引文格式:齐志军,方兴,吕志鹏.两种适用于线性回归EIV模型的高崩溃污染率算法[J].武汉大学学报(信息科学版),2025,50(1):74-82.DOI:10.13203/j.whugis20220441

Citation:QI Zhijun,FANG Xing,LÜ Zhipeng.Two Algorithms with High Breakdown Points Applied in Linear Regression EIV Model[J].Geomatics and Information Science of Wuhan University,2025,50(1):74-82.DOI:10.13203/j.whugis20220441

两种适用于线性回归EIV模型的高崩溃污染率算法

齐志军[1] 方 兴[1] 吕志鹏[2]

1 武汉大学测绘学院,湖北 武汉,430079

2 华东交通大学交通运输工程学院,江西 南昌,330013

摘 要:混合总体最小二乘是求解带有固定列的线性回归变量误差(errors-in-variables,EIV)模型的严密方法,结合M估计可以进一步增加其稳健性。但是M估计结果受初值影响,容易收敛错误。针对该问题,将两种高斯-马尔可夫模型下的抗差估计算法拓展到EIV模型中,提出两种高崩溃污染率的算法,即加权总体最小平方中值法(weighted total least median of squares,WTLMS)和加权截断总体最小二乘法(weighted total least trimmed squares,WTLTS)。分析两种算法的等变性质和崩溃污染率,给出单位权中误差的评定公式,分别通过重采样方法和可行集算法得到参数估计值。不同于已有的高崩溃污染率算法,所提算法考虑系数矩阵存在固定列的情况,同时减少对随机模型的限制。仿真数据和真实数据解算结果验证了两种算法在高粗差污染的观测数据中能够得到稳健可靠的估计结果。

关键词:加权总体最小平方中值法;加权截断总体最小二乘法;EIV模型;崩溃污染率;线性回归

中图分类号:P207 **文献标识码**:A **收稿日期**:2023-01-05

DOI:10.13203/j.whugis20220441 **文章编号**:1671-8860(2025)01-0074-09

Two Algorithms with High Breakdown Points Applied in Linear Regression EIV Model

QI Zhijun[1] *FANG Xing*[1] *LÜ Zhipeng*[2]

1 School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China

2 School of Civil Engineering and Architecture, East China Jiaotong University, Nanchang 330013, China

Abstract: **Objectives**: Linear regression model is a basic model in the field of geodesy. To consider the structure of the coefficient matrix with the fixed column, the mixed least squares and total least squares method is implemented. However, it is easily contaminated by outliers. The M-estimator results depend on the initial value and are extremely prone to convergence badly. To increase the robustness, we propose two algorithms with high breakdown points for linear regression errors-in-variables (EIV) models, namely, the weighted total least median of squares (WTLMS) method and the weighted total least trimmed squares (WTLTS) method. **Methods**: The two algorithms are extensions of traditional algorithms and use a more general stochastic model. Their breakdown points are near 50% and the two algorithms have two equivariant properties: scale equivariance and affine equivariance. The estimation formula of variance components is given. Since their objective functions are not differentiable, WTLMS and WTLTS get the solutions by the resampling algorithm and the feasible set algorithm in the EIV model respectively. **Results**: The results show that: (1) The result of the M-estimator is biased heavily from the real line, while the two proposed algorithms can obtain results close to the true value. Their performances are significantly better than M-estimator in terms of root mean square error and standard deviation. The efficiency of the two algorithms is not

基金项目:国家自然科学基金(41774009);国家自然科学基金青年科学基金(42204047)。

第一作者:齐志军,硕士,主要从事测量数据处理方面的研究。2016301610038@whu.edu.cn

通信作者:吕志鹏,博士,讲师。lv_zhipeng1989@qq.com

high, which can be further improved when the results of the two algorithms are used as the initial value of the M-estimator. The breakdown points of the two algorithms are close to 50% in the real data, which is extremely robust. (2) In the experiment of the LiDAR data, the performance of the proposed methods is better than that of the M-estimator. **Conclusions:** The two proposed algorithms have outstanding robustness, but their complexities are high and their efficiency is not ideal. We will focus to find an easy solution with higher efficiency.

