hsj-75014(ff)
hyl-75.10
内　部

科学技术成果报告

用样条函数和
统计技术作拟合曲线

刘瑞哲

核数据编辑组

一九七六年五月

用样条函数和统计技术作拟合曲线

本文介绍了B-样条及其主要性质，讨论了一般样条加权拟合公式，以及除对各实验值分配权重外，对各家实验数据也分配适当权重，两次应用正态分布给出的样条加权拟合公式，并给出了拟合误差和最佳拟合判据公式。最后简单分析了样条节点的选择对拟合曲线的影响。

在核数据评价工作中，要广泛收集各实验工作的数据。由于测量方法、技术以及其他条件的不同，得到的数据也往往各不相同。需要对这些数据经过分析比较和一定的数学处理，求出最佳值，以便应用。在许多情况下，需要找出一个最佳拟合函数，作为一系列实验点的描述。因此，必须对数据及其分布规律进行分析。

如果收集来的数据的误差服从高斯定律，即：

1.小误差比大误差出现的机会多；

2.大小相等、符号相反的正负误差的数目近于相等；

3.极大的正误差与负误差出现的机率均非常小；

则求最佳值就相当于求出现的机率极大，也就是相当于求分布密度函数对数的极值。

设某实验数据为 $\{x_i, y_i, \triangle y_i\}$ $i=1,2,\cdots M$，其中x_i为自变量，y_i为x_i点的测量值，$\triangle y_i$为y_i的标准偏差。

如果数据的最佳拟合函数为多项式

$$f(x)=\sum_{j=0}^{n} a_j x^j \qquad a_n \neq 0,$$

则

$$\sum_{i=1}^{M}[y_i-f(x_i)]^2/\triangle y_i^2$$

应为极小值。

一般都可以用上述形式的多项式 $f(x)$ 作拟合函数，但是因为 $f(x)$ 在整个区间$[a,b]$上都是同一个多项式，所以任何一点的大误差都会牵涉到全局。对于较复杂的图形，要想达到一定的拟合程度，往往需要次数 n 很高，而次数过高，计算量增大，又不易算准。

用样条函数代替$f(x)$，就会使误差局部化，个别误差过大也不会影响全局，并且计算也较简单。

一、三次B-样条函数[1]

三次样条函数$S(x)$是分段三次多项式，即在区间$[a,b]$上给定分割

$$\Delta:\ a=\xi_1<\xi_2<\cdots\cdots<\xi_N=b.$$

在每个子区间$[\xi_i,\xi_{i+1}]$ $i=1,2,\cdots N-1$上，$S(x)$是一个三次多项式。在整个区间$[a,b]$上，$S(x)$及其一阶、二阶导数处处连续。点ξ_i $(i=1,2,\cdots N)$称为样条节点，一般来说在不同的子区间，有不同的三次多项式。

这样的三次样条函数有多种形式，我们用的是一种叫做三次B-样条函数，即用下面所定义的函数$M_i(x)$作为一组基底构成的样条函数$S(x)$。

$M_i(x)$的定义是这样的：

对于分割Δ，在点

$$\xi_l,\ \xi_{l+1},\cdots\cdots\xi_{l+4}\qquad l=1,2,\cdots N-4$$

作函数$\psi(x,t)=4(t-x)_+^3$的四次差商（x为参数，对t差商），得

$$M_l(x)=4\sum_{r=0}^{4}(\xi_{l+r}-x)_+^3/w(\xi_{l+r}),\tag{1}$$

其中

$$w(\xi_{l+r})=\prod_{\substack{j=0\\ j\neq r}}^{4}(\xi_{l+r}-\xi_{l+j}),\tag{2}$$

$$(\xi_{l+r}-x)_+^3=\begin{cases}(\xi_{l+r}-x)^3, & \xi_{l+r}>x\\ 0, & \xi_{l+r}\leqslant x\end{cases}\tag{3}$$

这样构成的$M_i(x)$具有明确的局部化意义：对于每个节点上的B-样条只有三个非零，且当$|t-l|\geqslant 4$时，

$$M_l(x)\cdot M_t(x)=0.\qquad x\in[a,b]\tag{4}$$

以等距节点为例，当

$\xi_l=-2$，$\xi_{l+1}=-1$，$\xi_{l+2}=0$，$\xi_{l+3}=1$，$\xi_{l+4}=2$时

$$M_l(x)=\frac{1}{6}\{(x+2)_+^3-4(x+1)_+^3+6x_+^3-4(x-1)_+^3+(x-2)_+^3\}.\tag{5}$$

