统计推断在数模转换系统中的应用

50 组　孙煜　5130309236

摘要：本课程要运用 Matlab 软件，采用模拟退火算法统计 469 组数据（每组 51 个），并结合统计推断学的方法，完成从特征点的选取、特征点反推曲线关系到最后的方案评估等一系列过程，旨在提出一个合理的数学模型，学会从少量数据点推断大量数据点的方法。最终得出结论

关键词：拟合，算法实现，最优解

1. 引言

统计推断[statistical inference] 根据带随机性的观测数据（样本）以及问题的条件和假定（模型），而对未知事物做出的，以概率形式表述的推断。统计推断问题常表述为如下形式：所研究的问题有一个确定的总体，其总体分布未知或部分未知，通过从该总体中抽取的样本（观测数据）做出与未知分布有关的某种结论。

在工程实践中，我们常借助统计推断的理念，对变量间的函数关系进行分析。通常选取能代表变量之间关系的特征点，经过多种拟合方式进行拟合，从而求解出能够代表海量变量之间函数关系的最优化方式。

2. 课题概述

假定有某型投入批量试生产的电子产品，其内部有一个模块，功能是监测某项与外部环境有关的物理量（可能是温度、压力、光强等）。该监测模块中传感器部件的输入输出特性呈明显的非线性。本课题要求为该模块的批量生产设计一种成本合理的传感特性校准（定标工序）方案。

3. 研究目的

考虑到工业化生产下的成本、效率因素，针对该系统的某一实例，实测其 51 个输出电压对应的输入是不经济的。因而考虑跳点测定，在保证所要求精度的前提下，通过测定 7 组电压值（即特征点）对应的被检测物理量来推断整体的输入—输出电压(U)特性关系研究给定的样本的这一特性关系，确定这 7 个特征点，并提出由此特征点集推断出整体特性关系的函数逼近方法，最后评估该方案对给定样本的适用度。

4. 模型

为了对本课题展开有效讨论，需建立一个数学模型，对问题的某些方面进行必要的描述和限定。

监测模块的组成框图如图 1。其中，传感器部件（包含传感器元件及必要的放大电路、调理电路等）的特性是我们关注的重点。传感器部件监测的对象物理量以符号 Y 表示；传感部件的输出电压信号用符号 X 表示，该电压经模数转换器（ADC）成为数字编码，并能

被微处理器程序所读取和处理，获得信号 $\hat{Y}$ 作为 Y 的读数（监测模块对 Y 的估测值）。

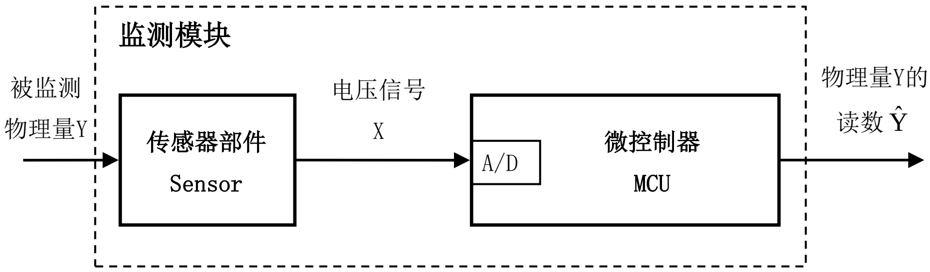


图 1 监测模块组成框图

所谓传感特性校准，就是针对某一特定传感部件个体，通过有限次测定，估计其 Y 值与 X 值间一一对应的特性关系的过程。数学上可认为是确定适用于该个体的估测函数 $\hat{y} = f(x)$的过程，其中 x 是 X 的取值，$\hat{y}$ 是对应 Y 的估测值。

考虑实际工程中该监测模块的应用需求，同时为便于在本课题中开展讨论，我们将问题限于 X 为离散取值的情况，规定

$$X \in \{x_1, x_2, x_3, ..., x_{50}, x_{51}\} = \{5.0, 5.1, 5.2, ..., 9.9, 10.0\}$$