Key words: weighted total least median of squares; weighted total least trimmed squares; errors-in-variables model; breakdown point; linear regression

线性回归模型是测绘领域中的一种基本模型,在直线拟合、点云平面拟合等方面应用十分广泛。传统方法只考虑观测向量误差,建立高斯-马尔可夫(Gauss-Markov,GM)模型,采用最小二乘(least square,LS)得到参数估值。此方法忽略系数矩阵的误差,估计结果在统计上是有偏的[1]。为了顾及观测向量和系数矩阵的误差,文献[2]拓展GM模型到变量误差(errors-in-variables,EIV)模型,并采用总体LS(total LS,TLS)求解参数。

当粗差污染观测向量和系数矩阵中的观测值时,TLS得到的估计值将会严重失真[3]。为了削弱粗差的不利影响,将稳健估计理论应用到EIV模型的解算,主要分为两类:(1)基于均值漂移模型识别粗差。文献[4]使用稳健加权TLS(robust weighted TLS,RWTLS)拟合点云平面,通过3倍中误差准则剔除粗差,文献[5]采用Baarda粗差探测法定位异常观测值。然而对于可疑粗差,基于识别的方法缺乏较好的处理手段。(2)基于方差膨胀模型控制粗差的影响。文献[6]使用同一单位权中误差对粗差进行降权,文献[7-8]对于观测向量和系数矩阵采用不同的单位权中误差进行处理,文献[9]和文献[10]分别针对三维坐标转换模型和多变量EIV模型提出相应的稳健方法。以上方法保留TLS处理正常观测数据的优良特性,但求解的是非凸优化问题,采用对粗差敏感的TLS结果作为初值,参数估值极其容易失真。

为了提高算法的崩溃污染率,文献[11]采用中位参数法建立RWTLS算法,但是算法稳健性和参数个数有关,实际中远达不到50%的最高崩溃污染率;文献[12]将传统的最小平方中值法(least median of squares,LMS)和截断最小二乘法(least trimmed squares,LTS)[12]应用到EIV模型中,提出总体LMS(total LMS,TLMS)[13]和截断总体最小二乘法(total LTS,TLTS)[14],但是两种方法对随机模型的结构有严格限制,前者要求独立等精度观测,后者系数矩阵对应的协因数矩阵满足特定结构,都不具备通用性。求解方法上,前者是一个抽样检验过程,参数精度较低,后者通过分支界定算法求得精确解,但是存在大量的冗余计算。

针对现有算法的不足,本文提出了两种高崩溃污染率的稳健估计算法,分别是加权TLMS(weighted TLMS, WTLMS)和加权TLTS(weighted TLTS,WTLTS)。从线性回归EIV模型出发,计算数据点到拟合超平面的加权残差。为了避免粗差的影响,分别基于中位数准则和截断准则得到两种改进算法的目标函数,进一步证明两种算法具有估计等变性和高崩溃污染率。不同于传统的高崩溃污染率算法,本文算法能够顾及系数矩阵的固定列,并且考虑同一个观测方程中的观测元素存在相关性的情况,随机模型更具通用性。因为目标函数不可导,基于梯度的优化算法将失效,借鉴LMS和LTS的求解思路,给出两种算法的求解过程。仿真实验和实测数据验证了本文算法的高崩溃污染率,但本文算法有效性不足,可以作为RWTLS的初值,获得有效性更佳的参数估值。

1　混合总体最小二乘法

求解线性回归的参数时,系数矩阵存在常数列,对应的模型被称为带有固定列的EIV模型,其函数模型和随机模型表达为[15]:

$$\boldsymbol{y}-\boldsymbol{e}_y=(\boldsymbol{A}-\boldsymbol{E}_A)\boldsymbol{\xi}=\left(\boldsymbol{A}^1-\boldsymbol{E}_{A^1}\right)\boldsymbol{\xi}_1+\boldsymbol{A}^2\boldsymbol{\xi}_2 \tag{1}$$

$$D(\boldsymbol{e})=\sigma_0^2\boldsymbol{Q} \tag{2}$$

式中,$\boldsymbol{y}$和$\boldsymbol{e}_y$分别是$n\times 1$维观测向量及其对应的随机误差向量;$\boldsymbol{A}$和$\boldsymbol{E}_A$分别是$n\times m$的固定系数矩阵及其对应的误差矩阵;$\boldsymbol{\xi}$是$m\times 1$待估参数向量;$\boldsymbol{A}^1$和$\boldsymbol{A}^2$分别为观测值组成的矩阵和常数组成的矩阵;$\boldsymbol{\xi}_1$和$\boldsymbol{\xi}_2$分别是与$\boldsymbol{A}^1$和$\boldsymbol{A}^2$对应的m_1维和$m-m_1$维参数向量;$D(*)$是计算向量方差矩阵

