国防科技大学学报
JOURNAL OF NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY
第16卷第3期 1994年9月 Vol. 16 No. 3

# 数据处理的实用方法和算法*

王正明

(国防科技大学系统工程与数学系 长沙 410073)

**摘 要** 通过对正交多项式的构造和逼近准则的研究，得到了一类数据处理问题的一种较理想的逼近准则和算法。理论分析和模拟计算表明，用本文方法处理这类问题有较高的拟合精度并容易计算。

**关键词** 数据处理，逼近准则，正交多项式，正交矩阵

**分类号** O241

本文考虑模型：

$$\begin{cases} y_i = f(x_i) + e_i, \quad i = 0,1,\cdots,m \\ -1 = x_0 < x_1 < \cdots < x_m = 1, \quad \mathrm{E}e_ie_j = \sigma^2\delta_{ij} \\ \int_{-1}^{1} |f''(x)|^2 \sqrt{1-x^2}dx \leqslant M \end{cases} \tag{1}$$

其中：$(x_i,y_i)(i=0,1,\cdots,m)$为观测数据，$f(x)$为被观测的真实函数，$\{e_i\}$为观测误差，$M$为已知正数。

本文的工作是寻找$n$次代数多项式$P(x)$作为$f(x)$的拟合函数。本文通过对切贝雪夫多项式的结构和逼近准则的研究，得到了模型(1)中$f(x)$的一种理想的拟合方法和算法；解决了估计$\sigma^2$和确定$n$的问题；当观测数据中有少量野值时，用本文方法仍能得到较好的效果。

## 1 正交多项式与光滑函数

代数多项式是逼近光滑函数的有力工具。由文[1]知，只要取$n$适当大，用$S_nf$可逼近光滑函数$f(x)$到足够高的精度。其中

$$S_nf = \sum_{i=0}^{n} a_kt_k(x), \quad a_0 = \frac{1}{\pi}\int_0^{\pi} f(\cos\theta)d\theta, \quad a_k = \frac{2}{\pi}\int_0^{\pi} f(\cos\theta)\cos k\theta d\theta, \quad (k \geqslant 1),$$

$$t_0(x) = 1,\ t_1(x) = x,\ t_k(x) = 2xt_{k-1}(x) - t_{k-2}(x),\ (k \geqslant 2)$$

数据处理问题只知道$f(x)$在区间$[-1,1]$中有限个点的观测值，要准确地得到$S_nf$

* 1993年5月25日收稿

是不可能的。本文旨在寻找尽可能接近 $S_nf$ 的多项式 $P(x)$作为 $f(x)$的拟合函数。为此，我们首先来分析系数 $a_i(i=0,1,\cdots,n)$的特点。

**引理 1** 设 $f^{(3)}(x)\in C[-1,1]$

$$f(\cos\theta)=\sum_{i=0}^{\infty}a_i\cos i\theta,\quad f''(\cos\theta)\sin\theta=\sum_{i=0}^{\infty}b_i\sin i\theta$$

则

$$\sum_{i=0}^{\infty}b_i^2=\int_{-1}^{1}|f''(x)|^2\sqrt{1-x^2}\mathrm{d}x$$

$$\frac{i}{i-1}b_{i-1}-\frac{i}{i+1}b_{i+1}=2i^2a_i,\quad i=2,3,\cdots,n.$$

根据对称矩阵的特征根的性质，可得：

**引理 2** 设 $q\leqslant 90$，$\lambda_G$ 为矩阵 $G'G$ 的最大特征根，

$$G=\begin{pmatrix}2&0&-\frac{2}{3}&0&\cdots&0&0&0\\0&\frac{3}{2}&0&-\frac{3}{4}&\cdots&0&0&0\\\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots\\0&0&0&0&\cdots&\frac{q}{q-1}&0&-\frac{q}{q+1}\end{pmatrix}$$

则 $\lambda_G\leqslant 4.837$.

**证明** 取 $q=90$，直接在 VAX 机上算得此时 $\lambda_G=4.837$，而由 $G$ 的定义和最大特征根 $\lambda_G$ 的性质可知，当 $q<90$ 时，$\lambda_G<4.837$.（证毕）

综合上述引理，我们有

**定理 1** 设 $f^{(2)}(x)\in C[-1,1]$，$q\leqslant 90$，则

$$\sum_{i=1}^{q}i^4a_i^2\leqslant 1.21\int_{-1}^{1}|f''(x)|^2\sqrt{1-x^2}\mathrm{d}x \tag{2}$$

由函数逼近论的有关结论可知，若 $M<10^4$，则取 $n<90$ 就能保证 $P(x)$逼近 $f(x)$到很高的精度。以下设 $n<90$.

