文章编号：1001-2486（2001）06-0042-06

弹道跟踪数据处理中的几个计算问题*

王正明，周海银，童丽，朱炬波，段晓君

（国防科技大学理学院，湖南 长沙 410073）

摘 要：结合弹道跟踪数据建模和数据处理的实践，提出了几个有典型应用价值的测量数据模型，在应用这些数学模型解决实际问题时，涉及统计、优化、函数逼近等几方面的理论和计算问题。这些问题和模型的研究，对解决许多相关的实际问题有重要意义。

关键词：弹道跟踪数据处理；异常数据识别；自变量选择；双正交基；试验鉴定

中图分类号：V421.4^{+}1 **文献标识码**：A

Several Calculating Problems of the Processing of Trajectory Tracking Data

WANG Zheng-ming, ZHOU Hai-yin, TONG Li, ZHU Ju-bo, DUAN Xiao-jun

(College of Science, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Based on our research and exploration in the field of data processing and modeling of trajectory tracking data, this paper puts forward several observing data models with typical application. These models are relative to many aspects of statistics, optimization, functional approximation in theory and calculation. The investigation of these problems and models is significant to the solution of many relevant practical problems.

Key words: trajectory tracking data processing; outlier identification; variable selection; double orthogonal basis; test and evaluation

弹道跟踪数据是分析和评估运载火箭飞行试验是否成功的重要依据，也是改进火箭控制系统设计、提高制导精度的重要依据。而弹道跟踪数据的精度则是保证弹道跟踪数据应用效果的关键。提高弹道跟踪数据的精度，一是从硬件设备上想办法；二是从数据处理上想方法[1]。

本文是作者多年实践工作的总结[2-9]，重点讨论了弹道跟踪数据处理中几个重要的大计算量问题，其中包括：提高测量数据精度的异常数据（测元）的诊断与识别；建立少参数、高精度参数模型的自变量选择；待估函数的双正交基逼近；基于全程试验的全程鉴定等。这些方法不仅可以联合使用，也可分别使用；不仅适用于各类弹道跟踪数据的处理问题，也适用于数据融合、经济分析、气象预报等领域的数据处理问题。其关系见图 1。

1 异常数据的批量识别问题

假设对 t_i 时刻的 $f(t_i)$ 值进行观测，得到测量数据 $(t_i, y_i)(i = 1,2,\cdots,M)$，设其模型为：

$$y_i = f(t_i) + e_i, \qquad i = 1,2,\cdots,M$$

其中 $Ee_i = 0$，$Ee_ie_j = \sigma^2\delta_{ij}$。测量数据处理的主要任务，就是要估计 $f(t_i)$。应用函数逼近方法可以将 $f(t)$ 表示为 $f(t) = \sum_{j=1}^{N}\beta_j\varphi_j(t)$，其中，$(\varphi_1,\varphi_2,\cdots,\varphi_N)$ 是一组(已知) 基函数，$\beta = (\beta_1,\beta_2,\cdots,\beta_N)^{\mathrm{T}}$ 是待估参数。记 $X = (x_{ij})_{M\cdot N}$，$x_{ij} = \varphi_j(t_i)$，$Y = (y_1,y_2,\cdots,y_M)^{\mathrm{T}}$，于是

$$Y = X\beta + e,\ e \sim (0,\sigma^2 I) \tag{1}$$

* 收稿日期：2001-04-01
基金项目：国家自然科学基金（69872039）和教育部高等院校骨干教师基金资助
作者简介：王正明（1962-），男，教授，博士。