相应的 Y 估测值记为 $\hat{y}_i = f(x_i)$， Y 实测值记为 y_i， $i = 1, 2, 3, ..., 50, 51$。

1.传感部件特性

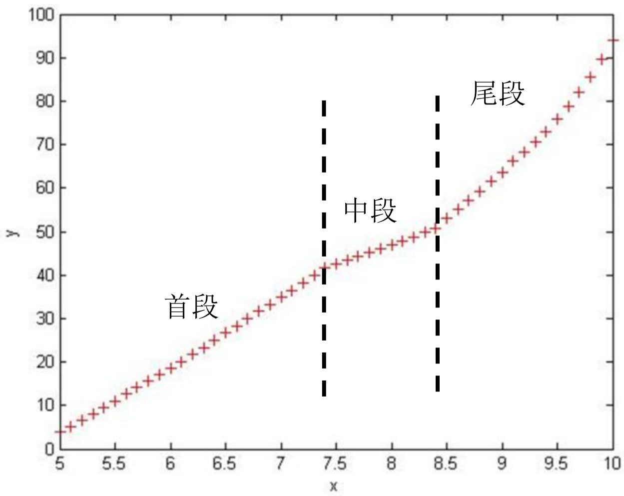


图 2 传感特性图示

一个传感部件个体的输入输出特性大致如图 2 所示，有以下主要特征：

- Y 取值随 X 取值的增大而单调递增；
- X 取值在[5.0,10.0]区间内，Y 取值在[0,100]区间内；
- 不同个体的特性曲线形态相似但两两相异；
- 特性曲线按斜率变化大致可以区分为首段、中段、尾段三部分，中段的平均斜率小于首段和尾段；
- 首段、中段、尾段单独都不是完全线性的，且不同个体的弯曲形态有随机性差异；

- 不同个体的中段起点位置、终点位置有随机性差异。

为进一步说明情况，图 3 对比展示了四个不同样品个体的特性曲线图示。

2.标准样本数据库

前期已经通过试验性小批量生产，制造了一批传感部件样品，并通过实验测定了每个样品的特性数值。这可以作为本课题的统计学研究样本。数据被绘制成表格，称为本课题的“标准样本数据库”。

该表格以 CSV 格式制作为电子文件。表格中奇数行存放的取值，偶数行存放对应的取值。第 2i－1 行存放第 i 个样本的 X 数值，第 2i 行相应列存放对应的实测 Y 数值。

