文章编号:1009－3486(2004)03－0066－03

小样本采样数据的预处理

吴文全，察 豪

(海军工程大学 电子工程学院，湖北 武汉 430033)

摘 要：一般情况下数据处理大多采用数理统计方法，该方法对于数据较少情况下处理和判别粗大误差不适用，提出了运用线性均方估计法和熵值判别法来处理和判别粗大误差．线性均方估计消除粗大误差是一种采用软化的方法处理粗大误差；熵值判别法是根据熵的上界对应最大的不确定度，利用所得数据的熵信息量判别数据是否含有粗大误差．这两种方法经过多个实例计算，结果表明，它们在处理小样本采样数据时更有效．

关键词：预处理；小样本；线性均方；熵

中图分类号：TG801 **文献标识码**：A

Preprocessing of small sampling data

WU Wen-quan, CHA Hao

(Electronic Eng. College, Naval Univ. of Engineering, Wuhan 430033, China)

Abstract: The method of mathematical statistic is usually used to process the data, but it is not suitable to process small quantity of data. Thus a new method based on the linear mean squares estimation and entropy judging method is used to distinguish the gross errors. LMSE is a way by softening the gross errors; the entropy distinguishing is based on the upper limit that is maximum error of the entropy, thus it can be used to distinguish the gross errors of data. Many examples prove that the two methods are more effective when they are used to process small sampling data.

Key words: preprocessing; small sample; linear mean square; entropy

在数据采集中，由于受条件的限制，对同一个量的采样数据通常个数较少，所得到的数据也可能含有粗大误差，这样得到的原始数据如果直接就进行数据处理(如：信号的估计和检测、参数的运算等)不仅得不到真实值，而且还可能引起误判断，导致误操作．为了减少这种无谓的失误，随着计算机的发展，对采样得到的数据常需要进行自动预处理，即消除含有粗大误差的数据，使所采样的数据用来估计测量量具有无偏性、一致性和有效性[1]．一般情况下数据处理大多采用数理统计的方法[2]，这种方法对于数据较少的情况不太适用，特别是对珍贵难采样数据处理基本无效．作者提出了运用线性均方估计法和熵值判别法来消除和判别粗大误差[3~5]．

1 线性均方估计方法

在线性均方(linear mean squares，LMS)估计中，所要得到的参数表示为采样数据的线性加权和．

设采样数据为：$x_1, x_2, \cdots, x_N, x_i$ 中可能含有粗大误差，$x = w_1x_1 + w_2x_2 + \cdots + w_Nx_N = \sum_{i=1}^{N} w_ix_i$，$w_i$ 为所需要确定的权系数．线性均方估计就是使采样数据的均方误差最小，从而得到权系数，也就是说根

收稿日期：2003-12-21；**修订日期**：2004-03-20

作者简介：吴文全(1972-)，男，讲师，硕士生．

据采样所得到的原始数据，不管其是否含有粗大误差，总可以在保留这些数据的前提下，求出其权系数，使得采样数据的均方误差最小，这样采样数据的加权和就可以较准确表示所需参数.

根据线性均方原理可得

$$\min\boldsymbol{E}\{(x-X)^2\}=\min\boldsymbol{E}\{(\sum_{i=1}^{N}w_ix_i-X)^2\}=\min\boldsymbol{E}\{(e)^2\} \tag{1}$$

式中：$e=x-X$ 为估计误差.

求(1)式相对于 w_k 的偏导，并令其为零，则得：

$$\frac{\partial\boldsymbol{E}[e^2]}{\partial w_k}=\boldsymbol{E}\{\frac{\partial e^2}{\partial w_k}\}=2\boldsymbol{E}\{e\frac{\partial e}{\partial w_k}\}=2\boldsymbol{E}\{ex_k\}=0$$

即

$$\boldsymbol{E}\{e\quad x_i\}=0\qquad i=1,2,\cdots,N \tag{2}$$

(2)式就是正交性原理，即要使均方误差最小，当且仅当估计误差 e 正交于每个采样数据 x_i，其中：$i=1,2,\cdots,N$.

由于 $\boldsymbol{E}\{x\}=\boldsymbol{E}\{\sum_{i=1}^{N}w_ix_i\}=\sum_{i=1}^{N}w_i\boldsymbol{E}\{x_i\}\boldsymbol{E}=\boldsymbol{E}\{X\}$，即 x 是无偏的.

考虑到 x 的无偏性和(2)式的正交条件，均方误差可写成：$\boldsymbol{E}\{e^2\}=\boldsymbol{E}\{e(x-X)\}=\boldsymbol{E}\{ex\}$，这表明线性均方估计误差等于数据估计误差与被估计参数乘积的均值.

