第 46 卷 第 4 期　　天 文 学 报　　Vol.46 No.4
2005 年 10 月　　ACTA ASTRONOMICA SINICA　　Oct., 2005

野值预剔除方法的改进

吴会英　吴连大

(中国科学院紫金山天文台 南京 210008)

摘要　在初轨有较大误差时, 传统的光学资料野值预剔除方法 —— 截断最小二乘 L 估计法不能有效地剔除野值. 从门限和模型两方面改进野值预处理方法, 即改用统计中误差定门限. 将残差的线性模型改为 2 次曲线模型, 结果为: 对于模拟产生的所有卫星弧段, 在初轨有 10^{-4} (半长轴 a 为其 5%, 以下不再说明) 误差情况下, 野值比例都能降低到 10% 以下.

关键词　人造卫星观测: 最小二乘, 野值, 门限
中图分类号: P 133;　　**文献标识码**: A

1 引言

由于所有的光学观测资料都不可避免地存在着野值, 为了在野值较少的情况下进行精密定轨, 资料预处理是很关键的. 如何选择资料处理的方法, 达到最好的剔除野值的效果, 是本文要研究的重点. 野值预剔除多用稳健估计法, 如文 [1] 中提到的稳健估计在激光测距资料中的应用. 资料处理的目的是判别观测目标是否正确, 尽可能多地剔除错误资料, 更好地进行轨道改进. 文 [2,3] 中提出了一种资料处理的方法 —— 截断最小二乘 L 估计法, 此方法对于初轨误差不大 (如 10^{-5}) 的情况是能满足实际需要的. 当轨道误差较大, 而又要求剔除较小野值时, 截断最小二乘 L 估计法剔除野值的效果并不十分理想, 这就需要我们对此方法进行改进, 使资料预处理的效果更好. 本文从门限和模型两方面改进了截断最小二乘 L 估计法.

2 截断最小二乘 L 估计法的剔除效果

文 [2] 中介绍的资料预处理方法为截断最小二乘 L 估计法, 这里不再赘述. 对于 $5''$ (令 $\sigma_0 = 5''$) 的资料精度、 10^{-4} 的初轨精度和 $6\sigma_0$ 的野值, 经过大量模拟计算, 我们选出 3 种门限, 野值剔除结果见表 1.

2004-09-28 收到原稿, 2005-01-14 收到修改稿

表 1 不同门限 $6\sigma_0$ 野值剔除效果

Table 1 The result of various thresholds ($6\sigma_0$)

编号	门限 1 (threshold 1)		门限 2(threshold 2)		门限 3(threshold 3)	
	误剔	漏剔	误剔	漏剔	误剔	漏剔
(ID)	(false)	(missing)	(false)	(missing)	(false)	(missing)
03576	0.657%	8.408%	7.396%	7.024%	5.409%	4.545%
05554	0.537%	6.743%	6.911%	3.419%	7.793%	3.824%
05679	0.045%	9.732%	4.106%	7.949%	4.469%	4.642%
06124	0.368%	7.266%	1.336%	2.730%	10.441%	4.458%
08818	0.497%	7.543%	5.616%	4.632%	8.309%	4.230%
13172	0.000%	13.283%	0.223%	11.864%	1.258%	5.678%
14401	0.039%	13.436%	0.208%	12.227%	3.605%	6.206%
17586	0.318%	3.520%	3.668%	1.286%	11.688%	1.436%
18944	0.701%	4.612%	0.927%	0.497%	15.497%	2.136%
19460	0.190%	13.171%	0.238%	11.594%	4.853%	6.301%
19468	0.144%	7.946%	0.494%	5.748%	8.023%	4.751%
20547	0.000%	16.051%	0.000%	15.403%	1.904%	10.040%
20642	0.741%	7.727%	6.371%	5.411%	6.212%	3.793%
21147	0.000%	13.571%	2.135%	12.418%	1.044%	5.781%
22519	0.645%	7.660%	5.373%	2.171%	4.869%	4.268%
25017	0.000%	13.299%	2.022%	12.238%	1.689%	6.492%
25615	0.042%	12.998%	0.140%	11.957%	3.303%	6.761%
26234	0.068%	13.121%	0.149%	12.126%	3.776%	7.338%
26906	0.067%	9.262%	4.751%	7.634%	3.467%	4.636%
27433	0.030%	12.643%	0.099%	11.451%	4.651%	7.106%
均值	0.254%	10.100%	2.608%	7.989%	5.613%	5.221%