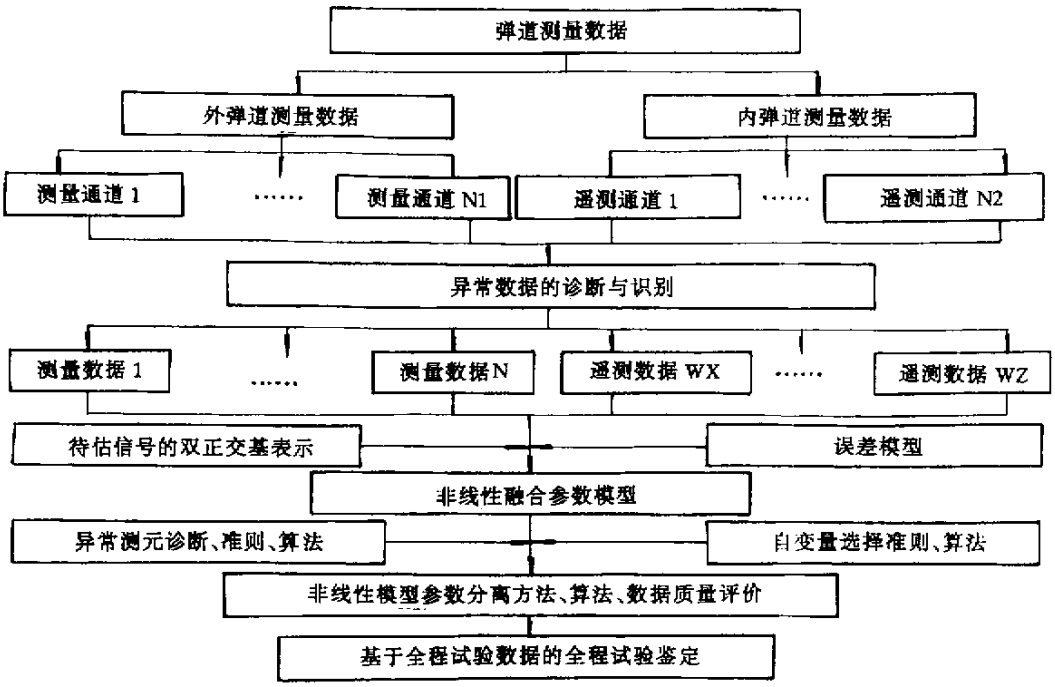

图 1 弹道跟踪数据处理中大计算量问题的关系

Fig.1 The relation among several calculating problems in the processing of trajectory tracking data.

根据最小二乘估计方法，可以得到线性模型（1.1）中参数 β 的估计 β_{LS}。但近代稳健统计学的研究表明，β_{LS} 的影响函数是一个无界函数，崩溃点 $\varepsilon^*(\beta_{LS}) = 0$。这就意味着，在应用 LS 估计时，观测数据中不能含有粗大误差，否则结果是不可靠的。但工程实践告诉我们，粗大误差是不可避免的。因此，在求 β 的估计以前，异常观测数据的识别与剔除是必需的，而且是十分重要的。

1.1 异常数据的识别与剔除

对于异常数据单个出现的情况，逐点剔除法的处理结果比较好[1]；但对于异常数据在某一时段连续出现或某一测量通道异常（即异常测元）的情况，逐点剔除法的处理效果就不够理想。在工程实践中，这种情况常有发生，例如测量设备、传输设备、接收设备中的任何一个在某一时间段内出现异常，就会有连续的异常数据出现。对于存在异常数据的观测数据来说，能一次识别并剔除所有的异常数据（测元），其处理效果是最好的。因此，构造异常数据批量识别的统计量，寻找计算量较小的方法来达到一次识别并剔除多个（最好是全部）异常数据的目的，是值得研究的问题。

对于异常数据的批量识别与剔除，采用如下方法：考虑在事先知道最多异常数据（测元）个数的情况下，首先应用逐点剔除法寻找明显的异常数据；然后利用并行计算，基于残差平方和最小的原则，采用全子集组合回归的方法寻找成片异常数据或异常测元，可以得到较好的效果。其具体实现见参考文献 [5]。并且，从文 [5] 算例可知，逐点剔除法只能寻找出其中一些明显的异常数据，应用一定范围内的全子集回归方法能找出全部的异常数据，估计结果有明显的改善，采用并行计算的时间较之串行计算的时间有很大程度的降低。