## 2 逼近准则和算法

我们的思想是寻找 $P(x)$尽可能接近 $S_nf$，也即我们要设法得到 $a_i(i=0,1,\cdots,n)$的较准确的估计。这需要建立适当的准则。由(1)知：

$$1.21\int_{-1}^{1}|f''(x)|^2\sqrt{1-x^2}\mathrm{d}x\leqslant\alpha^2\equiv 1.21\mathrm{M} \tag{3}$$

为方便，下记

$$a=(a_0,a_1,\cdots,a_n)',\quad T_i=(t_i(x_0),t_i(x_1),\cdots,t_i(x_m))'$$

$$P(x)=\sum_{i=0}^{n}\tilde{a}_it_i(x),\quad T=(T_0,T_1,\cdots,T_n)$$

$$a_P=(a_0,a_1)',\ a_R=(a_2,a_3,\cdots,a_n)'$$

$$T_P=(T_0,T_1),\quad T_R=(T_2,T_3,\cdots,T_n)$$

$$T_{PR}=T_P(T_P'T_P)^{-1}T_P'T_R,\quad T_{RR}=T_R-T_{PR}$$

$$\hat{a}_P = (T_P'T_P)^{-1}T_P'Y$$

$$\tilde{a} = \begin{pmatrix}\tilde{a}_P\\ \tilde{a}_R\end{pmatrix} = (\tilde{a}_0,\tilde{a}_1,\cdots,\tilde{a}_n)'; \quad \hat{\beta} = \begin{pmatrix}\hat{\beta}_P\\ \hat{\beta}_R\end{pmatrix} = (T'T)^{-1}T'Y$$

根据定理 1，我们建立如下准则确定 $\tilde{a}$：

$$\begin{cases}\sum_{i=2}^{n} i^4 a_i^2 \leqslant \alpha^2,\\ \| Y - Ta \|^2 = \min\end{cases} \tag{4}$$

直接从极值问题(4)中求 $\tilde{a}$ 是困难的，下面我们将问题逐步简化。

**引理 3** 设 $\tilde{a}$ 是(4)的解，$\hat{a}_R$ 是

$$\begin{cases}\Sigma i^4 a_i^2 \leqslant \alpha^2\\ \| T_{RR}(\hat{\beta}_R - a_R) \|^2 = \min\end{cases} \tag{5}$$

的解，则

$$\begin{cases}\tilde{a}_P = \hat{a}_P - (T_P'T_P)^{-1}T_P'T_R\hat{a}_R\\ \tilde{a}_R = \hat{a}_R\end{cases} \tag{6}$$

**证明** 注意到：

$$T'(Y-T\hat{\beta})=0, \quad T_P'T_{RR}=0,$$
$$T\hat{\beta}=T_P\hat{\beta}_P+T_{PR}\hat{\beta}_R+T_{RR}\hat{\beta}_R$$

于是，(4)可化为如下的等价形式：

$$\begin{cases}\sum_{i=2}^{n} i^4 a_i^2 \leqslant \alpha^2\\ \| T_P(\hat{\beta}_P - a_P) + T_{PR}(\hat{\beta}_R - a_R) + T_{RR}(\hat{\beta}_R - a_R) \|^2 = \min\end{cases} \tag{7}$$

显然，(7)等价于

$$\begin{cases}\| T_P(\hat{\beta}_P - a_P) + T_{PR}(\hat{\beta}_R - a_R) \|^2 = 0,\\ \sum_{i=2}^{n} i^4 a_i^2 \leqslant \alpha^2\\ \| T_{RR}(\hat{\beta}_R - a_R) \|^2 = \min\end{cases} \tag{8}$$

另一方面，设 $\hat{a}_R$ 为(5)的解，$\tilde{a}$ 由(6)式确定，则显然 $\tilde{a}$ 是(8)的解。综合上述即得引理。

求解(5)仍然有困难，下面我们将(5)进一步简化，记：

$$\Gamma = \mathrm{diag}(2^2,3^2,\cdots,n^2);$$
$$A = \Gamma^{-1}T_{RR}'T_{RR}\Gamma^{-1};$$
$$A = R \wedge R', \ R'R = \mathrm{diag}(1,1,\cdots,1);$$
$$\Lambda = \mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_{n-1});$$
$$V = T_{RR}\Gamma^{-1}R,$$
$$\hat{\gamma} = (\hat{\gamma}_1,\hat{\gamma}_2,\cdots,\hat{\gamma}_{n-1})' = R'\Gamma\hat{\beta}_R;$$
$$\gamma = (\gamma_1,\gamma_2,\cdots,\gamma_{n-1})' = R'\Gamma a_R.$$

在上述记号下，我们有：

**引理 4** 设 $\tilde{\gamma}$ 是极值问题：

$$\begin{cases} \| \gamma \|^2 \leqslant \alpha^2 \\ \sum_{i=1}^{n-1} \lambda_i (\gamma_i - \hat{\gamma}_i)^2 = \min \end{cases} \tag{9}$$

的解，$\hat{a}_R$ 是极值问题(5)的解，则 $\hat{a}_R = \Gamma^{-1} R \tilde{\gamma}$.