注：门限 1: $\nu_{Ti} > 0.002''$, $\nu_{\theta i} > 0.002^\circ$；门限 2: $\nu_{Ti} > 0.001''$, $\nu_{\theta i} > 0.002^\circ$；门限 3: $\nu_{Ti} > 0.001''$, $\nu_{\theta i} > 0.001^\circ$.

误剔：错误剔除的好资料占总的好资料的比例；漏剔：资料处理后所剩资料中的野值比例.

由表 1 易见：此种条件下的剔除效果并不理想，而且不同卫星差异很大. 当然我们可以通过调整门限来使漏剔比例降低，但对于较小的门限，误剔比例又会升高. 由此可见，对于初轨误差较大，且又要求剔除较小的野值的情况下，截断最小二乘 L 估计法存在着缺陷.

3 方法改进

由上文可知：在初轨误差 (10^{-4}) 较大时，用截断最小二乘 L 估计法剔除小野值 ($6\sigma_0$) 的效果并不理想，通过研究，我们认为有以下两点原因：

(1) 门限不合理

上文用的资料处理门限是通过分析计算所得，我们并未给出对于不同的卫星、不同弧段、不同资料精度均适用的门限，因此需要找到一种更合理、更普适的门限.

(2) 逼近模型有误差

对于初轨误差较大 (如 10^{-4}) 的长弧段，线性模型的拟合误差很大 (见图 1).

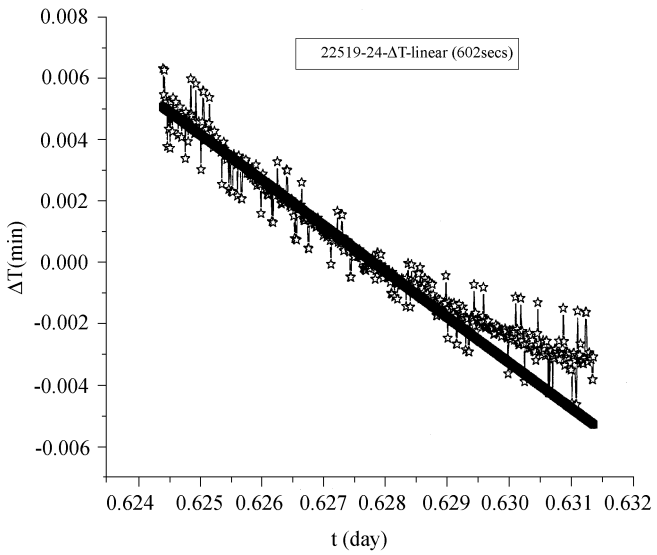

图 1　22519 第 24 圈资料线性拟合沿迹误差

Fig. 1　The linear fitting of ΔT (the 24th pass of 22519)

基于以上两点，我们进一步从门限和模型两方面研究了资料处理方法.

3.1 门限改进 — 统计中误差定门限法

仍假定线性模型成立，残差 $\nu_{Ti}, \nu_{\theta i}$ 符合正态分布，则根据最小二乘 3σ 剔除原则，就可依据 (1) 式剔除野值：

$$|\nu_{Ti}|\rangle 3\sigma_T, \quad \text{或} \quad |\nu_{\theta i}|\rangle 3\sigma_\theta\,. \tag{1}$$