这个方法对于解决异常数据的批量识别的困难在于：

(1) 全子集回归的计算量非常大，因此即使采用并行算法，它的计算量也是相当大的；

(2) 采用基于残差平方和最小的原则，没有构造出一个异常数据批量识别的统计量。

1.2 非线性融合模型中测元遴选的问题

对于多个测元融合的问题，首先要进行异常测元的诊断和测元遴选的工作[6]。

设有 K 个测元,每个测元进行了 M 次采样,N 为待估参数个数。假设用上全部测元，得到的非线性模型为

$$Y = F(\beta) + e, \quad e \sim (0,1) \tag{2}$$

而去掉一些测元后得到的模型为

$$Y_p = F_p(\beta_p) + e_p, \quad e_p \sim (0,1) \tag{3}$$

这里 Y_p, F_p, β_p, e_p 分别表示由 Y, F, β, e 的一部分分量构成的向量。

对于全模型(2)式,计算 $Q = \dfrac{\| Y - F(\hat{\beta}) \|^2}{KM - N}$ 然后,重复对模型(2)式每次去掉一个不同的测元。设共去掉了 l 个测元,待估参数个数变为 N_p,则得到相应的选模型(3)式,记 $Q_p = \dfrac{\| Y_p - F_p(\hat{\beta}_p) \|^2}{M(K-l) - N_p}$,由文献[1]可知,$Q$ 和 Q_p 均应接近于1。若去掉一个测元后,得到的 Q_B 使 $Q_B = \min Q_p$,且 $Q_B > 0.95Q$,那么应当使用全模型（2）式；否则使用选模型

$$Y_B = F_B(\beta_B) + e_B, \quad e_B \sim (0, I_B) \tag{4}$$

把模型（4）式当成模型（2）式，$l - 1 \Rightarrow l$ ，重复上述过程。

在测元遴选的过程中，同样存在上述的大计算量问题。

2 回归模型的自变量选择

在应用回归分析方法解决动态测量数据处理的实际问题时，对模型进行优选是必要的。应用 $X\beta = x_1\beta_1 + x_2\beta_2 + \cdots + x_n\beta_n$ 来近似 f,若 $x_1\beta_1$ 很接近于 0,那么剔除 $x_i\beta_i$ 无论对于估准 f 还是估准回归系数都是有益的[1]。自变量选择就是找出并剔除一部分 $x_i\beta_i$ ，使模型中待估参数的个数减少，从而提高参数估计的精度。

2.1 线性模型的自变量选择

考虑含 N 个自变量,M 个观测数据的线性回归模型

$$Y = X\beta + e, \quad e \sim (0, \sigma^2 I) \tag{5}$$

设全模型和最优选模型为

$$Y = X\beta + e = X_p\beta_p + X_R\beta_R + e, \quad e \sim (0, \sigma^2 I)$$

$$Y = X_p\beta_p + \tilde{e}, \quad \tilde{e} \sim (X_R\beta_R, \sigma^2 I)$$

$N = p + R$ 为全模型待估参数个数。定义 $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$,$\hat{\sigma}^2 = \dfrac{\| Y - X\hat{\beta} \|^2}{M - N}$,$\hat{\beta}_p = (X_p^{\mathrm{T}}X_p)^{-1}X_p^{\mathrm{T}}Y$,$C_p = \dfrac{\| Y - X_p\hat{\beta}_p \|^2}{\hat{\sigma}^2} + 2p - M$,那么从全模型中确定最优选模型的自变量选择准则[1]为：