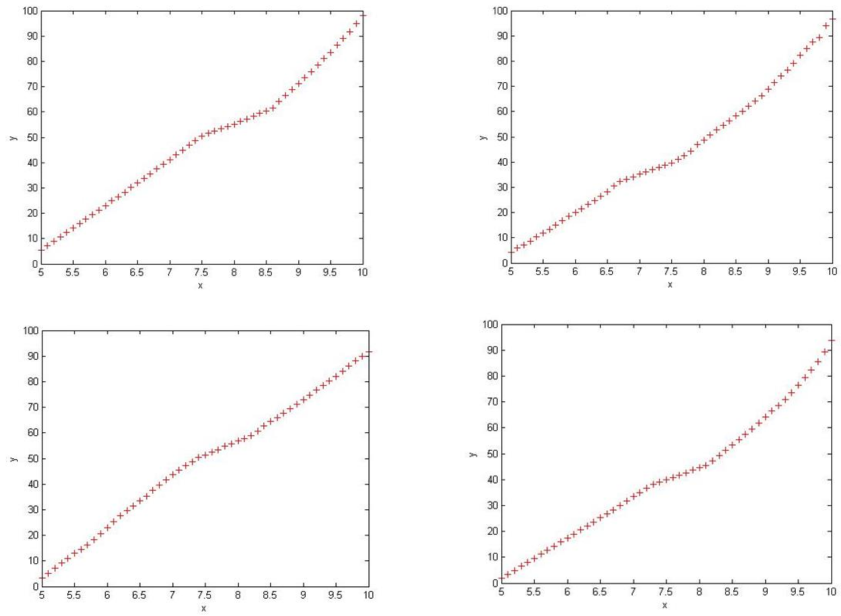


图 3 四个不同样本个体特性图示对比

3.成本计算

为评估和比较不同的校准方案，特制定以下成本计算规则。

- 单点定标误差成本

$$s_{i,j}=\begin{cases}0 & \text{if } \left|\hat{y}_{i,j}-y_{i,j}\right|\le 1\\ 1 & \text{if } 1<\left|\hat{y}_{i,j}-y_{i,j}\right|\le 2\\ 2 & \text{if } 2<\left|\hat{y}_{i,j}-y_{i,j}\right|\le 3\\ 4 & \text{if } 3<\left|\hat{y}_{i,j}-y_{i,j}\right|\le 5\\ 10 & \text{if } \left|\hat{y}_{i,j}-y_{i,j}\right|>5\end{cases}\qquad (1)$$

单点定标误差的成本按式（1）计算，其中 $y_{i,j}$ 表示第 i 个样本之第 j 点 Y 的实测值，$\hat{y}_{i,j}$ 表示定标后得到的估测值（读数），该点的相应误差成本以符号 $s_{i,j}$ 记。

- 单点测定成本

实施一次单点测定的成本以符号 q 记。本课题指定 q=20。

- 某一样本个体的定标成本

$$S_i = \sum_{j=1}^{51} s_{i,j} + q \cdot n_i \qquad (2)$$

对样本 i 总的定标成本按式（2）计算，式中 n_i 表示对该样本个体定标过程中的单点测定次数。

- 校准方案总体成本

按式（3）计算评估校准方案的总体成本，即使用该校准方案对标准样本库中每个样本个体逐一定标，取所有样本个体的定标成本的统计平均。

$$C = \frac{1}{M}\sum_{i=1}^{M} S_i \qquad (3)$$

总体成本较低的校准方案，认定为较优方案。

5. 方法探索

1. 暴力穷举法

对于此种问题，最简单也是最能反映整体的方法是暴力穷举法。但是由于数据过多，需要进行 C_{49}^{7} 次拟合或者求解方程，而这对于普通的计算机来说，是个巨大的数目，无法进行如此多的拟合，或者求解多项式。因此，必须改变算法，利用统计知识来求解。

2. 分段枚举法

对于给定的 51 个点，分成 7 段，每段 7-9 个点进行求取。由于在对一个样本进行测量时，我们需要兼顾在各个不同范围内的数据情况，因此，51 个点应当是离散的，又由于其整体图线线性程度较高，故可以较为均匀地对 51 个点进行分段。对于每一段的数值进行穷举，求得 C 的最小值对应点。可以减少暴力穷举法消耗的时间，但是所需时间依旧很多。

3. 模拟退火法

在温度较高的情况下，系统有较大的概率接受误差较大的分子（类比于温度较高时，各分子的动能波动范围大），此种模糊判断的好处是避免了系统过快地局部收敛。随着温度的降低，可接受变异误差范围逐渐减小，最终趋于稳定，即得出最优解

