统计推断在数模转换系统中的应用

组号　　姓名：骆雨 学号：5130309400，姓名：刘睿 学号：5130309381

摘要：为某产品寻求校准工序的优化方案。在问题的求解中需要用到MatLab软件通过编程来得到通过测定点的曲线方程，并且在软件中把所得到的结果画出来，然后计算测定点的测定成本以及误差损失的成本，然后改变测定点的组合把最终的成本计算出来进行比较，选择最佳的测定点的组合方案，其中或许会用到遗传算法以及退火算法。

关键词：MatLab，遗传算法，三次多项式拟合，三次样条插值法拟合

1 引言

1.1问题的提出

首先，通过三次课堂讲座老师对于问题的介绍以及求解的基本思路的讲解，我们认识到了这次的问题是通过对某一传感器的一部分数据点（离散点）上特征的测量来估计这一传感器的连续特性（即为某产品寻求校准工序的优化方案）。但是已知这个传感器的特征是分为三段，与对应的输入值分别成较好的线性关系，即三部分分为首段，中段以及尾段三部分。在问题的求解中需要用到MatLab软件通过编程来得到通过测定点的测定的三段曲线方程，并且在软件中把所得到的结果画出来，然后计算测定点的测定成本以及误差损失的成本，然后改变测定点的组合把最终的成本计算出来进行比较，选择最佳的测定点的组合方案，其中或许会用到遗传算法。

2 曲线拟合

然后，通过小组讨论，首先我们认为可以在首段、中段以及尾段上分别取若干个点，通过回归分析可以得到三段的回归曲线方程，并且对三段方程之间求解可以得到三段曲线之间的转折点，从而可以使用软件来得到产品的特征曲线。由于已知是线性的，因此仅使用一次方程作为最终的表达式。其次，在得到三段曲线方程以后就可以分别计算测定点的测定成本以及通过测定点的估算点的误差损失成本。再次，通过进行总成本的比较来确定最适合的测定点组合，从而得到最优解。

由最小二乘法线性回归公式:

$$Y=b+ax;$$

其中：

$$b=(\sum(x_i * y_i)-(1/n)*\sum x_i *\sum y_i)/(\sum(x_i\text{^}2)-(1/n)*(\sum x_i)\text{^}2)$$

$$a=\bar{y}-b*\bar{x}$$

$$\bar{y}=(\sum y_i)/n \quad \text{并且} \quad \bar{x}=(\sum x_i)/n$$

在画图时使用到函数plot（a1,b1,a2,b2…）其中以a为横坐标，以b为纵坐标进行画图，其中b

与a满足一定的函数关系。
在产生随机数时使用randi函数。Randi(a,b,c)可以产生b行c列的从1~a的随机数。

3 遗传算法

3.1 遗传算法的概述[1]

遗传算法 (Genetic Algorithm)是一种仿生算法,模拟了生物学中基因遗传的模式。它隶属于启发式 heuristic 搜索算法。

遗传算法中有三种重要的操作作为其算法核心:

a. 选择-复制(selection-reproduction)

b. 交叉(crossover，亦称交换、交配或杂交)

c. 变异(mutation，亦称突变)

a.选择-复制

对于一个规模为 N 的种群 S,按每个染色体 $x_i \in S$ 的选择概率 $P(x_i)$ 所决定的选中机会，分 N 次从 S 中随机选定 N 个染色体，并进行复制。