这里的 σ_T, σ_θ 并非已知量，下面将研究它们的估计方法. 我们可根据截断最小二乘法的目标函数计算 S_{T0} 、 $S_{\theta 0}$:

$$S_{T0} = \sqrt{\sum_{i=1}^{N\beta} \nu_{N,Ti}^2/(N\beta - p)}\,, \tag{2}$$

$$S_{\theta 0} = \sqrt{\sum_{i=1}^{N\beta} \nu_{N,\theta i}^2/(N\beta - p)}\,. \tag{3}$$

其中， N 为本圈资料个数，本文中的 β 取 0.5, $\{\nu_{N,Ti}\}$ 表示按残差 ν_{Ti} 的绝对值从小到大排序后的序列； $\{\nu_{N,\theta i}\}$ 表示按残差 $\nu_{\theta i}$ 的绝对值从小到大排序后的序列； p 为待估参数的个数，对于线性模型 $p = 2$. 显然 S_{T0} 、 $S_{\theta 0}$ 并非 $\nu_{Ti}, \nu_{\theta i}$ 的中误差，只是残差较小的、占好资料的比例为 5/8 (本文的模拟资料的好资料比例为 80%, S_{T0} 、 $S_{\theta 0}$ 对应于全部资料的 50%) 的资料的统计中误差. 但是，如果好资料的残差严格符合正态分布，我们

就可找出 S_{T0} 、 $S_{\theta 0}$ 与好资料中误差 σ_T, σ_θ 的关系. 根据统计理论, 正态分布残差较小的前 5/8 的资料的中误差为全部资料中误差的 0.48468389 倍, 即:

$$\sigma_T = \frac{S_{T0}}{0.48468389} = 2.063 S_{T0} = k_0 S_{T0} = \frac{k_1}{3} S_{T0}\,, \tag{4}$$

$$\sigma_\theta = \frac{S_{\theta 0}}{0.48468389} = 2.063 S_{\theta 0} = k_0 S_{\theta 0} = \frac{k_1}{3} S_{\theta 0}\,. \tag{5}$$

由于模型误差的存在, ν_{Ti} 、 $\nu_{\theta i}$ 并不严格符合正态分布, (4) 式和 (5) 式中的系数 k_0(或 k_1) 需修正, 为了得出线性模型待估系数 k_0 的修正量, 我们统计了 80% 的好资料中各种残差的比例分布, 见表 2:

表 2 线性模型残差比例

Table 2 The residual error proportion of linear model

误差限 (error limit) $\lvert\Delta\rvert \le \mu\sigma$[3]	置信水平 [4] (confidence level)	10^{-5} 初轨误差 (10^{-5} initial orbit error)		5×10^{-5} 初轨误差 (5×10^{-5} initial orbit error)		10^{-4} 初轨误差 (10^{-4} initial orbit error)	
		沿迹 (ΔT)	垂迹 ($\Delta\theta$)	沿迹 (ΔT)	垂迹 ($\Delta\theta$)	沿迹 (ΔT)	垂迹 ($\Delta\theta$)
3.000000σ	99.70%	99.21%	99.61%	99.04%	99.57%	99.07%	99.54%
2.575829σ	99.00%	98.26%	98.78%	97.90%	98.74%	97.66%	98.68%
2.000000σ	95.50%	94.69%	95.14%	94.20%	95.11%	93.40%	95.01%
1.959964σ	95.00%	94.28%	94.67%	93.79%	94.64%	92.99%	94.55%
1.644854σ	90.00%	89.87%	89.78%	89.65%	89.76%	88.79%	89.69%
1.281552σ	80.00%	81.29%	79.97%	81.81%	80.11%	81.48%	80.23%
1.036433σ	70.00%	72.68%	70.42%	74.03%	70.54%	74.71%	70.85%
1.000000σ	68.30%	71.23%	68.79%	72.71%	68.99%	73.55%	69.31%
0.887162σ	62.50%	66.38%	63.71%	68.29%	63.83%	69.79%	64.27%
0.841621σ	60.00%	64.30%	61.65%	66.34%	61.72%	68.07%	62.13%
0.674490σ	50.00%	55.84%	52.62%	58.41%	52.91%	61.13%	53.37%
0.524401σ	40.00%	45.88%	42.49%	48.88%	42.82%	52.84%	43.42%
0.385320σ	30.00%	35.12%	32.12%	38.21%	32.41%	42.84%	32.83%
0.253347σ	20.00%	23.90%	21.61%	26.17%	21.78%	30.60%	22.14%
0.125661σ	10.00%	12.22%	11.19%	13.55%	11.17%	16.26%	11.34%