$$\begin{cases} C_p \leqslant p + 1 \\ C_p = \min \end{cases}$$

2.2 线性模型自变量选择的并行算法

为了降低自变量选择的计算量，线性模型自变量选择一般分为三步进行[1]：

(1) 剔除多余的自变量 p_0；

(2) 确定必选变量 p_1；

(3) 将去掉必剔和必选变量后的模型看成是全模型,此时全模型的自变量个数为 $N - p_0 - p_1$,然后对所有可能的子集回归。

因此,从全模型中选择最优选模型的问题等价于从全模型(5)中寻找 X_p 使

$$\frac{\| \widetilde{X} - X_p(X_p^{\mathrm{T}}X_p)^{-1}X_p^{\mathrm{T}}\widetilde{Y} \|^2}{\hat{\sigma}^2} + 2p_1 + 2p_2 - M = \min$$

这一问题在本质上归结为 $2^{N-p_0-p_1}$ 个残差平方和的计算问题，这显然是一个计算量很大的工作。因此对这一过程的实现采取了快速并行算法。

快速算法：将求 $(X_p^{\mathrm{T}}X_p)^{-1}$ 和 $\| \widetilde{Y} - X_p(X_p^{\mathrm{T}}X_p)^{-1}X_p^{\mathrm{T}}\widetilde{Y} \|^2$ 用扫描运算代替，设计出节省计算量的字典序程序，具体的实现参见文献［1］和［6］：

并行算法：在快速算法的基础上，将 $2^{N-p_0-p_1}$ 个子集回归的计算用并行算法实现，设计出分布式并行算法，具体的实现参见文献［7］。

2.3 非线性模型自变量选择

在工程实践中，存在大量的非线性模型，它们也存在一个模型选择的问题。目前常用的方法是采用叠代格式近似成线性模型，然后再进行自变量选择，这样处理的问题在于：

(1) 线性模型的近似带来非线性模型选择的不准确；

(2) 用线性模型的叠代近似非线性模型，会使计算量进一步增大。

因此，值得研究的问题是：(1) 如何构造非线性模型自变量选择的统计量；(2) 如何处理非线性模型自变量选择的大计算量问题；(3) 叠代初值的选取，叠代格式的收敛性、稳定性的判定准则。

3 待估函数的双正交基逼近问题

3.1 频率的定义

为提高信噪分离的精度，文［8］运用按离散内积正交的样条定义一种贴近信号特征和先验信息的频率，使得在此定义下信号和噪声的频率特征差别较大，且信号集中在低频，以便于区分高低频段和实施计算。连续函数的频率定义方法如下。

定义 1 设 $f(t) \in C^{K-1}[t_1,t_m]$，$f^{(K)}(t)$ 分段连续并在 $[t_1,t_m]$ 上平方可积，记分划 T ：

$$[t_1,t_m] = \bigcup_{k=1}^{M}[T_k,T_{k+1}],t_1 = T_1 < T_2 < \cdots < T_{M+1} = t_m$$

定义函数 $f(t)$ 的频率为

$$F(f,T) = \frac{\sum_{k=1}^{M}\left\{\int_{T_k}^{T_{k+1}}\left| f^{(K)}(t) - \frac{[f^{(K-1)}(T_{k+1}) - f^{(K-1)}(T_k)]}{T_{k+1} - T_k}\right|^2 \mathrm{d}t\right\}}{\int_{t_1}^{t_m} | f(t) |^2 \mathrm{d}t}$$

3.2 双正交基的构造

针对实际工程背景中航天测量信号的主趋势有三阶多项式特征及特征点，且拟合时有保精度要求，故依此构造一组规范的双正交基，使得这些基的频率从小到大排列，而且信号由低频基的线性组合可以高精度表示，噪声则分布到各高频段上，从而依据基的特点可分离信号和噪声。