$$P(x_i)=\frac{f(x_i)}{\sum_{j=1}^{N} f(x_j)}$$

b. 交叉 就是互换两个染色体某些位上的基因。

交叉率(crossover rate)就是参加交叉运算的染色体个数占全体染色体总数的比例，记为 Pc,取值范围一般为 0.4～0.99。

c. 变异 就是改变染色体某个(些)位上的基因。

变异率(mutation rate)是指发生变异的基因位数所占全体染色体的基因总位数的比例，记为 Pm，取值范围一般为 0.0001～0.1。

通过对遗传算法的初步了解，我们知道：对于函数优化的实数编码与二进制编码各有优劣：二进制编码有稳定性高、种群多样性大的优点，而实数编码较容易理解，但是有着容易过早收敛，陷入局部最优的巨大缺陷，因此我们使用时应该使用二进制编码。
在遗传算法开始之前我们要对种群初始化，设种群大小为pop_size=30（在以后的工作中可以逐步缩小），每个染色体或个体的长度为chromo_size。
在选择操作中，应首先计算各个体的适应度，分布，在本问题中，认为其成本越小的，适应度越高。考虑其概率应为：abs(a/b-1)。其中abs是对一个数取其绝对值，a为一个个体的总成本，而b为整个种群的成本之和。这样就可以使用转盘方法选择适应度较好的个体，去掉适应度较差的个体。在繁衍下一代的过程中，对每一个没有淘汰的个体进行判断是否进行遗传与变异，然后随机生成一个实数0<=r<=1, 如果r<cross_rate， 0<cross_rate<1为交叉概率，则对这两个个体进行交叉，否则则不进行。如果需要进行交叉，再随机选择交叉位置，如果等于0或者1，将不进行交叉。否则将交叉位置以后的二进制串进行对换（包括交叉位置）。并在子代以后继续判断，直到找出最优解。
基于以上对遗传算法的初步分析，然后进行编码，解决问题。
通过使用遗传算法等其他算法完成组合测试点的选取
通过对matlab的进一步了解，了解对大数据的统一处理。
有关此次统计推断中所使用的是遗传算法，分别用三次函数（多项式）以及三次样条插值法

进行拟合统计。

3.2 三次样条插值法

将由 7 个点组成的曲线用 6 段 3 次曲线来近似表示。

首先选出取样点$S_1, S_2 \cdots S_7$；

对中间的点，根据左右最邻近的 4 个点进行三次函数插值作为中间点之间的曲线方程；

对于两边的点，则用靠近两端的三个点用二次函数对其插值。

以上过程在 matlab 中可用函数 interp1 实现。

在最初使用时，种群个体数目为 32-40 个（其大小与取点个数、方法有关），考虑到取点个数增加以后，取点成本会大大增加而拟合效果可能不会有太大变化（即使取了绝大多数点，拟合曲线与实际曲线不会完全重合，存在误差成本），因此，其中每个个体基因数为 5-7 不等。在调试代码阶段，为了测试代码的正确性及有效性，将遗传代数设为 5，便于观察。之后增加遗传代数，使拟合函数充分收敛，获得一个较好的取点结果。

在前期使用遗传算法时，种群数量不大，在每次计算了所得到的成本以后，取其成本的倒数为个体适应度，将最不适应环境的一半淘汰（使用遗传算法库函数 select），然后将剩余的一半作为父代繁衍后代，设置后代数量与父代数量一致，最后遗传以后的种群就是前一代种群中最优的一半以及其所繁殖产生的子代的总体。

在与老师进行交流之后，发现上述处理方法会使程序早熟，即过早收敛，其收敛的最终结果与初始种群取点的优劣有着较为密切的关系，如果初始种群取点方法较优，则所得到的成本也相对较低，反之，则较高。出现这样结果的原因为：1. 初始种群较小，使得每个基因数量有限，良好的基因不能够充分表现。2. 由于将较不适合环境的一半个体去掉，其体内的优良基因不能够遗传给子代，导致可以繁殖的个体之间基因相似，从而导致近亲繁殖，使算法过早收敛。

在改进的遗传算法中，将种群中个体数扩充为 200，在原有若干项不变的情况下，多取的个体使用以下方法：当基因数为 n 时，将所有点数（一共 51）基本均分为 n 个区间，然后在每个区间内随机取点（其中用到随机取点函数 randi(n)，输出为从 1 到 n 的随机整数）。对于遗传与变异的过程，使用多次重组变异，使种群产生较大的变化，便于取出优良的基因，达到较好效果。再有就是减少被淘汰的个体数目，即在筛选的过程中，留下的个体数占原有个体数的 90%，而不是 50%，这样可以在一定程度上避免算法过早收敛。