4. 遗传算法

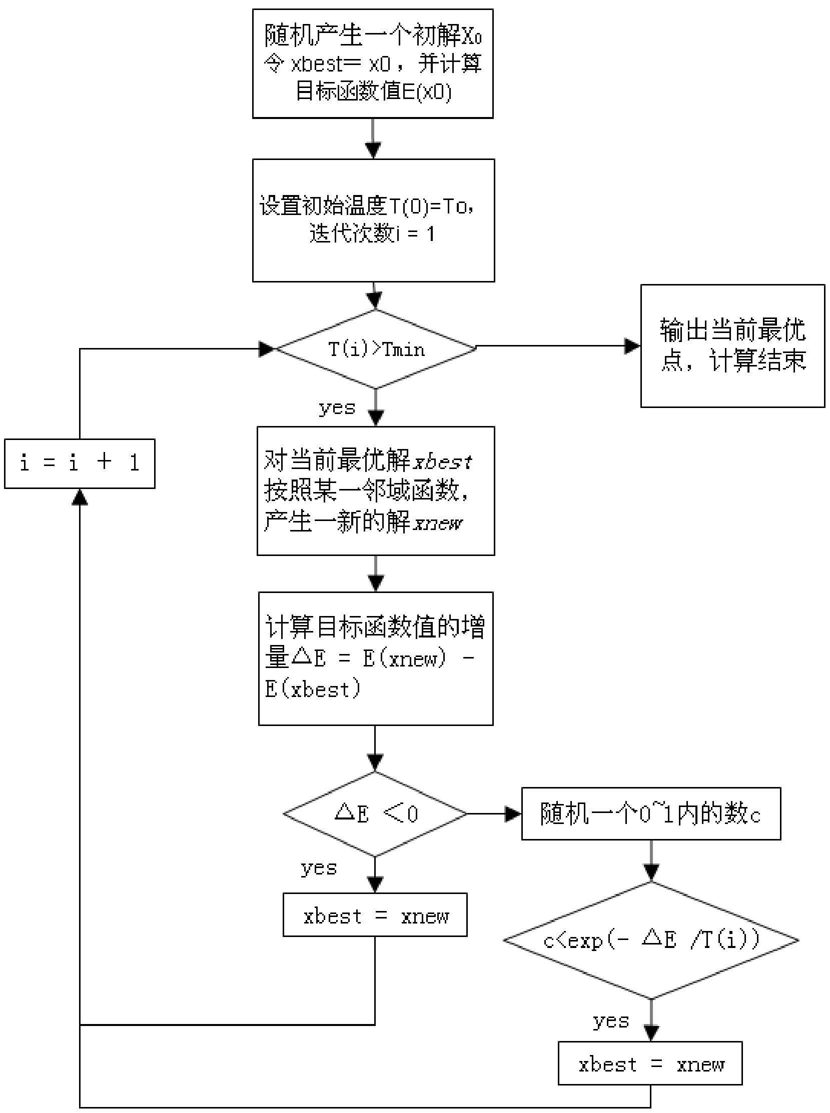
首先随机选择亲代样本，对其进行评估，筛选掉适应度较低的样本，留下适应度高的样本，并对适应度高的样本进行交叉配对，产生第二代。再从整体中随机取得另一部分样本加入第二代。通过不断产生下一代，适应度高的样本逐渐留下来，适应度低的样本被淘汰，最后得到适应度最高的一组。此种算法可以大大减少计算时间，不过由于程序较为繁琐，对于编程人员要求较高，未能实现。

6. 方法选择及算法实现

1. 方法选择

本组选用模拟退火算法，模拟退火算法来源于固体退火原理，将固体加温至充分高，再让其徐徐冷却，加温时，固体内部粒子随温升变为无序状，内能增大，而徐徐冷却时粒子渐趋有序，在每个温度都达到平衡态，最后在常温时达到基态，内能减为最小。根据 Metropolis 准则，粒子在温度 T 时趋于平衡的概率为 e-ΔE/(kT)，其中 E 为温度 T 时的内能，ΔE 为其改变量，k 为 Boltzmann 常数。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，这是基于蒙特卡罗迭代求解法的一种启发式随机搜索过程。

其算法流程如下



2. 计算 C 值

设矩阵 Xij（i=1,2,…,419; j=1,2,…,51），Yij （i=1,2,…,419; j=1,2,…,51），

某个解 $(s_1, s_2, \cdots, s_7)$

6.2.1 对应关系 f

为方便,设对应关系为 Y=f(X)。

三次拟合法下,对每个 Xi,只需对取得的七个点进行三次多项式拟合,即可得 fi(X)。（其中 i=1,2,3....418）。

三次样条插值法中，Xi,sk-1 到 Xi,sk 之间的 Y-X 关系由过 Sk-1,Sk,Sk+1,Sk+2 四点的三次曲线确定,设曲线方程为 $Y=a1X^3+a2X^2+a3X+a4$。由四点的值可以求得对应的系数值。所以对每个 i,可以得到一个分段函数形式的 Y=f (X)。

其它拟合法下类似。

6.2.2 由 f 得到对应 M 值

对于矩阵 Xij（i=1,2,…,419; j=1,2,…,51）中的每一个值，代入 Y=f(X)中，得到矩阵 Y‘ij （i=1,2,…,419; j=1,2,…,51)。