**证明** 注意到：

$$\sum_{i=2}^{n} i^4 a_i^2 = \| \gamma \|^2$$

$$\| T_{RR}(\hat{\beta}_R - a_R) \|^2 = \sum_{i=1}^{n-1} \lambda_i (\gamma_i - \hat{\gamma}_i)^2$$

于是，由以上两式及引理 3 即得引理。

下面，我们来求解极值问题(9).

**引理 5** 设 $\tilde{\gamma}$ 是极值问题(9)的解，则

$$\tilde{\gamma}_i = \lambda_i \hat{\gamma}_i / (\lambda^0 + \lambda_i), \quad i = 1, 2, \cdots, n-1 \tag{10}$$

当 $\| \hat{\gamma} \| \leqslant \alpha$ 时，$\lambda^0 = 0$；当 $\| \hat{\gamma} \| > \alpha$ 时，$\lambda^0$ 是方程

$$\sum_{i=1}^{n-1} \left( \frac{\lambda_i \hat{\gamma}_i}{\lambda + \lambda_i} \right)^2 = \alpha^2$$

在区间 $(0, \lambda_{n-1} \| \hat{\gamma} \| \alpha^{-1})$ 中的唯一解。

**证明** 不妨设 $\| \hat{\gamma} \| > \alpha$，记

$$\mathcal{T} = \{ \gamma \mid \| \gamma \| \leqslant \alpha \}, \quad D(\gamma) = \sum_{i=1}^{n-1} \lambda_i (\gamma_i - \hat{\gamma}_i)^2$$

$$F(\gamma) = D(\gamma) + \lambda ( \| \gamma \|^2 - \alpha^2 + \tau^2 )$$

由松驰变量法得：

$$\begin{cases} \lambda \tau = 0, \quad \gamma_i = \dfrac{\lambda_i \hat{\gamma}_i}{\lambda + \lambda_i} \\ f(\lambda) \equiv \sum_{i=1}^{n-1} \left( \dfrac{\lambda_i \hat{\gamma}_i}{\lambda + \lambda_i} \right)^2 - \alpha^2 = 0 \end{cases}$$

因 $f(0) > 0$，故 $\tau = 0$. 下证 $\lambda^0 > 0$. 若不然，设 $\lambda^* < 0$，

$$\gamma_i^* = \frac{\lambda_i \hat{\gamma}_i}{\lambda^* + \lambda_i}, \quad i = 1, 2, \cdots, n-1$$

$\gamma^*$ 为 $D(\gamma)$ 在 $\mathcal{T}$ 中的最小值点，取

$$\gamma_i^{**} = \frac{\lambda_i \hat{\gamma}_i}{\lambda_i - \lambda^*}, \quad i = 1, 2, \cdots, n-1$$

则显然有：

$$\| \gamma^{**} \| < \alpha, \quad D(\gamma^{**}) < D(\gamma^*)$$

这与 $\gamma^*$ 为 $D(\gamma)$ 在 $\mathcal{T}$ 中的最小值矛盾，故 $\lambda^0 > 0$. 另一方面

$$f'(\lambda) < 0, \quad \forall \lambda > 0,$$

$$f(0) > 0, \quad f(\lambda_{n-1} \| \hat{\gamma} \| \alpha^{-1}) < 0$$

综合上述即得引理。

在实算时，可用二分法求得 $\lambda^0$。综合上述引理，我们有：

**定理 3** 设 $\tilde{a}$ 是(4)的解，$\tilde{\gamma}$ 由(10)式给出，则

$$\begin{cases}\tilde{a}_R = \Gamma^{-1}R\tilde{\gamma} \\ \tilde{a}_P = \hat{a}_P - (T_P'T_P)^{-1}T_P'T_R\tilde{a}_R\end{cases}$$