为了推导出权系数，将(2)式改写为：

$$\boldsymbol{E}\{(\sum_{i=1}^{N}w_ix_i-X)x_i\}=0\qquad i=1,2,\cdots,N \tag{3}$$

令 $g_i=\boldsymbol{E}\{Xx_i\}$ 和 $R_{ij}=\boldsymbol{E}\{x_ix_j\}$，则(3)式可以简化为：

$$\sum_{k=1}^{N}R_{ik}w_k=g_i\qquad i=1,2,\cdots,N \tag{4}$$

若记 $\boldsymbol{R}=[R_{ij}]_{i,j=1}^{N}$，$\boldsymbol{W}=[w_1,w_2,\cdots,w_n]^{\mathrm{T}}$，$\boldsymbol{G}=[g_1,g_2,\cdots,g_n]^{\mathrm{T}}$. 则(4)式可表示为：$\boldsymbol{W}=\boldsymbol{R}^{-1}\boldsymbol{G}$.

需说明的是，由于采样数据 $x_1,x_2,\cdots,x_N$ 相互独立，所以相关矩阵 $\boldsymbol{R}$ 是非奇异，在计算 $\boldsymbol{G}$ 时由于未知，这时可用中间值的期望对其进行逼近[5].

线性均方估计步骤如下：

首先求出采样数据 $x_1,x_2,\cdots,x_N$ 的中间值(去掉最大值及最小值)的期望值，作为计算 g_i 的 X 值；

然后计算 $\boldsymbol{R}^{-1}=[\boldsymbol{E}(x_ix_j)]^{-1}=\left\{\boldsymbol{E}\begin{bmatrix}x_1x_1 & x_1x_2 & \cdots & x_1x_N\\ x_2x_1 & x_2x_2 & \cdots & x_2x_N\\ \vdots & \vdots & & \vdots\\ x_Nx_1 & x_Nx_2 & \cdots & x_Nx_N\end{bmatrix}\right\}^{-1}$；

再计算 $\boldsymbol{G}=\begin{bmatrix}g_1\\ g_2\\ \vdots\\ g_N\end{bmatrix}=\boldsymbol{E}\begin{bmatrix}xx_1\\ xx_2\\ \vdots\\ xx_N\end{bmatrix}$；

最后计算 $\boldsymbol{W}=\boldsymbol{R}^{-1}\boldsymbol{G}$.

2　熵判别方法

在信息论中，信息量 $I(x_k)$ 表示观测到一个以概率 p_k 发生的事件 x_k 的信息.

定义 $I(x_k)=\log(\frac{1}{p_k})=-\log p_k$，定义熵为 $H(x)=\boldsymbol{E}\{I(x_k)\}=\boldsymbol{E}[-\log p_k]$，熵与方差之间存在 $H(x)=\log(A\sigma)$，A 为与 p_k 有关的常数，σ 为方差. 若取以 e 为底的对数，$H(x)=\ln(A\sigma)$. 可得 $e^{H(x)}=A\sigma$，当置信概率为 95% 时，不确定度 $\Delta x=\pm\frac{3}{4}e^{H(x)}$[6].

由于小样本采样得到的数据是离散的，均信息量的熵也应该是离散熵 $H(x)$，$H(x)=-\sum_{k=i}^{N}p_k\ln p_k$.

由于采样的样本较少，不能用统计频数代替概率估计，这时应采用秩估计的方法进行熵估计. 具体方法如下：

(1) 将采样数据 $x_1,x_2,\cdots,x_N$ 按从小到大的顺序排成新的序列 $x_{(1)},x_{(2)},\cdots,x_{(N)}$；

(2) 定义秩 r_k 为 $r_k=\int_{-\infty}^{x(k)}p(x)\mathrm{d}x=\int_{-\infty}^{x(k)}\mathrm{d}p(x)=P(x(k))$，$p(x)$ 为 x 的概率分布函数，$P(x(k))$ 的估计 $\overline{p(x(k))}=\overline{r_k}=k/(n+1)$.

(3) $H(x)$ 的估计

$$\overline{H(x)}=-\sum_{k=1}^{N}\ln\left[\frac{\Delta p(x(k))}{\Delta x(k)}\right]\Delta p(x(k))=-\sum_{k=1}^{N}\ln\left[\frac{\overline{r_{k+1}}-\overline{r_k}}{x(k+1)-x(k)}\right](\overline{r_{k+1}}-\overline{r_k}).$$

(4) $\Delta x=\pm\frac{3}{4}e^{\overline{H(x)}}$.

如果 $\Delta x_i=x_i-\frac{1}{N}\sum_{i=1}^{N}x_i$ 超过 Δx 的范围，则判定 x_i 含有粗大误差.

3 实例验证及比较

用数字电压表对某一高精度稳压电源(12 V)电压进行6次测量，测得数据如表1所示.