然后根据 3.3 的（1)（2)（3）式得出对应的 M。

3. 获得最优解

对所得的算法，考虑其运算时间和成本，可信度等，总结出最优的解。由于算法部分仍未全部完成，所以其具体分析留待算法完成后进行。

4. 算法的实现及程序运行结果

详细算法请见附录,其中代码选用模拟退火法和三次样条插值拟合,取点个数为 7 个,（由于在其他拟合法或取点个数的情况下，代码差别不大，因此仅附上得到最优解的代码。若要测试其他情况,只需修改 best_array 元素个数和初值,或者修改 get_score 中的拟合函数）

最好的运行结果为三次样条插值拟合下

取[2，9，20，26，34，45，50]

成本为 93.8998

7. 结果分析

通过模拟退火法和三次样条插值法得出的最低成本为 93.8998。事实上，通过多组数据的观察，该种取点与拟合法得到的成本固定在 94 左右。在我将 spline 插值换作 cubic 插值后，结果并无明显变化，所以在后续的运行中沿用 spline 插值，同时也便于 test 程序的检验。

由于取点的个数受到已有经验的影响，因此一开始便选作 7 个，但是仅凭单一种取点个数得到的结论具有局限性,于是我又尝试了其他个数的情况,其中换作取 5 个点和 6 个点时,均不能得到成本为 110 以下的解（在我的程序中，初始最优解的成本应小于 110，可在代码内看出)，只有在取点个数为 8 的时候，可得到成本为 105 左右的解，仍大于已得的最小成本。故三次样条插值法下 7 个点的取法最优。

然后我又考虑改变拟合的方法，即将三次样条插值法改作多项式拟合。首先尝试的是 3 次多项式拟合。结果不尽人意，在取点个数为 5，6，7，8 的情况下都未能取得成本小于 110 的解。接着我又尝试了 4 次的拟合，本来在以往的观测中 4 次似乎较 3 次更优，可是还是未能得出成本小于 110 的解，在 7 个点的情况下最好，但成本仍在 115 左右。不知道是否是运行代数的缘故，多项式拟合未能取得好的结果。

所以我最后得出的结论为三次样条插值法下取 7 个点的计算成本最小。

8. 心得体会及工作展望

第一次接触到 Matlab 这个软件，难免感到十分生疏，不过好在 Matlab 的设计十分的人性化，上手起来并不是很难。虽然说起来 Matlab 比 c++感觉更人性化，但是毕竟我接触的还只是最基本的一些操作，即使对于 Matlab 中数组和矩阵，其灵活的操作方法总是容易让我混淆，不过好在到最后还是成功得到了结果。

但在这一过程中还存在一些问题，比如在模拟退火的代码中，对于变动的那个元素，我仅限定了它的粗略范围，但没有添加防止其与其他元素重复的情况，这样在运行中，出现重复情况后，会导致三次样条插值错误，使程序停止。我有考虑过添加代码进行判断，可是修改后的代码由于多了这一判断的操作会是运行的效率降低，运行时间变得很长（可能是我写的判断代码还不够好），考虑到运行时间问题以及出现重复的概率非常低，因此我还是选择沿用开始的代码。这一问题是我在以后的过程中需要优化的，毕竟对于使用者，无论出错的概率多小，程序都应避免。

然后，在与老师的免谈是十分重要的一环。因为在面谈之前，我的代码存在漏洞，得出的结果和测试程序得到的大不相同，本来我以为只是因为与测试程序的拟合法不一样，可是老师指出先解决这个才是首要，于是我又回去经过调试，发现原代码中确实存在问题，原来得到的结果也不为最优解，最后经修改终于得到了和测试程序一样的结果。

虽然最后得出了结果，但是该问题的解决还存在诸多路径去探寻。比如遗传算法，本来考虑在面谈之后尝试下，但是由于面谈中发现了问题，解决之后对我来说时间就不太充足了，于是没有做遗传算法这一块。事实上，遗传算法算法与退火法都是对当前的最优解进行扰动，而避免所得到最优解为局部最优解的过程，只是遗传算法更精细，结果受随机影响的概率较小。当然，这并不代表退火法不够好，它计算过程简单，算法通用，可用于求解复杂的非线性优化问题。其实，在编写遗传算法算法的过程中，可以局部应用退火法，使算法效率更高。不过，这就是在完成遗传算法以后的事了。

