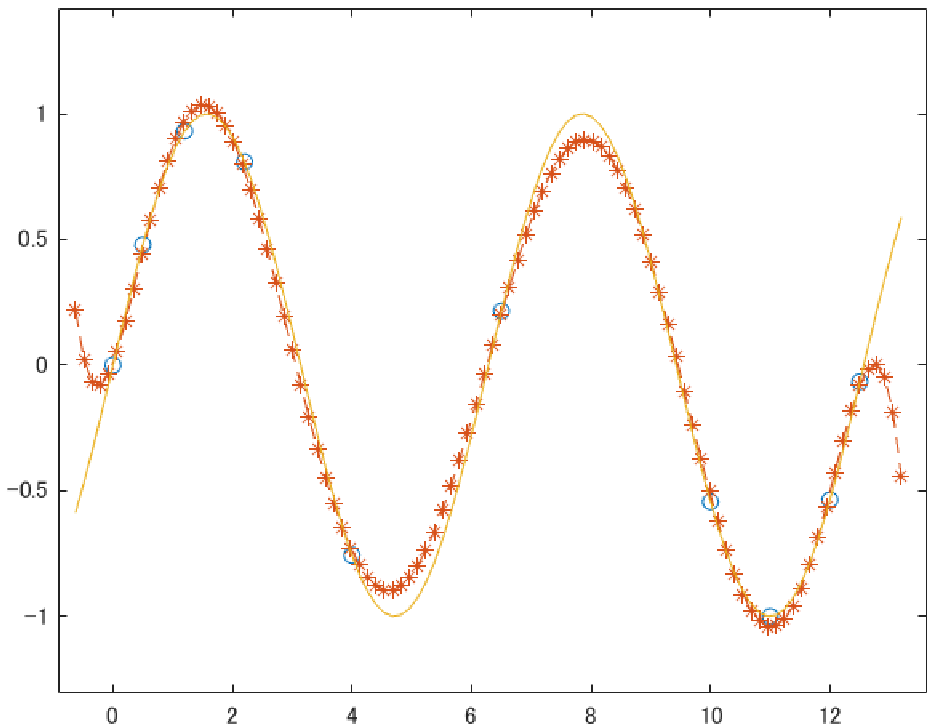
```
x = linspace(0,4*pi,10);
y = sin(x);
p = polyfit(x,y,5);
q = polyfit(x,y,7);
x1 = linspace(-0.5*pi,4.3*pi);
yp = polyval(p,x1);
yq = polyval(q,x1);
y1 = sin(x1);
%figure
plot(x,y,'o');hold on
plot(x1,yp,'--*');hold on
plot(x1,yq,'-.+');hold on
plot(x1,y1);hold off
```



例２

同じく不等間隔10点で与えられたsin関数を7次多項式で近似する：

```
x=[0 0.5 1.2 2.2 4 6.5 10 11 12 12.5];
y = sin(x);
p = polyfit(x,y,7);
x1 = linspace(-0.2*pi,4.2*pi);
yp = polyval(p,x1);
y1 = sin(x1);
%figure
plot(x,y,'o');hold on
plot(x1,yp,'--*');hold on
plot(x1,y1);hold off
```



課題

（1）「簡単な例題」の問題で2次式のフィッティングを行え．

（2）上の例で，多項式の点の数に比べてずっと小さな次数に変えて、同じことを実行してみよ．また点の数に比べて大きな次数に変えたらどうか．

（3）上の例で，例示された領域の外ではどうなっているか．xの広い範囲でプロットしてみよ．

（4）様々な初等関数（多項式，三角関数，指数関数）について同様なことをやってみよ．

（5）polyfit を使わず、^tMM を計算して解を求めるプログラムを書いてみよ．