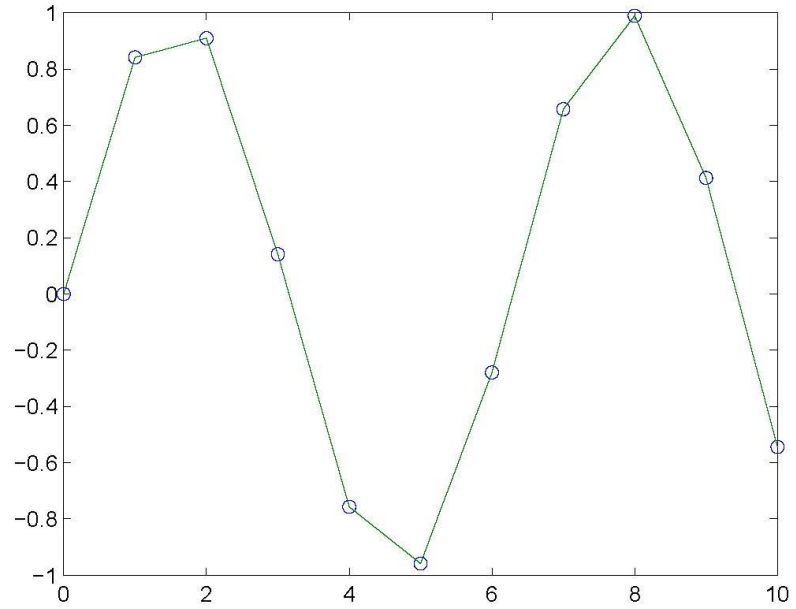
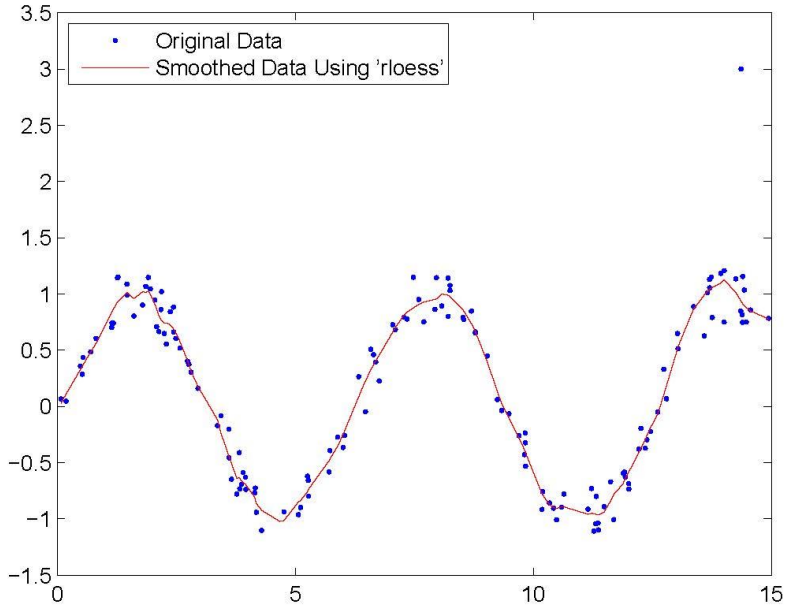
3 数据拟合
Curve Fitting

何军辉
hejh@scut.edu.cn

计算机科学与工程学院
School of Computer Science & Engineering
华南理工大学
South China University of Technology

插值与拟合


1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
0
2
4
6
8
10


Original Data
Smoothed Data Using 'rloess'
3.5
3
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
0
5
10
15

3.1 单变量数据拟合及最小二乘法

- 若给定的数据表表示的是一个量与另一个量的关系，则可以使用单变量数据拟合法寻找一个近似函数来代替函数$f(x)$
- 通常$F(x)$称为拟合函数，$f(x)$称为被拟合函数
  - 与插值法不同，数据拟合并不要求近似函数$F(x)$通过已知数据点
  - 希望能找到一个最好的函数来近似代替$f(x)$
  - 好坏的标准?

3.1 单变量数据拟合及最小二乘法

定义：若记$\delta_i = f(x_i) - F(x_i)$，$i = 1,2,\cdots,n$，则称$\delta_i$为$f(x)$与$F(x)$在$x_i$的偏差.