将$M_i(x)$展成三次多项式，即

$$M_l(x)\begin{cases}0, & x\leqslant -2\\ \frac{1}{6}(x+2)^3, & -2\leqslant x\leqslant -1\\ \frac{1}{6}(x+2)^3-\frac{4}{6}(x+1)^3, & -1\leqslant x\leqslant 0\\ \frac{1}{6}(-x+2)^3-\frac{4}{6}(-x+1)^3, & 0\leqslant x\leqslant 1\\ \frac{1}{6}(-x+2)^3, & 1\leqslant x\leqslant 2\\ 0. & x\geqslant 2\end{cases}\tag{6}$$

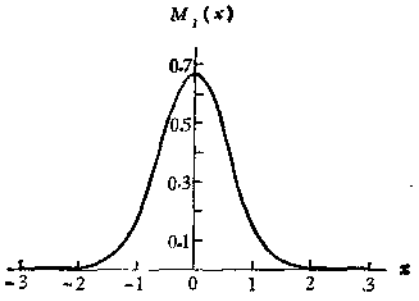

图1　等距节点函数$M_i(x)$图形

它的函数图形如图1。非等距节点也类似，只是图形不象那么左右对称而已。

由图可见，当$|t-l|\geqslant 4$时，$M_t(x)$和$M_l(x)$至少有一个为零，所以公式(4)成立。

既然$M_i(x)$是三次样条函数的一组基底，那么，在$[a,b]$上以$\xi_i(i=1,2,\cdots N)$为节点的样条函数可以唯一地写成

$$S(x)=\sum_{i=1}^{N-4}C_iM_i(x), \tag{7}$$

其中C_i为待定系数。

二、用样条函数作拟合曲线[2]

对于离散点$\{x_i,y_i,\Delta y_i\}$　$i=1,2,\cdots M$，其中x_i为自变量，y_i为在x_i的测量值，Δy_i为y_i在点x_i的标准偏差。样条节点沿x轴取在点集

$$\xi_j \qquad j=1,2,\cdots N$$

当$x\in[\xi_4,\xi_{N-3}]$时，

令
$$\chi^2=\sum_{i=1}^{M}[y_i-S(x_i)]^2/\Delta y_i^2\ , \tag{8}$$

其中$S(x)$用公式(7)代入，由

$$\frac{\partial\chi^2}{\partial C_l}=0\ , \qquad l=1,2,\cdots N-4 \tag{9}$$

得
$$\sum_{i=1}^{N-4}C_i\sum_{i=1}^{M}(\Delta y_i)^{-2}M_i(x_i)M_l(x_i)=\sum_{i=1}^{M}(\Delta y_i)^{-2}y_iM_l(x_i)\ . \tag{10}$$

$$l=1,2,\cdots N-4\ .$$

解线性方程组(10)得$C_i(l=1,2,\cdots N-4)$，然后代入公式(7)，$S(x)$就是所求拟合曲线的函数。

由公式(4)知方程组(10)的系数矩阵是七对角线矩阵，即一行的非零元素最多只有七个，计算简单。

三、多家数据的一种拟合方法

为了评价和推荐一条曲线，往往需要收集许多家的数据，这些数据很可能是彼此偏离的。引起这种偏离的原因往往是不清楚的，但经过仔细地分析比较，会使评价者对各家数据得到一个“可信程度”的印象，并合理地分配每家的权重。一般各家对于一个物理量 y 的测量，都给出一个标准偏差 Δy，它出现的可能性用权重 $1/\Delta y^2$ 表示。另外，在评价者看来，不同的实验类型 C_i 有不同的标准偏差 Σc_i。如果某家实验数据属于类型 C_i，那么，它的所有测量值都具有偏差 b_k。b_k 是具有零中项和标准偏差 $\Sigma_k(=\Sigma c_i)$ 的高斯变量，它出现的可能性用权重 $1/\Sigma_k^2$ 表示。

设有实验数据

$$(x_{ik},\ y_{ik},\ \Delta y_{ik},\ \Sigma_k) \quad i=1,2,\cdots n_k,\ k=1,2,\cdots K\ .$$

脚标“ik”表示第 k 家数据的第 i 次测量，Σ_k 是评价者自己给的。如果这些数据能被样条函数 $S(x)$ 表示，那么，对于每个 x_{ik} 和 y_{ik}，量 $y_{ik}-S(x_{ik})$ 是一个具有中项 b_k 和标准偏差 Δy_{ik} 的高斯变量。对于固定的 b_k，第 k 家各数据同时出现的似然函数与

$$\prod_{i=1}^{n_k}\exp\left\{-\frac{1}{2}(\Delta y_{ik})^{-2}[y_{ik}-b_k-S(x_{ik})]^2\right\} \tag{11}$$