由表 2 可见, 残差并不严格符合正态分布, 尤其是当初轨误差较大时, 而且沿迹方向尤为明显, 还有将近 1% 的好资料残差大于 3σ. 可见, 我们统计的 S_{T0} 、 $S_{\theta 0}$ 偏小, 这说明系数 k_0 将比 2.063 大, 假定其在 2.5 到 3.5 之间, 乘上 3, 可近似认为 $k_1 \in$[7,10], 表 3 为用此门限计算的 $6\sigma_0$ 野值剔除结果, 并与表 1 的结果进行了比较.

由表 3 可见: 误剔、漏剔比例随着 k_1 取值变化而变化, k_1 愈大误剔比例愈小, 而漏剔比例愈大, 在实际使用时, k_1 值可根据数据的实际情况选取. 显然, 对于所有卫星的总体效果, 可调门限的剔除效果明显提高.

表 3　不同门限 $6\sigma_0$ 野值剔除效果

Table 3　The result of various thresholds ($6\sigma_0$)

编号	门限 3 (threshold 3)		门限 4 (threshold 4)		门限 5(threshold 5)	
	误剔	漏剔	误剔	漏剔	误剔	漏剔
(ID)	(false)	(missing)	(false)	(missing)	(false)	(missing)
03576	5.409%	4.545%	10.153%	2.800%	4.135%	5.506%
05554	7.793%	3.824%	8.291%	1.966%	3.053%	5.586%
05679	4.469%	4.642%	10.821%	1.970%	5.172%	5.258%
06124	10.441%	4.458%	4.375%	0.102%	1.397%	2.271%
08818	8.309%	4.230%	9.628%	1.252%	4.260%	4.205%
13172	1.258%	5.678%	7.020%	1.101%	2.901%	4.452%
14401	3.605%	6.206%	9.397%	0.741%	4.373%	3.785%
17586	11.688%	1.436%	5.844%	0.604%	1.937%	3.183%
18944	15.497%	2.136%	2.123%	0.000%	0.587%	0.985%
19460	4.853%	6.301%	8.254%	0.644%	3.247%	3.830%
19468	8.023%	4.751%	6.789%	0.439%	3.045%	3.402%
20547	1.904%	10.040%	10.313%	1.696%	5.355%	4.978%
20642	6.212%	3.793%	9.000%	1.996%	3.503%	6.008%
21147	1.044%	5.781%	9.891%	2.314%	4.246%	5.854%
22519	4.869%	4.268%	6.357%	1.247%	2.538%	3.881%
25017	1.689%	6.492%	10.134%	1.992%	4.837%	5.164%
25615	3.303%	6.761%	8.188%	0.334%	4.059%	3.383%
26234	3.776%	7.338%	8.785%	0.634%	4.264%	4.030%
26906	3.467%	4.636%	9.779%	2.847%	3.563%	5.836%
27433	4.651%	7.106%	8.427%	0.604%	3.061%	4.005%
均值	5.613%	5.221%	8.178%	1.264%	3.477%	4.280%

注：门限 3: $\nu_{Ti} > 0.001\,\mathrm{min}$　　$\nu_{\theta i} > 0.001^\circ$, 门限 4: $k_1 = 7$, 门限 5: $k_1 = 10$; 误剔、漏剔的计算方法和表 1 相同.