定义：以“偏差的平方和最小”为原则选择近似函数的方法称为最小二乘法.

$$min \sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} [f(x_i) - F(x_i)]^2$$

3.1 单变量数据拟合及最小二乘法

- 单变量数据拟合法的一般步骤：
  1. 按给定数据表画出散点图
  2. 分析散点图，确定近似函数$F(x)$的类型，以反映给定数据的一般趋势
  3. 用最小二乘法确定近似函数$F(x)$的未知参数，从而得到最小二乘拟合函数$F(x)$

3.1 单变量数据拟合及最小二乘法

□ 最小二乘直线

定理： 给定$y = f(x)$的数据表$(x_i, y_i)(i = 1,2,\cdots,n)$如下，若点$(x_i, y_i)$大体上满足线性函数，即最小二乘拟合函数为

$$F(x) = a + bx$$

则待定参数a和b是正规方程组方程组

$$\begin{cases} na + \left(\sum_{i=1}^{n} x_i\right) b = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right) a + \left(\sum_{i=1}^{n} x_i^2\right) = \sum_{i=1}^{n} x_i y_i \end{cases}$$

3.1 单变量数据拟合及最小二乘法

例：已知一组实验数据如表所示，试用单变量数据拟合法求其拟合函数.

x	-1	0	1	2	3	4	5	6
$y=f(x)$	10	9	7	5	4	3	0	-1

$$F(x)=a+bx$$

3.1 单变量数据拟合及最小二乘法

单变量线性拟合法算法

① 读入数据x_i和$y_i(i=1,2,\cdots,n)$

② 计算

$$s_x=\sum_{i=1}^{n}x_i\,,s_y=\sum_{i=1}^{n}y_i\,,s_{xx}=\sum_{i=1}^{n}x_i^2\,,s_{xy}=\sum_{i=1}^{n}x_iy_i$$

③ 解正规方程组

$$\begin{cases}na+s_xb=s_y\\s_xa+s_{xx}b=s_{xy}\end{cases}$$

④ 输出a和b

$$a=\frac{s_{xx}s_y-s_xs_{xy}}{ns_{xx}-s_x^2},b=\frac{ns_{xy}-s_xs_y}{ns_{xx}-s_x^2}$$

3.2 多变量数据拟合

- 在实际问题中，很多问题反映的不是一个量与一个量的关系，而是一个量与若干个量的关系.
  - 一个量由若干个量确定
  - 其中：若干个量通常称为自变量，由这些自变量确定的量通常称为因变量
  - 记自变量为$x_1, x_2, \cdots, x_k$，因变量为y
  - n次实验或测量得到n组数据

次数	x_1	x_2	$\cdots$	x_k	$y = f(x_1, x_2, \cdots, x_k)$
1	x_{11}	x_{12}	$\cdots$	x_{1k}	y_1
2	x_{21}	x_{22}	$\cdots$	x_{2k}	y_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nk}	y_n

3.2 多变量数据拟合

多变量线性拟合

$$F(x_1, x_2, \cdots, x_k) = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_k x_k$$

- 使用最小二乘法确定待定参数$a_0, a_1, \cdots, a_k$
  - 偏差平方和

$$\varphi(a_0, a_1, \cdots, a_k) = \sum_{m=1}^{n} (y_m - a_0 - a_1 x_{m1} - a_2 x_{m2} - \cdots - a_k x_{mk})^2$$

  - 根据多元函数求极小值方法，对$\varphi(a_0, a_1, \cdots, a_k)$分别求关于$a_0, a_1, \cdots, a_k$的偏导数并令其等于0

$$\frac{\partial \varphi}{\partial a_i} = 0 \ (i = 0,1, \cdots, k)$$

  - 解方程组得到$a_0, a_1, \cdots, a_k$

3.2 多变量数据拟合

多变量线性拟合

$$F(x_1, x_2, \cdots, x_k) = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_k x_k$$