应用本文方法时，$M-\int_{-1}^{1}|f''(x)|^2\sqrt{1-x^2}\,dx$ 愈小愈好。$n$ 的大小与 $M$ 有关；当 $M<10^4$ 时，取 $n<90$，实用时 $n$ 可取得稍大些。准则(4)保证了不会出现 Runge 现象。用本文方法的明显的好处是不用估计 $\sigma^2$ 和确定 $n$，而这正是多项式拟合的难点。用本文方法的另一优点是，观测误差可以是色噪声；并且当观测数据中存在少量野值时，用 $a$ 的稳健最小二乘估计 $\beta^M$ 取代 $\hat{\beta}$ 后，仍可得到较满意的效果。

## 3 数值例子

以下的数值例子较好地说明了本文方法的优点，该例在 VAX/VMS 上实现计算。

**例 1** 我们用模型(1)进行模拟计算，其中：$x_i=0.01i-1$，$i=0,1,\cdots,200$；$e_i\sim N(0,0.01)$，$Ee_ie_j=0$；$f(x)=\sin x+e^x+(1+5x^2)^{-1}$

我们取：

$n=31$，　$M=20$；　$y(x_i)=f(x_i)+e_i$，　$i=0,1,\cdots,200$；

$$P(x)=\sum_{i=0}^{31}\tilde{a}_it_i(x)；\quad \hat{P}(x)=\sum_{i=0}^{31}\hat{\beta}_it_i(x)$$

计算结果见表 1. 从表 1 中可看到，用 $P(x)$拟合 $f(x)$有较好的效果。

**表 1**

| $x_i$ | $y(x_i)$ | $f(x_i)$ | $P(x_i)$ | $\hat{P}(x_i)$ |
|---|---|---|---|---|
| −1.0 | −0.3439 | −0.3069 | −0.3130 | −0.3348 |
| −0.9 | −0.0425 | −0.1787 | −0.1482 | −0.1550 |
| −0.8 | 0.0511 | −0.0299 | −0.0040 | −0.0293 |
| −0.7 | 0.1343 | 0.1422 | 0.1324 | 0.1549 |
| −0.6 | 0.2791 | 0.3413 | 0.3196 | 0.2915 |
| −0.5 | 0.6884 | 0.5715 | 0.5644 | 0.5577 |
| −0.4 | 0.8128 | 0.8365 | 0.8601 | 0.8806 |
| −0.3 | 1.1149 | 1.1350 | 1.1575 | 1.1330 |
| −0.2 | 1.3362 | 1.4534 | 1.4702 | 1.4850 |
| −0.1 | 1.6747 | 1.7574 | 1.7608 | 1.7467 |
| 0.0 | 2.1625 | 2.0000 | 1.9992 | 2.0222 |
| 0.1 | 2.1906 | 2.1574 | 2.1506 | 2.1481 |
| 0.2 | 2.0855 | 2.2534 | 2.2443 | 2.2433 |
| 0.3 | 2.2229 | 2.3350 | 2.3308 | 2.3207 |
| 0.4 | 2.4459 | 2.4368 | 2.4409 | 2.4498 |
| 0.5 | 2.5895 | 2.5726 | 2.5757 | 2.5648 |
| 0.6 | 2.6929 | 2.7439 | 2.7544 | 2.7640 |
| 0.7 | 2.8554 | 2.9478 | 2.9769 | 2.9627 |
| 0.8 | 3.2646 | 3.1810 | 3.1952 | 3.1998 |
| 0.9 | 3.6430 | 3.4409 | 3.4430 | 3.4640 |
| 1.0 | 3.7911 | 3.7264 | 3.7441 | 3.7837 |

## 参 考 文 献

1 E.W.切尼. 逼近论导引. 上海科技出版社，1981
2 黄俊钦，刘整社. 多项式回归的快速算法. 应用数学学报，1986，(2)
3 Б.Ф.日丹纽克. 无线电外弹道测量结果统计处理基础. 北京：宇航出版社，1987
4 G.G.洛伦茨. 函数逼近论. 上海科技出版社，1981
5 程正兴. 数据拟合. 西安交通大学出版社，1986
6 王正明. 回归系数的改进主成分估计. 数学的实践与认识，1990(1)


# A Practical Method and Its Algorithms of Data Processing


Wang Zhengming

(Department of System Engineering and Applied Mathematics)



**Abstract**

In this paper the approximation rules and the structures of orthogonal polynomials have been studied and a practical method and its algorithms of a class of data processing problems have been obtained. In dealing with the above problems. it is easier to calculate by our methods and with more accurate results at that. The advantages of our methods are ittustrated by theory analysis and simulations.



**Key words** data processing, approximation rules, orthogonal polynomial, orthogonal matrice