统计推断在 ADC 中的应用

组号：19　　张家杰 5130309445　吕东亮 5130109046

摘要：本文以某型产品内部的检测模块为研究对象，在确保测量精度的前提下，运用统计推断方法，以 MATLAB 为工具，通过对原始数据库中 469 组密集选点的实验数据进行分析，来寻找校准工序的优化方案，在确保精度前提下，尽量降低测定成本。考虑到这是一个组合优化问题，我们希望通过学习数种统计方法，比较取较优方法来获得校准工序的优化方案，并且通过测试与分析，评价方案的有效性。

关键词：统计推断，组合优化，曲线拟合，插值，MATLAB

1 引言

在原始数据库的数据条件下，研究被检测物理量与所得示数的关系。标准化测量无法精确到各点，经过抽样研究后得到如下特点：

（1）有确定的对应关系

（2）局部非线性或局部线性

（3）个体之间存在明显差异

所以需要通过几个特殊的测定点来有效拟合出产品的示数的特定曲线（本实验中为所测物理量与检测模块的特性曲线），并由此确定校准工序的优化方案。本实验中，我们通过原始数据库中 469 组密集选点的实验数据，讨论如何在误差限定范围内完成标定，需要讨论测定点的个数，测定点的选取，关系曲线的表达式确定方法。由于数据组数叫为庞大，我们不可能使用暴力穷举等传统方法来得出结论。考虑到拟合效果与运行时间，我们选定多项式与插值法先进行拟合，然后通过退火法拟合，最终根据结果优化得到最终的标定方法。

2 拟合或插值法选取探究

根据所给的469组Y-X数据以及其曲线形状，我们发现大多数数据有着相类似的曲线形状。考虑到对每组数据都是用同样一个最优解，故我们选择使用一个能够满足所有数据组的方法来得到曲线。已经给出的成本计算规则如下所示，据此研究每一组数据，我们将得到的曲线在每一个数据点与其原数据值比较并得出相应的测定成本，取所有组成本的平均值，为本组解的成本（测定成本应当越小越好）。

$$s_{i,j}=\begin{cases}0 & \text{if } \left|\hat{y}_{i,j}-y_{i,j}\right|\le 0.5\\ 0.5 & \text{if } 0.5<\left|\hat{y}_{i,j}-y_{i,j}\right|\le 1\\ 1.5 & \text{if } 1<\left|\hat{y}_{i,j}-y_{i,j}\right|\le 2\\ 6 & \text{if } 2<\left|\hat{y}_{i,j}-y_{i,j}\right|\le 3\\ 12 & \text{if } 3<\left|\hat{y}_{i,j}-y_{i,j}\right|\le 5\\ 25 & \text{if } \left|\hat{y}_{i,j}-y_{i,j}\right|>5\end{cases}$$　（1）式，单点成本测定式