$$\begin{cases} n a_0 + a_1 \sum_{m=1}^{n} x_{m1} + a_2 \sum_{m=1}^{n} x_{m2} + \cdots + a_k \sum_{m=1}^{n} x_{mk} = \sum_{m=1}^{n} y_m \\ a_0 \sum_{m=1}^{n} x_{m1} + a_1 \sum_{m=1}^{n} x_{m1} x_{m1} + a_2 \sum_{m=1}^{n} x_{m2} x_{m1} + \cdots + a_k \sum_{m=1}^{n} x_{mk} x_{m1} = \sum_{m=1}^{n} y_m x_{m1} \\ a_0 \sum_{m=1}^{n} x_{m2} + a_1 \sum_{m=1}^{n} x_{m1} x_{m2} + a_2 \sum_{m=1}^{n} x_{m2} x_{m2} + \cdots + a_k \sum_{m=1}^{n} x_{mk} x_{m2} = \sum_{m=1}^{n} y_m x_{m2} \\ a_0 \sum_{m=1}^{n} x_{mk} + a_1 \sum_{m=1}^{n} x_{m1} x_{mk} + a_2 \sum_{m=1}^{n} x_{m2} x_{mk} + \cdots + a_k \sum_{m=1}^{n} x_{mk} x_{mk} = \sum_{m=1}^{n} y_m x_{mk} \end{cases}$$

3.2 多变量数据拟合

例：已知一组测量数据，求其线性拟合函数.

测量次数	x_1	x_2	$y = f(x_1, x_2)$
1	1	1	7
2	1	2	9
3	2	1	10
4	2	2	11
5	2	3	12

$$F(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2$$

3.3 非线性数据线性化

- 原始数据之间并不呈现线性关系，无法直接应用最小二乘线性拟合.

① 直接应用最小二乘思想可能得到非线性方程组，不便于求解.

② 可以尝试将原始数据作一定的变换，使经过变换后的数据呈现线性关系.

3.3 非线性数据线性化

例：钢包容量与使用次数之间关系的测试数据

i	次数	容量	i	次数	容量	i	次数	容量
1	2	6.42	6	7	10.00	11	12	10.60
2	3	8.20	7	8	9.93	12	13	10.80
3	4	9.58	8	9	9.99	13	14	10.60
4	5	9.50	9	10	10.49	14	15	10.90
5	6	9.70	10	11	10.59	15	16	10.76

$$\frac{1}{y} = a + b\frac{1}{x} \text{ (双曲线)}$$

$$X = \frac{1}{x}, Y = \frac{1}{y}$$

$$Y = a + bX \text{ (直线)}$$

3.3 非线性数据线性化

例：已知一组数据如下，求一个经验函数，形如 $y = ae^{bx}(a, b$为常数$)$，使之与数据相拟合.

x	0	1	2	3	4
$y = f(x)$	1.5	2.5	3.5	5	7.5

$$y = ae^{bx}$$

$$\ln y = \ln a + bx$$

$$Y = \ln y, A = \ln a, B = b, X = x$$

x	0	1	2	3	4
$Y = lny_i$	0.405465	0.916291	1.252763	1.609438	2.014903

$$Y = A + BX$$

3.3 非线性数据线性化

- 采用非线性数据线性化方法拟合多项式
  - 设有两个量z和y基本满足m次多项式，实验测量数据如下表：

z	z_1	z_2	$\cdots$	z_n
$y=f(z)$	y_1	y_2	$\cdots$	y_n

$$y = a_0 + a_1 z + a_2 z^2 + \cdots + a_m z^m$$

$$x_1 = z, x_2 = z^2, \cdots, x_m = z^m$$

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_m x_m$$

$$= F(x_1, x_2, \cdots, x_m)$$

3.4 正交多项式拟合

- 多项式拟合
  - 先把多项式拟合函数变成多变量拟合函数
  - 利用多变量拟合法求出多项式系数a_i
    - 需要求解一个正规方程组
    - 当多项式次数较高时，正规方程组可能会病态

定义：如果方程组$\boldsymbol{Ax = b}$中的系数矩阵$\boldsymbol{A}$和常数项$\boldsymbol{b}$有微小变化，就会引起方程组的解很大变化，则方程组$\boldsymbol{Ax = b}$称为病态方程组

3.4 正交多项式拟合

- m次多项式函数：

$$y^* = \Psi_m(x) = a_0 p_0(x) + a_1 p_1(x) + \cdots + a_m p_m(x)$$

$$= \sum_{k=0}^{m} a_k p_k(x)$$

其中a_i为待定参数，$p_k(x)(k = 0,1,\cdots,m)$是k次多项式.