成正比。

但是，因为 b_k 也是一个高斯变量，所以这些点混合出现的似然函数的对数去掉一个常数就是

$$F=-\frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\Delta y_{ik})^{-2}\left\{y_{ik}-b_k-S(x_{ik})\right\}^2-\frac{1}{2}\sum_{k=1}^{K}(b_k^2/\Sigma_k^2)\ . \tag{12}$$

令

$$\frac{\partial F}{\partial b_k}=0\ , \tag{13}$$

得

$$b_k=\sum_{i=1}^{n_k}(\Delta y_{ik})^{-2}\left\{y_{ik}-S(x_{ik})\right\}\cdot\left[\sum_{i=1}^{n_k}(\Delta y_{ik})^{-2}+\Sigma_k^{-2}\right]^{-1}\ . \tag{14}$$

将 (14) 式代入 (12) 式，得

$$F'=-\frac{1}{2}\sum_{k=1}^{K}\left[\sum_{i=1}^{n_k}(\Delta y_{ik})^{-2}\left\{y_{ik}-S(x_{ik})\right\}^2-\right.$$
$$\left.-\left\{\sum_{i=1}^{n_k}(\Delta y_{ik})^{-2}[y_{ik}-S(x_{ik})]\right\}^2\cdot\left\{\sum_{i=1}^{n_k}(\Delta y_{ik})^{-2}+\Sigma_k^{-2}\right\}^{-1}\right]\ . \tag{15}$$

令

$$\frac{\partial F'}{\partial C_t}=0\ ,\qquad t=1,2,\cdots N-4 \tag{16}$$

$S(x)$ 用 (7) 式代入，得

$$\sum_{k=1}^{K}\sum_{l=1}^{N-4}C_l\left[\frac{\left\{\sum_{i=1}^{n_k}(\Delta y_{i,k})^{-2}M_l(x_{i,k})\right\}\cdot\left\{\sum_{i=1}^{n_k}(\Delta y_{i,k})^{-2}M_t(x_{i,k})\right\}}{\sum_{i=1}^{n_k}(\Delta y_{i,k})^{-2}+\Sigma_k^2}\right.$$

$$\left.-\sum_{i=1}^{n_k}(\Delta y_{i,k})^{-2}M_l(x_{i,k})\,M_t(x_{i,k})\right]$$

$$=\sum_{k=1}^{K}\left[\frac{\left\{\sum_{i=1}^{n_k}(\Delta y_{i,k})^{-2}y_{i,k}\right\}\cdot\left\{\sum_{i=1}^{n_k}(\Delta y_{i,k})^{-2}M_t(x_{i,k})\right\}}{\sum_{i=1}^{n_k}(\Delta y_{i,k})^{-2}+\Sigma_k^2}\right.$$

$$\left.-\sum_{i=1}^{n_k}(\Delta y_{i,k})^{-2}y_{i,k}M_t(x_{i,k})\right],$$

$$t=1,2,\cdots N-4 \tag{17}$$

(17) 式是一个关于C_l的$N-4$个线性方程组成的方程组。解 (17) 式得$C_l(l=1,2,\cdots N-4)$，代入 (7) 式得拟合曲线的函数$S(x)$。

Σ_k是未知的系统误差，它使曲线移动一常量b_k，但一般系统误差是使曲线按某一百分数移动，所以 (17) 式有待改进。

如果参加计算的只有一家数据，或者几家数据$\Sigma_k=0$，(17) 式同 (13) 式，所以 (13) 式可看成是 (17) 式的特殊情况。

四、拟合误差和最佳拟合的判据

把公式(10)和公式(17)写成如下形式：

$$\boldsymbol{AC}=\boldsymbol{D}. \tag{18}$$

其中$\boldsymbol{A}$为$N-4$阶方矩阵，$\boldsymbol{C}$、$\boldsymbol{D}$为$N-4$列的列矩阵。这样，在特殊点x'的拟合方差为

$$\sigma^2(x')=\boldsymbol{u}^T\boldsymbol{A}^{-1}\boldsymbol{u}. \tag{19}$$

其中$\boldsymbol{A}^{-1}$为(18)式中矩阵$\boldsymbol{A}$的逆矩阵，也叫误差矩阵。

$$\boldsymbol{u}^T=\left(\frac{\partial S(x')}{\partial C_1}\quad\frac{\partial S(x')}{\partial C_2}\cdots\cdots\frac{\partial S(x')}{\partial C_{N-4}}\right) \tag{20}$$

拟合程度的提高是通过选择节点实现的，通过选择节点，使

$$\frac{\sum_{i=1}^{M}\left[y_i-S(x_i)\right]^2(\Delta y_i)^{-2}}{M-N+4} \tag{21}$$