$$S_i=\sum_{j=1}^{51}s_{i,j}+q\cdot n_i$$　（2）式，单个样本成本测定式，式中q为单点测定成本，q=12

$$C=\frac{1}{M}\sum_{i=1}^{M}S_i$$ （3）式，总体成本测定式

2.1 多项式拟合

2.1.1 概述

在广泛观察了大量的实验数据之后，我们总结出数据实验曲线在中部的数据变化值偏大，而在两端的数据具有一个较为良好的线性关系，如图1所示，第1组数据图像。

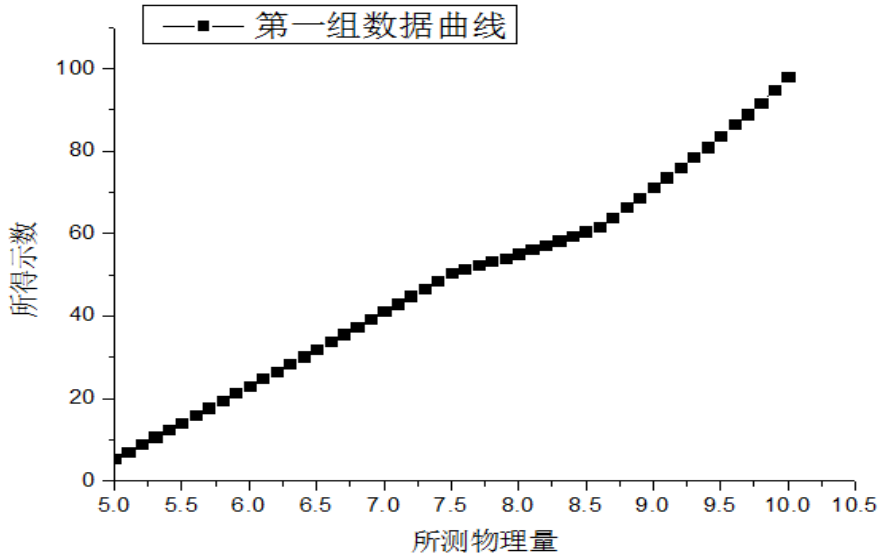

图 1 第一组数据曲线

因此我们认为可以通过多项式的性征来拟合出数据曲线。由于曲线两端的线性较好，中段曲线的性状较为不明显，考虑在中部多选取点，而在两边选取较少的点，根据选取的样本数据的曲线以及成本计算公式，考虑选取不大于七个点来定标，而七个点最多只能确定六次多项式，我们对二次至六次的各阶多项式进行了尝试。

我们对各阶多项式拟合都编写了成本计算函数，其输入值为一个7元素数组，其元素为由小到大排列的范围在2至51的7个数，代表所选取的7个点的最优解。通过MATLAB矩阵算法提取每一组的数据进行成本计算并平均，其输出为所的成本值。

2.1.2 各阶多项式拟合结果比较

利用随机抽取的50组数据样点组合对二次至六次曲线的相应拟合结果进行分析，得到如下表格1中的数据。

表格1 多项式拟合所得结果

	平均运行时间/s	拟合组数	计算所得成本
二次多项式	0.25	51	641.1
三次多项式	0.34	51	146.7
四次多项式	0.55	51	140.7
五次多项式	0.56	51	143.5
六次多项式	0.72	51	184.6

通过对表格中数据进行分析可知，当拟合阶数为三阶、四阶、五阶时，其平均成本处于130左右，说明其拟合效果较好，能较为准确的反映出应有的数据走势，并且成本控制的较好。

而二阶、六阶的曲线，其拟合所得成本明显偏高，故在实验中应当选取三、四、五次的多项式，进行数据拟合。

同时可从表中观察到，对于不同阶次的多项式而言，其运行时间也有较大的差异。对二阶、三阶的拟合，其用时较短，而四五六阶时间均在0.6s以上，其原因可能为是多项式次数升高导致的运算复杂度增加，从而增加了拟合时间。

2.1.3 结果分析

根据已有的数据进行分析，在多项式拟合曲线中，能够良好的反应出曲线的性状并保证运行速度的应为三、四、五次多项式，所以在寻找最优解的过程中，我们应当以上述三种多项式来进行拟合，从而达到最好的效果。然而同时多项式拟合也反映出了其精度不高的问题，即便是效果最好的三次、四次多项式，随机取点的样本中拟合出来并进行自行评估后所得到的最低成本也没未达到期望值，因此我们认为需要考虑更为精确的拟合手段。

2.2 三次样条插值拟合

2.2.1 概述

样条插值是使用一种名为样条的特殊分段多项式进行插值的形式。由于样条插值可以使用低阶多项式样条实现较小的插值误差，这样就避免了使用高阶多项式所出现的龙格现象。对于(n+1)个给定点的数据集{Xi}，我们可以用n段三次多项式在数据点之间构建一个三次样条。用图2表示对函数f进行插值的样条函数，需要：

（1） 插值特性，S(Xi)= f(Xi)