- 当$x = x_i$时，产生的偏差为：

$$\delta_i = y_i - y_i^* = y_i - \sum_{k=0}^{m} a_k p_k(x_i)$$

  - 最小二乘法：偏差平方和最小

3.4 正交多项式拟合

- 由于从实验或测量中得到的不同数据精度不同，为了反映这种不同，通常在每一个δ_i前面乘上一个表示数据精度的权数α_i
- 使加权偏差平方和最小

$$\sum_{i=1}^{n}(\alpha_i\delta_i)^2=\sum_{i=1}^{n}\alpha_i^2\delta_i^2=\sum_{i=1}^{n}\omega_i\delta_i^2$$

其中$\omega_i=\alpha_i^2$称为权因子.

$$\varphi(a_0,a_1,\cdots,a_m)$$

$$\stackrel{\text{def}}{=}\sum_{i=1}^{n}\omega_i\delta_i^2=\sum_{i=1}^{n}\omega_i\left[y_i-\sum_{k=0}^{m}a_kp_k(x_i)\right]^2$$

3.4 正交多项式拟合

- 选择a_i使加权偏差平方和$\sum_{i=1}^{n}\omega_i\delta_i^2$最小的问题转化为求函数$\varphi(a_0,a_1,\cdots,a_m)$极小值的问题
  - 对函数$\varphi(a_0,a_1,\cdots,a_m)$分别求关于$a_0,a_1,\cdots,a_m$的导数，令其等于0，联立得到方程组：

$$\frac{\partial\varphi}{\partial\alpha_j}=-2\sum_{i=1}^{n}\omega_i\left[y_i-\sum_{k=0}^{m}a_kp_k(x_i)\right]p_j(x_i)=0$$

其中$(j=0,1,\cdots,m)$

$$\sum_{i=1}^{n}\omega_i\left[\sum_{k=0}^{m}a_kp_k(x_i)p_j(x_i)\right]=\sum_{i=1}^{n}\omega_iy_ip_j(x_i)$$

3.4 正交多项式拟合

- 交换求和顺序得到

$$\sum_{k=0}^{m} a_k \left[\sum_{i=1}^{n} \omega_i p_k(x_i) p_j(x_i) \right] = \sum_{i=1}^{n} \omega_i y_i p_j(x_i)$$

令$c_{jk} = \sum_{i=1}^{n} \omega_i p_k(x_i) p_j(x_i)$，$b_j = \sum_{i=1}^{n} \omega_i y_i p_j(x_i)$

- 正规方程组

$$\sum_{k=0}^{m} c_{jk} a_k = b_j$$

定义：对数据x_i和加权因子ω_i的正交多项式簇.

$$c_{jk} = \sum_{i=1}^{n} \omega_i p_k(x_i) p_j(x_i) = 0 \quad (j \neq k)$$

$$c_{jj} = \sum_{i=1}^{n} \omega_i p_j(x_i)^2 > 0 \; (j, k = 0,1,\cdots,m)$$

3.4 正交多项式拟合

$$\sum_{k=0}^{m} c_{jk}a_k = b_j,\ c_{kk}a_k = b_k,\ a_k = \frac{b_k}{c_{kk}}$$

例：正交多项式簇

- 假设给定一组$n+1$个等距节点$\xi_i(i=0,1,\cdots,n)$，间隔为h，选取加权因子$\omega_i = 1$
- 引入变换 $x = \frac{\xi-\xi_0}{h}$，则 ξ_i变为$x_i = i$，x_i是$n+1$个整数等距节点
- 构造多项式

$$p_{m,n}(x) = \sum_{k=0}^{m}(-1)^k \binom{m}{k}\binom{m+k}{k}\frac{x^{(k)}}{n^{(k)}}$$

其中$x^{(k)} = x(x-1)\cdots(x-k+1)$且$x^{(0)} = 1$


Thank You!