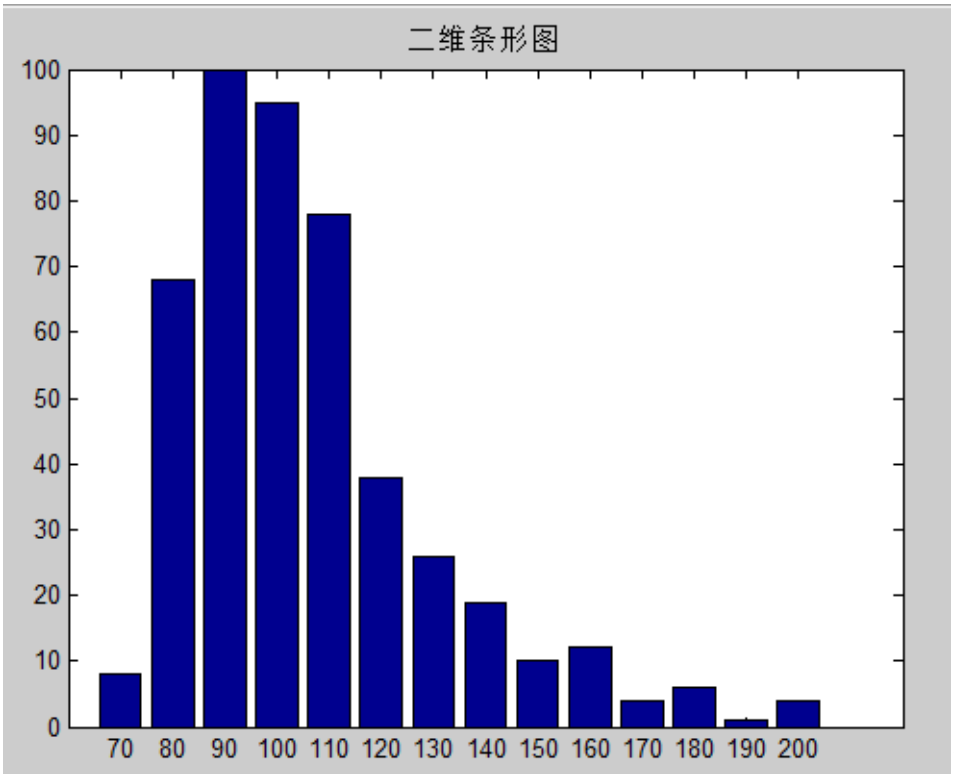
本次使用遗传算法对数据进行处理

所设种群中个体数目为 200 个，遗传代数为 30 代，基因数（取点数目）分别为 5，6，7 所得到的最优解及其成本为：

当基因数为 5 时，得到的最优解为[4 15 26 37 48]，平均成本为 111.5，最佳个体成本为 72.0 当基因数为 6 时，所得到的最优解为[4 13 22 31 40 49]，平均成本为 121.61，最优个体成本为 84.5 当基因数为 7 时所得到的最优解为[3 11 18 26 35 40 48]，平均成本为 133.6，最佳个体成本为 96。当基因数为 8 时所得到的最优解为[4 10 16 24 30 36 40 48]，平均成本为 146.6，最佳成本为 108。可以想象，当取点数继续增加时，成本也会逐渐增加。

因此可以认为： 采用三次函数的拟合方法，通过遗传与变异所得的最佳取点为 5 个点，所得到的取点为[4 15 26 37 48]，所获得的最佳平均成本为 111.5