（2） 样条相互连接，$S_{i-1}(Xi)=S_i(Xi)$,i=1,2...n-1

（3） 两次连续可导,$S_{i-1}'(Xi)=S_i'(Xi)$,同时二阶导数也相等

$$S(x)=\begin{cases} S_0(x), & x\in[x_0,x_1] \\ S_1(x), & x\in[x_1,x_2] \\ \cdots & \\ S_{n-1}(x), & x\in[x_{n-1},x_n] \end{cases}$$

图2

由于每个三次多项式需要四个条件才能确定曲线形状，所以对于组成S的n个三次多项式来说，这就意味着需要4n个条件才能确定这些多项式。但是，插值特性只给出了（n+1）个条件，内部数据点给出n+1-2=n-1个条件，总计是4n-2个条件。我们还需要另外两个条件，根据不同的因素我们可以使用不同的条件。

在本课题中我们使用六段三次曲线样条来近似表达曲线性状，取法如图3所示。

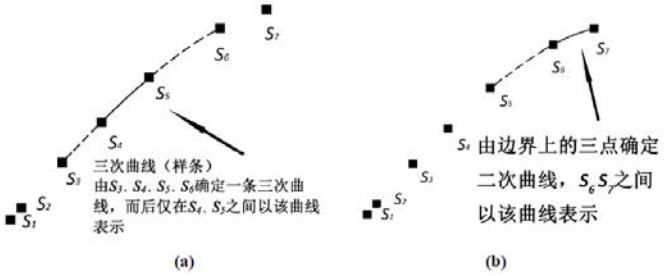

图3

2.2.2 MATLAB自带三次样条插值工具

运用MATLAB软件自带的三次样条拟合函数spline()，其输入值为两组已知的数据以及所求点，返回值为所求点在已知数据的三次样条拟合下的数值。MATLAB自带的函数可以直接解决在非中间点的数据拟合，故可直接使用。由三次样条插值所得成本略微低于多项式拟合的最佳结果，特别是运行时间较短。可能原因是，直接使用MATLAB系统自带的函数其拟合精度优于多项式拟合，同时拟合时间较短。

2.2.3 结果分析

通过与多项式拟合结果的比较，三次样条插值拟合的方式相对于多项式拟合在精确度方面有了明显的提升。但是感觉MATLAB自带的三次样条插值的spline()函数在成本计算所得结果仍可通过改进来得到提升，因此在接下来的拟合中，将采用退火法进行优化。

3 模拟退火算法（SA）

方法介绍：用固体退火模拟组合优化问题，将内能 E 模拟为目标函数值 f，温度 T 演化成控制参数 t，即得到解组合优化问题的模拟退火算法：由初始解 i 和控制参数初值 t 开始，对当前解重复“产生新解→计算目标函数差→接受或舍弃”的迭代，并逐步衰减 t 值，算法终止时的当前解即为所得近似最优解。模拟退火算法可以分解为解空间、目标函数和初始解三部分。退火过程由冷却进度表(Cooling Schedule)控制，包括控制参数的初值 t 及其衰减因子△t、每个 t 值时的迭代次数 L 和停止条件 S。

算法实现过程：

(1) 初始化：初始温度 T(充分大)，初始解状态 S(是算法迭代的起点)， 每个 T 值的迭代次数 L

(2) 对 k=1，……，L 做第(3)至第(6)步：

(3) 产生新解 S′

(4) 计算增量△t′=C(S′)-C(S)，其中 C(S)为评价函数

(5) 若△t′<0 则接受 S′作为新的当前解，否则以概率 exp（-△t′/T）接受 S′作为新的当前解.