以下说明规范正交基的构造。若在 (t_1,t_m) 中有 r 个特征点 $T_1^*,T_2^*,\cdots,T_r^*$，即 $f(t) \in C^2[t_1,t_m] \cap C^4\{[t_1,t_m] \setminus T^*\}$。样条内节点集为 $\{\tau_5,\tau_6,\cdots,\tau_n\} = \{T_1^*,T_2^*,\cdots,T_r^*\} \cup \{\tau_1^*,\tau_2^*,\cdots,\tau_{n-r-4}^*\}$，$\{T_1^*,T_2^*,\cdots,T_r^*\} \cap \{\tau_1^*,\tau_2^*,\cdots,\tau_{n-r-4}^*\} = \varnothing$，使得特征点都是样条节点，且使 $\max\{\max_{5\leqslant i\leqslant n-1} | \tau_{i+1} - \tau_i |,\tau_5 - t_1,t_m - \tau_n\} \triangleq h$ 充分小。记

$$\begin{cases}\psi_1^*(t) = 1, \quad \psi_2^*(t) = t, \quad \psi_3^*(t) = t^2, \quad \psi_4^*(t) = t^3,\\ \psi_j^*(t) = (t - \tau_j)_+^3, \quad j = 5,6,\cdots,n\end{cases}$$

记 $\psi_j^* = (\psi_j^*(t_1),\psi_j^*(t_2),\cdots,\psi_j^*(t_m)),\psi_n^* = (\psi_1^*,\psi_2^*,\cdots,\psi_n^*)$，令 $\psi_1 = \frac{1}{\sqrt{m}}\psi_1^*,\Psi_{j-1} = (\psi_1,\psi_2,\cdots,\psi_{j-1})$，$j \geqslant 2$。第 j 个规范正交基为 $\psi_j = \frac{\psi_j^* - \Psi_{j-1}\Psi_{j-1}^{\mathrm{T}}\psi_j^*}{\| \psi_j^* - \Psi_{j-1}\Psi_{j-1}^{\mathrm{T}}\psi_j^* \|}$。可验证满足规范化要求。记

$\psi_j^{\#}(t_i) \triangleq \psi_j^{*(3)}(t_i) - \frac{1}{N_{k+1} - N_k}\left[\sum_{l=N_k}^{N_{k+1}}\psi_j^{*(3)}(t_l)\right]$，$l \leqslant i \leqslant m$，$\psi_j^{\#} = (\psi_j^{\#}(t_i),\psi_j^{\#}(t_2),\cdots,\psi_j^{\#}(t_m))^{\mathrm{T}}$，$j = 1,2,\cdots,n$，$\Psi_n^{\#} = (\psi_1^{\#},\psi_2^{\#},\cdots,\psi_n^{\#})$，$Q = \Psi_n^{\#\mathrm{T}}\Psi_n^{\#}$，由于 Q 是对称半正定矩阵，于是存在正交矩阵 $P = (P_1,P_2,\cdots,P_n)$ 和对角矩阵 $A = \mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_n)$，$0 < \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$，使得 $Q = P^{\mathrm{T}}\Lambda P$。

定义 2 记 $v_i = \Psi_m P_j = \Psi_n^* L_n P_i,v_i^{\#} = \psi^{\#}P_i,i = 1,2,\cdots,n$，而 λ_i 为 v_i 由定义 1 算得的对应

频率。若 $v_i^{\mathrm{T}}v_j = \begin{cases}1, & i=j\\ 0, & i\neq j\end{cases}$，$v_i^{\#\mathrm{T}}v_j^{\#} = \begin{cases}\lambda_i, & i=j\\ 0, & i\neq j\end{cases}$ $1\leqslant i, j\leqslant n$，则称$(v_1, v_2, \cdots, v_n)$为双正交基。

由规范双正交基的构造，知 v_i 中 i 越小，则 v_i 对应的频率越低。并且可以证明，低频双正交基的线性组合仍为低频。规范双正交基的逼近性质可用 Q 的特征值递增速度来推断。