在这种取点情况下，对 469 组数据计算成本的分布情况如图（1）所示



图（1）

分析图（1）不难看出，成本在 80 以下的数目较少，大部分成本均在 80-120 之间，其中成本在 90-100 之间的数目最多。

在使用三次样条插值拟合时，对于分别取 6，7 个点：

当取 6 个点时，得到的最优解为[4　13　21　31　39　49]，所得到的成本为 98.85

当取 7 个点时，得到的最优解为[5　14　19　25　32　39　49]，所得到的成本为 114.45

因此可以认为在使用三次样条插值法拟合时所得的最优解为取 6 个点，取点情况为[4　13　21　31　39　49]，可以得到较小的平均成本为 98.85

分析认为，不论是三次样条插值法还是三次多项式拟合，当取点数较多时，尽管有可能减少误差成本，但是测定成本会成线性增长，大大增大最终平均成本。当取点数较少时，尽管测定成本会减少，但是由于拟合的不准确度提高，会使误差成本大大增加，也不利于产生较好的一个取点方法，因此认为，在取测定点是以 5、6、7 个点为宜，然后计算最优的取点方法。

本次设计中所采用的遗传算法函数库中的函数有：`select` 用于选择更为适应环境的个体，`xovmp recombine` 用于多点交叉重组

本次设计中所采用的拟合函数为 polyfit(a,b,c)，其中参数 a 表示横坐标的值，b 表示纵坐标的值，参数 c 为整型数，表示所拟合曲线的最高次数，即 c 次多项式。输出为 c 次多项式系数。

三次样条插值法

由图形直观观测不难发现，其数据点走势近似可以划分为三段直线段，因此可以分为三段分别用一次函数拟合，即三次样条插值法。

当取点为 5 个点时，首段使用前两个点拟合，中段使用第 3，4 个点拟合，尾段使用第 4，5 个点拟合。

当取点为 6 个点时，每两个点分为一组，分别使用 polyfit(a,b,1)进行一次函数拟合

当取点为 7 个点时，首段使用前两个点进行拟合，中间一段为三个点进行拟合，最后一段使用最后的 2 个点进行拟合。

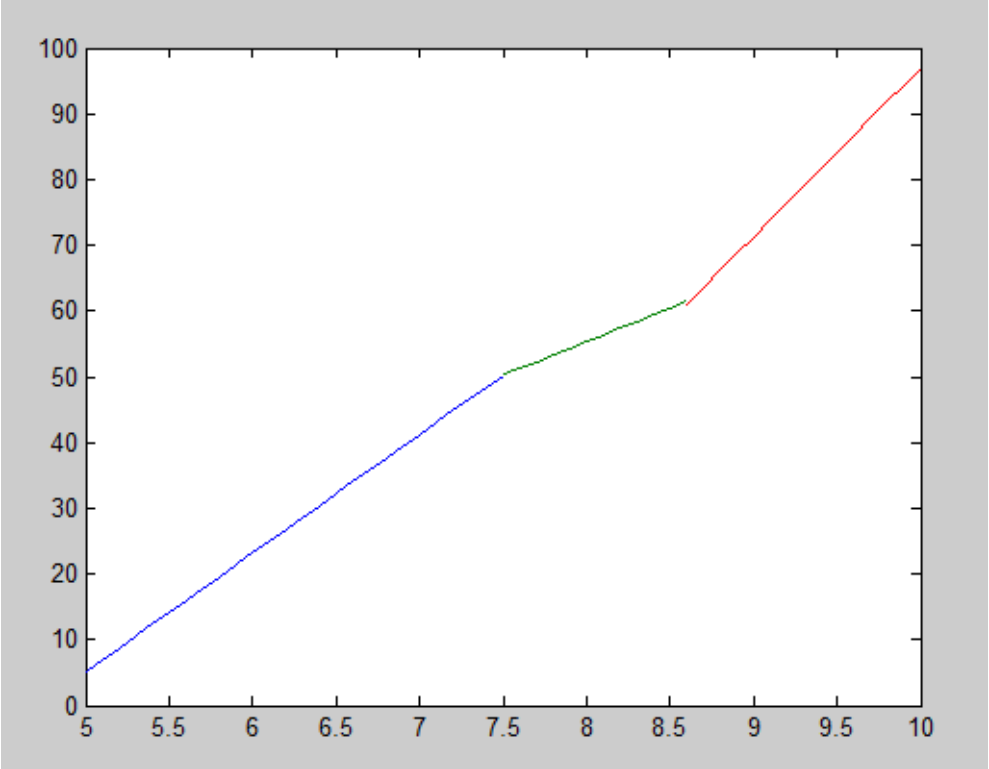
继续观测图形，其走势是逐渐上升的，首尾两段增加幅度较大，而中间一段的变化幅度较小，考虑到它与三次函数的走势相近，因此使用三次多项式进行拟合。