3.2　模型改进 —— 截断最小二乘 2 次曲线拟合法

为了更有效地剔除野值，可以考虑采用残差的 2 次曲线模型，或称之为截断最小二乘 2 次曲线拟合法. 本方法假定观测量为时间的 2 次函数：

$$y_i = a + b(t_i - t_0) + c(t_i - t_0)^2 + e_i\,. \tag{6}$$

截断最小二乘 2 次曲线拟合法的目标函数与 L 估计法相同，由于此资料处理方法需要估计 3 个参数，为了在节省计算量的前提下更好地使样点平均分布，我们采用了一种新的抽样方法：资料平均分组随机抽样法.

将资料分成 k 组，每组 m 份资料 $(m = int(N/k)$, 第 k 组 $N - (k-1)m$ 个), 在每组资料中随机抽取 1 份资料，用所选的 k 份资料估计 a, b, 抽取样点的次数就是目标函数的计算次数，这样，我们就可以控制计算量. 此种抽样方法能至少得到一次抽样均是好资料的概率推导如下：

(1) 设好资料比例为 x, 而且 k 组比例相同，从 k 组中随机选出的 k 个资料 (每组抽一个点) 都是好资料的概率为 x^k;

(2) 一次 k 点抽样，起码有 1 个野值的概率为 $1-x^k$;

(3) N_0 次抽样均有野值的概率为 $(1-x^k)^{N_0}$;

(4) N_0 次抽样起码有一次没有野值的概率为 $1-(1-x^k)^{N_0}$, 记为:

$$P(x,k,N_0)=1-(1-x^k)^{N_0}\,. \tag{7}$$

易见， $P(x,k,N_0)$ 为 x 、 N_0 的增函数， k 的减函数. 但是这里有 3 个待估参数，分组数起码要大于等于 3, 为了使抽样点有较好的分布，我们选择了 6 点抽样法，沿迹方向拟合图见图 2, 总体剔除效果见表 4 (2 次曲线拟合只是针对长弧，短弧资料仍然采用线性模型，并且和线性模型的剔除效果进行了比较).

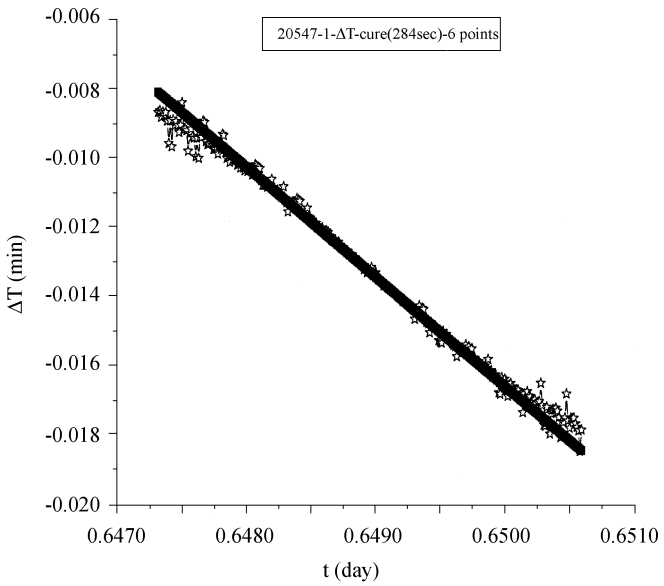

图 2 20547 第 1 圈资料 6 点抽样法 2 次曲线拟合沿迹误差

Fig. 2 The quadratic curve fitting of ΔT (the first pass of 20547-6 points)

由图 2 易见， 2 次模型能较好地反映误差的实际情况.

由表 4 易见， 2 次曲线拟合的总体效果要比线性模型有明显改善，采用截断最小二乘 2 次曲线拟合法，误剔比例和漏剔比例均已得到满意的结果；但对于线性模型，抽样方法的改变并未使剔除的总体效果有显著提高.

这里需要对线性模型中的 2 点随机抽样做一说明：当抽取的 2 个样点的时间差小于 1/3 弧长时，取消此次抽样，重新随机抽样.