3.3 分片光滑函数的双正交基逼近

离散观测数据 $Y = f + e$ 可以表示为

$$\sum_{i=1}^{n}(Y^{\mathrm{T}}v_i)v_i = \sum_{i=1}^{n}(f^{\mathrm{T}}v_i)v_i + \sum_{i=1}^{n}(e^{\mathrm{T}}v_i)v_i$$

的形式，v_i 为双正交基。记 $c_i = f^{\mathrm{T}}v_i$，$u_i = e^{\mathrm{T}}v_i$，假设 h 充分小，样条内节点集满足条件，那么

$$F(f,T) = \sum_{i=1}^{n}\lambda_i c_i^2, \qquad \| f - \sum_{i=1}^{M_q} c_i v_i \|^2 = \| \sum_{i=M_k+1}^{n} c_i v_i \|^2 = \sum_{i=M_0+1}^{n} c_i^2$$

而若同时知道 $F(f,T)\leqslant\mu$，则

$$\sum_{i=1}^{n}\lambda_i c_i^2 \leqslant \mu \Rightarrow \Big\| f - \sum_{i=1}^{M_0} c_j v_j \Big\|^2 = \sum_{i=M_0+1}^{n} c_i^2 \leqslant \frac{1}{\lambda_{M_0+1}}(\mu - \sum_{i=1}^{M_k}\lambda_i c_j^2)$$

3.4 双正交基构造中的计算问题

由于在实际测量数据处理问题中，采样节点密，数据量大，双正交基的构造过程中有一个计算量很大的问题存在。希望找到一种方法能够快速计算出特定情况下的双正交基以便应用。

4 试验鉴定中的高维积分问题

4.1 小子样鉴定

由于战略武器试验的特殊性，考虑到经济和政治因素，因此只能采取小子样试验的方式。在试验鉴定过程中，普遍采用的方法均以落点散布特性为检验依据，验前信息也仅局限于试验的落点信息；而对试验过程中的全程弹道信息未予以考虑，在鉴定结论中未得到应有的体现。这种方法的局限主要在于验前信息的确定受主观因素影响太大，先验分布的合理性一直是鉴定结果可信性的争议焦点。可见，传统的鉴定方法有如下的弱点：(1) 验前信息的确定受主观因素影响大；(2) 先验分布的合理性及其鉴定结果一直有争议；(3) 全程试验的大量测量信息搁置未用，是一种很大的信息资源浪费。

总的看来，充分利用全程弹道试验数据资源进行靶场鉴定的关键技术和系统介绍仍然少见；针对试验鉴定的实用且高精度的弹道误差模型及算法尚未建立，关于落点偏差的验前信息常有争议，难以保证试验结论的高精度和高可靠性。实现小子样鉴定有以下一些有利条件：(1) 阵地试验与测试信息较多；(2) 全弹道跟踪，有多套、多站的跟踪数据；(3) 可以对试验测量数据进行建模；这是我们进行小子样鉴定的基础。

4.2 基于全程试验信息的全程鉴定模型

弹道差的位置及速度参量用双正交基表示如下：

$$\begin{cases}\Delta x(t) = \sum_{j=1}^{N} b_j\psi_j(t), & \Delta y(t) = \sum_{j=1}^{N} b_{j+N}\psi_j(t), & \Delta z(t) = \sum_{j=1}^{N} b_{j+2N}\psi_j(t)\\ \Delta\dot{x}(t) = \sum_{j=1}^{N} b_j\dot{\psi}_j(t), & \Delta\dot{y}(t) = \sum_{j=1}^{N} b_{j+N}\dot{\psi}_j(t), & \Delta\dot{z}(t) = \sum_{j=1}^{N} b_{j+2N}\dot{\psi}_j(t)\end{cases} \tag{6}$$

落点纵横向偏差 $(\Delta L, \Delta H)$ 可以表示为：

$$\Delta L = \eta_1 \overset{\mathrm{def}}{=} \phi_L b = \sum_{i=1}^{3M}\phi_{Li}b_i, \qquad \Delta H = \eta_2 \overset{\mathrm{def}}{=} \phi_H b = \sum_{i=1}^{3M}\phi_{Hi}b_i$$