或

$$\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\Delta y_{i,k})^{-2}\left[y_{i,k}-b_k-S(x_{i,k})\right]^2}{\sum_{k=1}^{K}n_k-N+4} \tag{22}$$

达到极小值，这时即为最佳拟合。但是，因为用机器自动调节点的问题尚未解决，所以要找(21)、(22)式的极小值也是不容易的。尽管如此，计算这些量对于人工调节点还是可以提供一些依据的。

五、具体例子

曲线拟合程度如何，与合理选择节点的位置和个数关系很大。节点一般取在曲线变化复杂的地方。如果节点取得太少，逼近程度不好，应有的峰或谷往往被拉平；如果节点取得过多，又可能出现不应有的峰或谷。

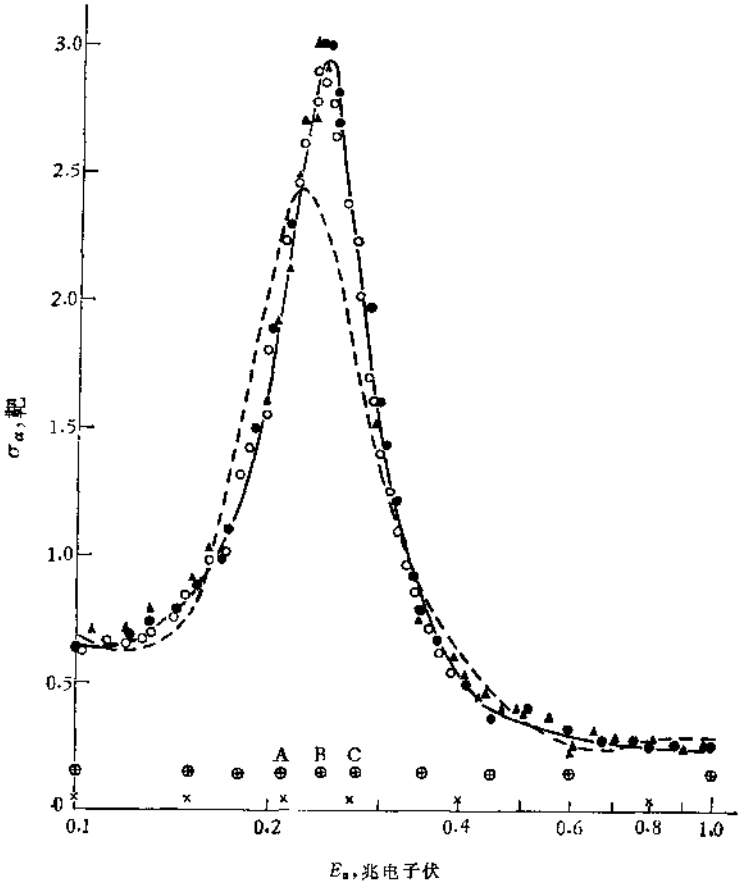

图2　节点选择对拟合程度的影响

▲○●——^{6}Li (n,α) 截面数据；＋×——节点选择位置。

我们用第二节中的方法作了^{239}Pu在10千电子伏—20兆电子伏能区裂变截面的拟合曲线[3]，也作了^{6}Li在1—1700千电子伏能区（n,α）截面和全截面的拟合曲线[4]，均拟合得较好。

在图2上举例说明了节点选择对拟合程度的影响。由于$^{6}Li(n,\alpha)$反应在100—400千电子伏能区有共振，所以对节点的位置反应很灵敏。图中实线是以“+”为节点的拟合曲线。节点B取在峰值的横座标位置，节点A和C取在与B有适当距离的左右两侧，这样做拟合较好。若A与B和B与C之间的距离过远或过近，拟合曲线的峰值将会不适当地下降或上升。图中虚线是以“×”为节点的拟合曲线，节点只取在峰值横座标两侧，拟合得不好。

对^{6}Li（n,α）截面数据拟合的情况说明，因两节点间是以三次多项式逼近的，曲线走向不能太复杂，因此节点应取在曲线走向有突变的位置，使两节点间的曲线走向较平滑，这样可以得到较好的拟合。

总之，节点的选择是很重要的，在这方面我们还缺乏经验，有待进一步探索。

参 考 文 献

[1]孙家昶，计算机应用与应用数学，6，1（1974）。

[2]J.B.Ahorsley et al., *Nucl. Instrum. Methods*, 62, 29 (1968)。

[3]刘继才，1千电子伏—20兆电子伏^{239}Pu的裂变截面评价，hsj-75005/hyl-75.10，核数据编辑组，1976年。

[4]周德林，1.7兆电子伏以下^{6}Li核中子数据评价，hsj-75006/hyl-75.10，核数据编辑组，1976年。