表 4 不同方法 $6\sigma_0$ 野值的剔除效果

Table 4 The result of quadratic curve fitting – 6 points

编号	线性 2 点文 [1] 抽样 (linear 2points [1] sample)		线性 2 点随机抽样 (linear 2 points stochastic sample)		2 次 6 点随机抽样 (square 6 points stochastic sample)	
ID	误剔 (false)	漏剔 (missing)	误剔 (false)	漏剔 (missing)	误剔 (false)	漏剔 (missing)
03576	4.135%	5.506%	4.087%	5.533%	0.184%	0.670%
05554	3.053%	5.586%	3.079%	5.552%	0.038%	0.230%
05679	5.172%	5.258%	5.309%	5.115%	0.159%	0.901%
06124	1.397%	2.271%	1.373%	2.247%	0.184%	1.619%
08818	4.260%	4.205%	4.298%	4.225%	0.248%	0.836%
13172	2.901%	4.452%	2.921%	4.491%	0.304%	1.740%
14401	4.373%	3.785%	4.464%	3.738%	0.989%	3.342%
17586	1.937%	3.183%	2.001%	3.154%	0.016%	0.063%
18944	0.587%	0.985%	0.608%	0.945%	0.113%	0.615%
19460	3.247%	3.830%	3.187%	3.828%	0.702%	3.402%
19468	3.045%	3.402%	3.086%	3.404%	0.165%	2.334%
20547	5.355%	4.978%	5.276%	5.087%	1.626%	5.416%
20642	3.503%	6.008%	3.724%	5.948%	0.062%	0.440%
21147	4.246%	5.854%	4.269%	5.745%	0.237%	1.752%
22519	2.538%	3.881%	2.422%	3.829%	0.008%	0.000%
25017	4.837%	5.164%	4.803%	5.184%	0.448%	2.113%
25615	4.059%	3.383%	3.961%	3.243%	0.280%	2.783%
26234	4.264%	4.030%	4.372%	4.047%	1.385%	3.340%
26906	3.563%	5.836%	3.601%	5.794%	0.201%	0.762%
27433	3.061%	4.005%	3.001%	4.012%	0.745%	4.127%
均值	3.477%	4.280%	3.492%	4.256%	0.405%	1.824%

注：可调门限的系数 k_1=10, 误剔、漏剔的计算方法同表 1.

4 结论

当初轨误差较大, 弧段较长时, 会导致截断最小二乘 L 估计法不能有效的剔除野值, 我们从门限和模型两方面改进了此方法，得到结论：

(1) 用统计中误差定门限的方法优于原来的经验门限；

(2) 2 次曲线拟合模型优于线性模型.

对于模拟资料来说，虽然用线性模型的可调门限去剔除野值已能得到很好的效果, 我们仍然研究了具有更好剔除效果的 2 次曲线拟合，原因是在实际应用中可能还会遇到许多难以预测的问题. 当然，如果实际工作需要，我们还可以尝试 3 次、甚至更高次的曲线拟合. 在实际应用中，可调门限的系数仍然可以根据需要调整.

参 考 文 献

1 贾沛璋，谭德同，张忠平，等．测绘学报， 1994,23(4): 302
2 Wu Lian-da, Jia Pei-zhang. AcASn, 1994, 35(2): 113 (吴连大，贾沛璋．天文学报， 1994, 35(2): 113)
3 Wu Hui-ying, Wu Lian-da. AcASn, 2005, 46(2): 173 (吴会英，吴连大．天文学报， 2005, 46(2): 173)
4 丁月蓉．天文数据处理方法．南京：南京大学出版社， 1998

An Improvement Method for Data Preprocess

WU Hui-ying WU Lian-da
(*Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing* 210008)

Abstract When initial orbit errors are too big, the traditional optical data preprocess method-interceptive least-square L-estimation method can not delete outliers effectively. This method is improved by thresholds and model: the thresholds are determined by statistical mean error; the linear model is changed to quadratic curve model. The conclusions are: for all satellite passes which are simulated, when the initial orbit error is 10^{-4}(semi-major is 5 percent of it, which don't explain in the following) ,the proportion of outliers is reduced less than 10% in all cases.

Key words observation of artificial satellite: least-square, outliers, thresholds