在使用三次样条插值法分析时，当获得三段直线的相关系数以后，（设其中相邻两段的函数为 $y_1 = a_1 \times x + b_1$， $y_2 = a_2 \times x + b_2$）则其交点横坐标为

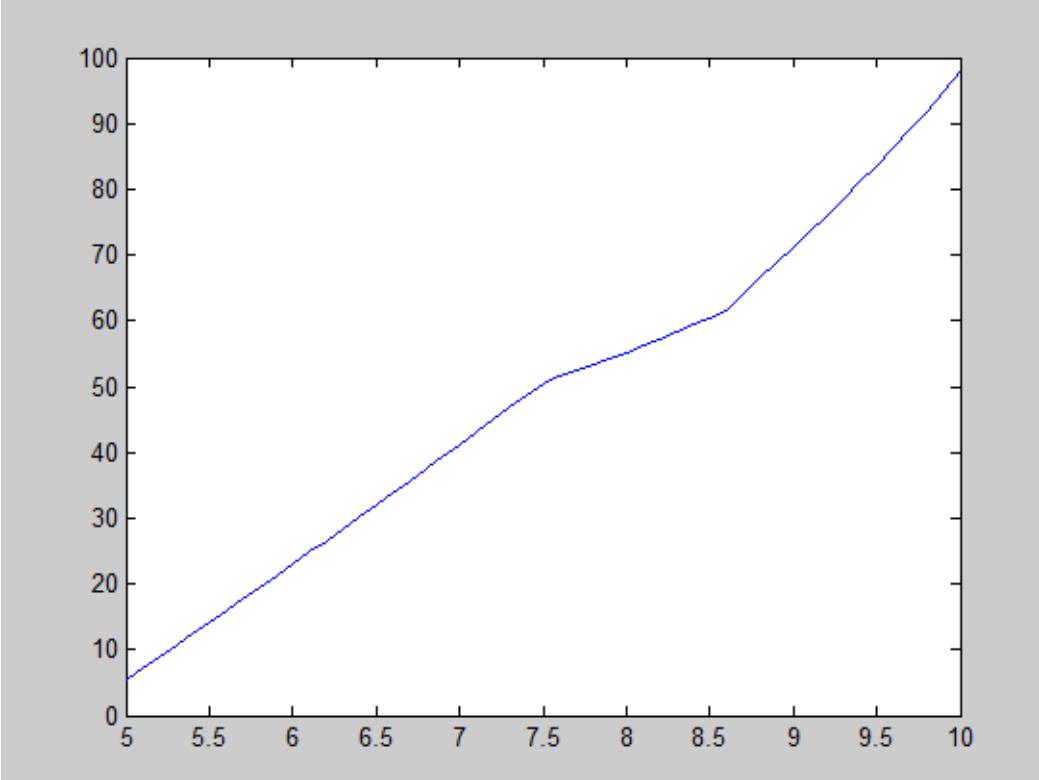
$$x = \frac{b_2 - b_1}{a_1 - a_2}$$

在计算误差成本时，当横坐标值小于交点横坐标时，则使用前一段函数计算所得到的值，否则使用后一段一次函数计算所得到的值。在将每一个点对应的估计值计算出以后，与标准值比较计算误差值，从而计算成本。