(6) 如果满足终止条件则输出当前解作为最优解，结束程序。

终止条件通常取为连续若干个新解都没有被接受时终止算法。

(7) T 逐渐减少，且 T 趋向于 0，然后转第(2)步。

优缺点分析：

使用 MATLAB 编程，相比有遗传算法，代码易写，计算时间也较短。但是缺点是在实际执行中，结果成本并不是稳定达到所能达到的最小值分，具有一定的偶然性。

4 最终确定测定方法

4.1 确定具体方法

通过分析比对，我们确定对每组 51 个数据，选取其中的确定点数进行拟合并计算适用度。每组数据选取的点的下标一致。考虑到难易程度，时间和效率的问题，我们最终确定使用模拟退火算法，拟合过程同时使用三次样条插值法和三次多项式拟合法。

4.2 解决方法实现

(1) 读入老师提供的 469 组数据。

(2) 通过之前的拟合，从选取 3 个点开始进行，考虑到 9 个点所产生的测定成本已经大于之前所得的结果，确定于 9 个点结束。

(3) 随机函数随机排序并选取前确定个数的点个点，再次排序，得到随机的几个点。

(4) 用模拟退火算法作为大循环。

(5) 在循环内前部分加入随机变化一个点生成新的确定点个数的点组合。并用拟合法得到对应的成本。
(6) 循环中间部分加入计算去掉两个端点的 49 个点的评价函数的分值，并计算 469 组数据的平均成本。
(7) 循环的后半部分决定是否接受这确定个数的点。若成本小于最小成本，则接受；若大于，则以 exp（-△t′ /T）的概率接受。其中 T 为温度。
(8) 输出最小成本。
(9)通过比较所得取不同个数测定点的结果，确定最终选取点的个数，得出最终结果。

4.3 结果分析

经过上述流程，最终得出以此方法所得的最小测定成本为 105.2，所选取的测定点为 1,5,22,29,36,48,50.可以清楚的看到，相比于以上的直接拟合结果，测定成本已经降低了很大一部分，表明这种测定方法相对比较优秀。

5 参考文献

[1] 袁炎. 统计推断讲座 2：课题由来和基本问题的提出
[2] 袁炎. 统计推断讲座 2：课题由来和基本问题的提出
[3] 网络资源 China-pub.com 《matlab 教程》
[4] 网络资源 http://baike.baidu.com/view/18185.htm

附页

最终方法对应程序

```
function go( )
KCdata=xlsread('20141010dataform.csv');
D=KCdata(1:2:end,1:end);
U=KCdata(2:2:end,1:end);
A=randperm(51);
B=sort(A(1:7));
B_min=B;
cost_min=10000;
Tf=0.01;
Tk=100;
tic;
while Tk>Tf
n=0;
while n<Tk/10;
n=n+1;
```

```
remain=setdiff(A,B);
E=remain(randperm(44));
F=randperm(7);
S=B;
S(1,F(1))=E(1,F(1)+1);
S=sort(S);
cost=zeros(1,469);
M=zeros(469,51);
for i=1:469
X=D(i,S);
Y=U(i,S);

M(i,:)=U(i,:)-interp1(X,Y,D(i,:),'spline');
for j=1:51
if abs(M(i,j))<=0.5
cost(1,i)=cost(1,i)+0;
elseif abs(M(i,j))<=1 && abs(M(i,j))>0.5
cost(1,i)=cost(1,i)+0.5;
elseif abs(M(i,j))<=2 && abs(M(i,j))>1
cost(1,i)=cost(1,i)+1.5;
elseif abs(M(i,j))<=3 && abs(M(i,j))>2
cost(1,i)=cost(1,i)+6;
elseif abs(M(i,j))<=5 && abs(M(i,j))>3
cost(1,i)=cost(1,i)+12;
else cost(1,i)=cost(1,i)+25;
end
end
cost(1,i)=cost(1,i)+12*7;
if (cost(1,i)<0)
cost(1,i)=0;
end
end
ave_cost=sum(cost,2)/469;
if ave_cost<cost_min
cost_min=ave_cost;
cost_save=ave_cost;
B_min=S;
B=S;
elseif rand<exp((ave_cost-cost_save)/Tk)
cost_save=ave_cost;
B=S;
end
cost_min;ave_cost;
end
```

```
Tk=Tk*0.97;
end
cost_min
B_min
toc;
```