表1 数字电压表测稳压电源数据

测量次序	1	2	3	4	5	6
电压值/V	12.13	11.97	12.03	12.01	11.98	12.04

(1) 运用线性均方估计求权系数

$$\boldsymbol{W}=\boldsymbol{R}^{-1}\boldsymbol{G}=\left\{\boldsymbol{E}\begin{bmatrix}x_1x_1 & x_1x_2 & \cdots & x_1x_N\\ x_2x_1 & x_2x_2 & \cdots & x_2x_N\\ \vdots & \vdots & & \vdots\\ x_Nx_1 & x_Nx_2 & \cdots & x_Nx_N\end{bmatrix}\right\}^{-1}*\left(\boldsymbol{E}\begin{bmatrix}xx_1\\ xx_2\\ \vdots\\ xx_N\end{bmatrix}\right)$$

式中：$X=12.015$，为去掉最大值和最小值后的均值.

将数据代入后可得 $\boldsymbol{W}=[0.164\ 9\quad 0.167\ 1\quad 0.166\ 3\quad 0.166\ 5\quad 0.167\ 0\quad 0.166\ 1]^{\mathrm{T}}$；

计算 $x=w_1x_1+w_2x_2+\cdots+w_6x_6=12.001$.

(2) 运用熵判别法

测量电压重新排序为：11.97，11.98，12.01，12.03，12.04，12.13. 减去最小测量值后可得：0，1，4，6，7，16(暂不考虑小数点).

计算熵估计值 $\overline{H(x)}=-\sum_{k=1}^{6}\ln\left[\frac{\overline{r_{k+1}}-\overline{r_k}}{x(k+1)-x(k)}\right](\overline{r_{k+1}}-\overline{r_k})=1.959\ 78$；

计算 $\Delta x=\pm\frac{3}{4}e^{\overline{H(x)}}=\pm0.053\ 2$，这样可以判断第一个电压测量值 12.13 含有粗大误差；

计算 $\overline{x}=12.006$.

(3) 用数理统计方法处理

莱以特准则判别可得 $\overline{x}\pm3\sigma=12.026\ 6\pm3\times0.057\ 5=12.026\ 6\pm0.172\ 5$；

最小值和最大值都不含粗大误差，$\overline{x}=12.026\ 6$.

格拉布斯准则可得：最小值和最大值都不含粗大误差，$\overline{x}=12.026\ 6$.

从上面的验证可以看出，用这两种方法对小样本采样数据估算信号参数所得到的结果比用数理统计方法处理更接近参数的真值.

(下转第73页)

动调整,因此滤波效果较差.而采用MPAEKF时,由于滤波过程中,对系统虚拟噪声进行了动态估计,在实时修正噪声方差的同时,也对系统模型线性化误差进行了补偿,因而减小了滤波误差,验证了MPAEKF算法的有效性.因此,本文提出的一种修正极坐标系下的自适应卡尔曼滤波算法,提高了滤波的稳定性、快速性和精确性.

参考文献:

[1] Aidala V J, Hammel S E. Utilization of modified polar coordinates for bearings-only tracking [J]. IEEE Transactions on Automatic Control,1983,28(3):283—294.

[2] Balarishnan S N. Extention to modified polar coordinates and application with passive measurements [J]. Journal of Guidance, Control and Dynamic,1980,12(6):906—912.

[3] Taek L S, Jason L. A stochastic analysis of modified gain extend Kalman filter with application to estimation with bearings only measuremnets [J]. IEEE Transactions on Automatic Control,1985,30(10):940—949.

[4] Ahmed N U. Modified extended Kalman filtering [J]. IEEE Transaction on Automatic Control,1994,39(6):1322—1326.

[5] 石章松,王树宗,刘　忠.基于SVD的机动目标自适应滤波研究与仿真 [J].海军工程大学学报,2003,15(2):53—56.

[6] 邓自立,王建国.非线性系统的自适应推广Kalman滤波 [J].自动化学报,1987,13(5):375—379.

[7] 周　荻,胡振坤,胡恒章.自适应推广Kalman滤波应用于导弹的被动制导 [J].宇航学报,1997,18((4):31—36.

(上接第68页)

4 结　论

小样本采样由于采样的样本数较少,通常不适合用数理统计方法进行预处理.这两种基于非统计方法的线性均方估计法和熵值判别法可以有效地解决这个问题.特别是线性均方估计法,它采取误差软化的方法,为珍贵采样数据的进一步处理提供了原始数据,为信号处理提供了更多的信息.

参考文献:

[1] 刘云生,孙丰瑞,张仁兴,等.燃气轮机的实时仿真及数据预处理 [J].海军工程大学学报,2002,14(2):76—79.

[2] 费业泰.误差理论与数据处理 [M].北京:电子工业出版社,1997.

[3] 向敬成.信号检测与估计 [M].北京:电子工业出版社,2001

[4] 林洪桦.现代测量误差分析及数据处理 [J].计量技术,1997,(6):41—44.

[5] 黄幼才.数据探测与抗差估计 [M].北京:测绘出版社,1999.

[6] 林洪桦.测量不确定度评定的熵方法[J].宇航计测技术,1997,(增刊):20—27.