以第一组数据为例所得到的拟合曲线（对于一个特例，取 8 个点：图（2）为：



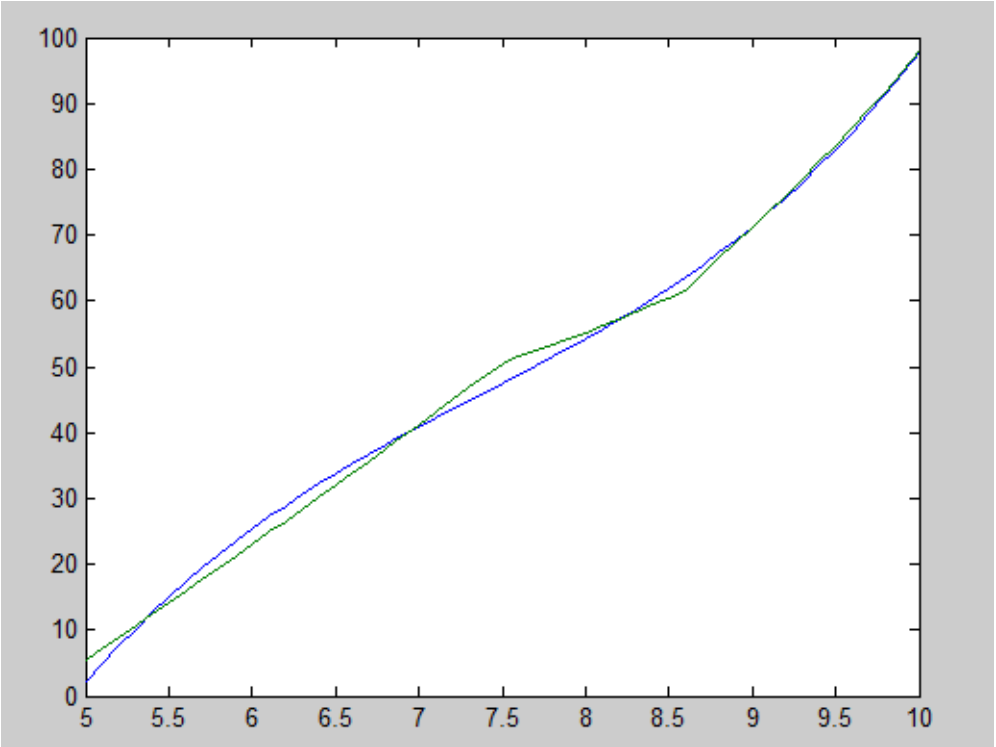
图（2）：对特例取 8 个点准确拟合图像



图（3）：原有数据所得图像（连点为线）

与图（2）为比较发现，拟合较为准确

当使用三次多项式拟合时，与原图比较发现，还是存在较大差距（见图（4），其中蓝线为拟合曲线，红线为原有数据的曲线）



图（4）：三次函数拟合与原图比较图

而在使用三次多项式拟合分析时，其系数之间可以得到。由于不存在分段，可以直接使用一个循环计算成本。

4. 拓展探究

不论是以上所提到的三次多项式拟合还是三次样条插值拟合，均不能够较为智能地根据每一个图形的大致走向来进行个性化的拟合，因此提出以下改进方法（主要针对三次样条插值）：首先在首段、尾段各取一个点 a,b，再在所有数据的中间取一个点 c，在 a,c 中间取一点 d，通过观测这 4 个点，判断三段直线段的大致分布，通过取不同的点达到计算出直线段方程的目的。