总的来说，通过这个课程，学到了许多，不仅是数学工具的使用，也有研究方法等，对报告的编写也有了一定认识。比较遗憾的是，程序不尽完美，遗传算法没有完成。不过就当作是以后的努力的方向。

9. 参考文献

[1]统计与推断讲座 3：问题的求解路经.ppt

[2]统计推断：第 nn 组（组长姓名）课程报告设计.doc

[3]“统计推断”课程设计的要求 V2.0 2014-10-10.doc

[4] 陈华根,吴健生,王家林,陈冰　模拟退火算法机理研究　同济大学学报(自然科学版) 第 32 卷第 6 期 2004 年 6 月

[5]曲强，陈雪波　基于ＭＡＴＬＡＢ的模拟退火算法的实现　鞍山科技大学学报 26 卷第 3 期　2003 年 6 月

[6]雷英杰.张善文。 MATLAB 遗传算法工具箱及应用。西安电子科技大学出版社

附录

程序源代码：

Main.m

```
data=csvread('20141010dataform.csv');%读取数据
best_array=zeros(1,7);%best_array 用来储存结果相对较好的数组
best_cost=0;
sum_cost=0;
T=100;
cost=200;%初始化平均得分

while cost>=110  %找到一个平均成本小于 110 的数组，用于稍后微调
    sum_cost=0;
    arr1=randi([43,51]);
    arr2=randi([36,42]);
    arr3=randi([29,35]);
    arr4=randi([22,28]);
    arr5=randi([15,21]);
    arr6=randi([8,14]);
    arr7=randi([1,7]);
    random_array=[arr7,arr6,arr5,arr4,arr3,arr2,arr1];%产生初始解
    for i=1:469;
        temp=data([2*i-1,2*i],:);
        y_base=temp(2,:);
        x_base=temp(1,:);
        x_tmp=x_base(1,random_array);
        y_tmp=y_base(1,random_array);
        sum_cost=sum_cost+Get_cost(x_base,y_base,x_tmp,y_tmp);
    end
    cost=sum_cost/469;
    best_cost=cost;
    best_array=random_array;
end

while T > 0.1
    T = T/(1+0.05);
    sum_cost=0;
    disp(cost)
```

```
    for i=1:7
        while true
            b=random_array(i);
            c=randi(3);
            b=b+c-2;
            if (b>0&&b<=51)%保证产生变动时不会超过范围
                break
            end
        end
        random_array(i)=b;
    end
    for i=1:469;
        temp=data([2*i-1,2*i],:);
        y_base=temp(2,:);
        x_base=temp(1,:);
        x_tmp=x_base(1,random_array);
        y_tmp=y_base(1,random_array);
        sum_cost=sum_cost+Get_cost(x_base,y_base,x_tmp,y_tmp);
    end
    cost=sum_cost/469;
    e=exp(-(cost-best_cost)/T*100);
    if (cost<=best_cost)
        best_cost=cost;
        best_array=random_array;
    elseif(rand>e)%以 1-exp（-成本差 /温度）的概率不接受改变
        random_array=best_array;
    end
end
disp(best_array);
disp(best_cost);
```

Get_cost.m

```
function re=Get_cost(x_base,y_base,x_tmp,y_tmp)
per_cost=0;
x=[5.0:0.1:10.0];
y=interp1(x_tmp,y_tmp,x,'spline');
for w=1:51
    d=abs(y(w)-y_base(w));
        if d<=0.5,
            per_cost=per_cost;
        elseif d<=1,
            per_cost=per_cost+0.5;
        elseif d<=2,
            per_cost=per_cost+1.5;
```

```
        elseif d<=3,
            per_cost=per_cost+6;
        elseif d<=5,
            per_cost=per_cost+12;
        elseif d>5,
            per_cost=per_cost+25;
        end;
end;
per_cost=per_cost+12*7;
re=per_cost;
```