其中 (η_1, η_2) 的密度函数为 $q(y_1, y_2) = \int_{R^{3M-2}} \frac{\bar{h}(y\mid a)\mathrm{d}y_3\mathrm{d}y_4\cdots\mathrm{d}y_{3M}}{\phi_{L1}\phi_{H2} - \phi_{L2}\phi_{H1}}$。其期望值和方差分别为：

$$\mathrm{E}[\Delta L] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y_1 q(y_1, y_2)\mathrm{d}y_1\mathrm{d}y_2, \mathrm{E}[\Delta H] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y_2 q(y_1, y_2)\mathrm{d}y_1\mathrm{d}y_2, \tag{7}$$

$$\begin{cases} \mathrm{Var}(\Delta L) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(y_1 - \mathrm{E}[\Delta L])^2 q(y_1, y_2)\mathrm{d}y_1\mathrm{d}y_2, \\ \mathrm{Var}(\Delta H) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(y_2 - \mathrm{E}[\Delta H])^2 q(y_1, y_2)\mathrm{d}y_1\mathrm{d}y_2 \end{cases} \tag{8}$$

4.3 高维积分的计算问题

在计算落点偏差的分布过程中，如在（7）～（8）式中，存在一个高维积分的计算问题。目前使用的方法是 Monte-Carlo 方法。具体实现方法如下：

考虑多元函数 $f(\xi)(\xi \in R^n)$ 的高维积分计算，积分区域为有界闭区域 D，则用下面的方法进行计算：取一充分大的立方体 $[a,b]^n$，使得 $D \subset [a,b]^n$，但 $D \not\subset [a+10^{-5}, b-10^{-5}]^n$。定义 $f^*(\xi) = \begin{cases} f(\xi), \xi \in D \\ 0, \quad \xi \notin D \end{cases}$，产生 $\xi^{(i)} \sim U[a,b]^n$，于是 $\int_D f(\xi)\mathrm{d}\xi = \int_{[a,b]^n} f^*(\xi)\mathrm{d}\xi \approx \frac{1}{N}(b-a)^n \sum_{i=1}^{N} f^*(\xi^{(i)})$。

仿真计算表明[9]：对于不同的 n，收敛速度是不一样的，当 n 较大（如 $n \geqslant 50$）时，积分的计算比较困难，需要研究更为科学的高维积分的快速和高精度计算方法。

参考文献：

[1] 王正明，易东云，周海银等．弹道跟踪数据的校准与评估［M］．长沙：国防科技大学出版社，1999.

[2] Wang Zhengming，Zhou Haiyin. Mathematical Proceedings of Range and Rang Rate Tracking Data［R］. AD－A310796，1996.

[3] Wang Zhengming，Zhu Jubo. Reduced Parameter Model on Trajectory Tracking Data with Applications［J］. Science in China（Series E），1999（2）.

[4] Wang Zhengming，Zhou Haiyin. A New Method of Estimating System Error of Guidance instrument［J］. Science in China（Series E），1998（2）.

[5] 童丽，曾泳泓，王正明．异常点剔除及其并行算法［J］．数值计算与计算机应用，2000（3）.

[6] Wu Yi，Zhu Jubo. A Fusion Method for Estimate of Trajectory［J］. Science in China，1998（6）.

[7] 童丽，王正明，曾泳泓．自变量选择及其并行计算［J］．数值计算与计算机应用，2001（3）.

[8] Wang Zhengming，Duan Xiaojun. Frequency－Domain Method for Separating Signal and Noise［J］. Science in China（Series E），1999（6）.

[9] Wang Zhengming，Duan Xiaojun. Overall test and evaluation based on trajectory data［J］. Science in China（Series E），2000（6）.