由于时间有限，未对本次课程设计提出较为详细的统计方面的结论，只是粗略计算了在一种情况（即用三次函数拟合，取 5 个点）下的每个样本的成本。没有计算其方差等有关数据。计算方差的公式如下：

$$Var(a)=\frac{\sum_{i=1}^{469}(a_i-\bar{a})^2}{469}$$

其中 a_i 表示每一个样本的成本，$\bar{a}$ 表示平均成本。

由于对 Matlab 本身以及遗传算法的了解度不够，在编码过程中主要使用 C 风格的编码，使得程序运行较为复杂，没有简化代码。

5 代码附录

```
%对第一组数据处理（没有遗传算法）
a=[5.00 5.30 5.40 5.50 5.60 5.70 5.80 5.90 6.00 6.10 6.20 6.30 6.40 6.50
6.60 6.70 6.80 6.90 7.00 7.10 7.20 7.30 7.40 7.50 7.60 7.70 7.80 7.90 8.00
8.10 8.20 8.30 8.40 8.50 8.60 8.70 8.80 8.90 9.00 9.10 9.20 9.30 9.40 9.50
9.60 9.70 9.80 9.90 10.00 ];
b=[5.33 10.74 12.45 14.23 16.05 17.77 19.58 21.31 23.07 24.87 26.59 28.34
30.18 31.92 33.78 35.64 37.50 39.31 41.14 42.99 44.91 46.79 48.63 50.53
51.47 52.41 53.27 54.12 55.12 56.23 57.27 58.34 59.41 60.51 61.67 64.04
66.38 68.72 71.14 73.62 76.00 78.46 81.00 83.59 86.43 88.97 91.76 94.92
98.05 ];
k=[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0];
h=[0 0 0];%kÎªÐ±ÂÊ hÎªËù±ê¼ÇµÄ×ªÕÛµã
aa=[5.00 5.10 5.20 5.30 5.40 5.50 5.60 5.70 5.80 5.90 6.00 6.10 6.20 6.30
6.40 6.50 6.60 6.70 6.80 6.90 7.00 7.10 7.20 7.30 7.40 7.50 7.60 7.70 7.80
7.90 8.00 8.10 8.20 8.30 8.40 8.50 8.60 8.70 8.80 8.90 9.00 9.10 9.20 9.30
```

```
9.40 9.50 9.60 9.70 9.80 9.90 10.00 ];
bb=[5.33 7.03 8.92 10.74 12.45 14.23 16.05 17.77 19.58 21.31 23.07 24.87
26.59 28.34 30.18 31.92 33.78 35.64 37.50 39.31 41.14 42.99 44.91 46.79
48.63 50.53 51.47 52.41 53.27 54.12 55.12 56.23 57.27 58.34 59.41 60.51
61.67 64.04 66.38 68.72 71.14 73.62 76.00 78.46 81.00 83.59 86.43 88.97
91.76 94.92 98.05 ];
numa=nnz(a);
constcost=30;
numaa=nnz(aa);
numb=nnz(b);
for i=1:(numa-1)
    k(i)=(a(i+1)-a(i))\(b(i+1)-b(i));
end
i=1;j=1;
for i=1:(numa-2)
    if (abs(k(i+1)-k(i))>5)
        h(j)=i;j=j+1; %have at least 2 numbers in h
    end
end
rr=1:(h(1)+1); %µÚÒ»¶ÎÇúÏß
rs=(h(1)+1):(h(2)+1);  %µÚ¶þ¶ÎÇúÏß
rt=(h(2)+1):numa;  %µÚÈý¶ÎÇúÏß
r1=rr;    rxy1=rr; rr1=rr;
r2=rs;    rxy2=rs; rs2=rs; %rr1,rs2 rt3±íÊ¾x*x
r3=rt;    rxy3=rt; rt3=rt; %
for i=1:(h(1)+1)
    rr(i)=a(i);r1(i)=b(i); %rr for x, r1 for y
    rr1(i)=a(i)*a(i);rxy1(i)=b(i)*a(i); %rr1 for x^2, rxy1 for xy
end
for i=(h(1)+1):(h(2)+1)
    rs(i-h(1))=a(i);r2(i-h(1))=b(i);
    rxy2(i-h(1))=b(i)*a(i); rs2(i-h(1))=a(i)*a(i);
end
for i=(h(2)+1):numa
    rt(i-h(2))=a(i);r3(i-h(2))=b(i);
    rxy3(i-h(2))=a(i)*b(i); rt3(i-h(2))=a(i)*a(i);
end
xx=sum(rr)/(1+h(1));yx=sum(r1)/(1+h(1)); %xx for the first average of x
lxxx=sum(rr1)-(sum(rr))*(sum(rr))/(1+h(1));
lxxy=sum(rxy1)-sum(rr)*sum(r1)/(1+h(1));

xy=sum(rs)/(h(2)-h(1)+1);yy=sum(r2)/(h(2)-h(1)+1);%the second
lyxx=sum(rs2)-(sum(rs)*sum(rs))/(h(2)-h(1)+1);
lyxy=sum(rxy2)-sum(rs)*sum(r2)/(h(2)-h(1)+1);
```

```
xz=sum(rt)/(numa-h(2));yz=sum(r3)/(numa-h(2));
lzxx=sum(rt3)-(sum(rt)*sum(rt))/(numa-h(2));
lzxy=sum(rxy3)-sum(rt)*sum(r3)/(numa-h(2));

k1=lxxy/lxxx;d1=yx-k1*xx;
k2=lyxy/lyxx;d2=yy-k2*xy;
k3=lzxy/lzxx;d3=yz-k3*xz;
x1=rr;y1=k1*x1+d1;
x2=rs;y2=k2*x2+d2;
x3=rt;y3=k3*x3+d3;
plot(x1,y1,x2,y2,x3,y3);
cost=0;cha=0;ccost=0;
for i=1:numaa
    if aa[i]<=a[h(1)+1]
        if abs(k1*aa[i]+d1-bb[i])<=1
            ccost=0;

        elseif abs(k1*aa[i]+d1-bb[i])<=2 && abs(k1*aa[i]+d1-bb[i])>1
            ccost=1;
        end
        if abs(k1*aa[i]+d1-bb[i])<=3 && abs(k1*aa[i]+d1-bb[i])>2
            ccost=2;
        end
        if abs(k1*aa[i]+d1-bb[i])<=4 && abs(k1*aa[i]+d1-bb[i])>3
            ccost=3;
        end
    end
    if aa[i]<=a[h(2)+1] && aa[i]>a[h(1)+1]
        if abs(k2*aa[i]+d2-bb[i])<=1
            ccost=0;
        elseif abs(k2*aa[i]+d2-bb[i])<=2 && abs(k2*aa[i]+d2-bb[i])>1
            ccost=1;
            elseif            abs(k2*aa[i]+d2-bb[i])<=3            &&
abs(k2*aa[i]+d2-bb[i])>2
                    ccost=2;
                elseif          abs(k2*aa[i]+d2-bb[i])<=4          &&
abs(k2*aa[i]+d2-bb[i])>3
                        ccost=3;
        end
    end
    if aa[i]>a[h(2)+1]
        if abs(k3*aa[i]+d3-bb[i])<=1
            ccost=0;
```

```
        elseif abs(k3*aa[i]+d3-bb[i])<=2 && abs(k3*aa[i]+d3-bb[i])>1
            ccost=1;
            elseif            abs(k3*aa[i]+d3-bb[i])<=3            &&
abs(k3*aa[i]+d3-bb[i])>2
                    ccost=2;
                elseif          abs(k3*aa[i]+d2-bb[i])<=4          &&
abs(k3*aa[i]+d2-bb[i])>3
                        ccost=3;
        end
    end
    cost=cost+ccost;
end
cost=cost+numa*constcost; &所得成本
```

6 结论

正文内容

7 参考文献

[1] 上海交大电子工程系. 统计推断在数模转换系统中的应用课程讲义