的算子;$\boldsymbol{e}=\begin{bmatrix}\mathrm{vec}(\boldsymbol{E}_{A^1})\\ \boldsymbol{e}_y\end{bmatrix}$是所有观测量组成的残差向量,其中vec(*)是矩阵按列向量化算子;σ_0^2是单位权方差;$\boldsymbol{Q}$是$\boldsymbol{e}$的协因数矩阵。

混合总体最小二乘法(mixed least squares and total least squares,mixed LS-TLS)是求解带有固定列EIV模型的严密方法,令$\boldsymbol{B}=\begin{bmatrix}\boldsymbol{\xi}_1^{\mathrm{T}}\otimes\boldsymbol{I}_n & -\boldsymbol{I}_n\end{bmatrix}$,数值计算上等价于求解带有非线性约束的优化问题:

$$\min:\boldsymbol{e}^{\mathrm{T}}\boldsymbol{Q}^{-1}\boldsymbol{e}\quad \text{s.t. } \boldsymbol{y}-\boldsymbol{A}\boldsymbol{\xi}+\boldsymbol{B}\boldsymbol{e}=0 \tag{3}$$

该优化问题可以转换为无约束的非线性优化问题[16]:

$$\min:(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\xi})^{\mathrm{T}}\left(\boldsymbol{B}\boldsymbol{Q}\boldsymbol{B}^{\mathrm{T}}\right)^{-1}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\xi}) \tag{4}$$

该转换使得优化问题不需要服从约束,可以看作是最小瑞利商估计的拓展,统计上可以看作残差$\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\xi}$及其权阵$\left(\boldsymbol{B}\boldsymbol{Q}\boldsymbol{B}^{\mathrm{T}}\right)^{-1}$的二次型运算。如果观测值是不相关,即$\boldsymbol{Q}$为对角矩阵,式(4)可以用分量求和的形式表示为:

$$\min:\sum_{i=1}^{n}\frac{\left(y_i-\boldsymbol{A}_i\boldsymbol{\xi}\right)^2}{Q_{y_i}+\boldsymbol{\xi}_1^{\mathrm{T}}\boldsymbol{Q}_{A_i^1}\boldsymbol{\xi}_1} \tag{5}$$

式中,$\boldsymbol{A}_i$为矩阵$\boldsymbol{A}$的第i行元素;y_i和Q_{y_i}分别为向量$\boldsymbol{y}$的第i个分量及其协因数;$\boldsymbol{A}_i^1$和$\boldsymbol{Q}_{A_i^1}$分别为矩阵$\boldsymbol{A}_1$的第i行元素及其协因数矩阵。mixed LS-TLS的准则是最小化所有数据点的残差平方和,所以该方法极其容易被粗差污染,即使单个粗差也可能使得估计崩溃。

2 WTLMS估计和WTLTS估计

求和运算不能抵抗粗差的干扰,可以使用其他数学运算取代。文献[12]提出了两种在GM模型中的稳健估计方法,分别是LMS估计和LTS估计。文献[13-14]将这两种方法拓展到EIV模型,但仍然存在一些不足。在此基础上,针对不等精度的带有固定列EIV模型,提出WTLMS估计和WTLTS估计这两种高崩溃污染率估计算法,讨论其估计特性和崩溃污染率,并给出两种算法的求解步骤。

2.1 估计准则

用中位数运算代替求和运算,WTLMS估计的准则为:

$$\min:\underset{i=1}{\overset{n}{\mathrm{med}}}\frac{\left(y_i-\boldsymbol{A}_i\boldsymbol{\xi}\right)^2}{Q_{y_i}+\boldsymbol{\xi}_1^{\mathrm{T}}\boldsymbol{Q}_{A_i^1}\boldsymbol{\xi}_1} \tag{6}$$

式中,$\underset{i=1}{\overset{n}{\mathrm{med}}}(*)$表示取中位数操作。需要解释的是,协因数矩阵$\boldsymbol{Q}$不局限于是对角矩阵,可以推广到同一观测方程中的随机量是相关的情况,即$\boldsymbol{Q}_{A_i^1y_i}\neq 0$,对应式(6)的分母为$Q_{y_i}+\boldsymbol{\xi}_1^{\mathrm{T}}\boldsymbol{Q}_{A_i^1}\boldsymbol{\xi}_1+2\boldsymbol{\xi}_1^{\mathrm{T}}\boldsymbol{Q}_{A_i^1y_i}$,为了方便表达,依旧将$\boldsymbol{Q}$视为对角矩阵。与文献[13]中$\boldsymbol{Q}=\boldsymbol{I}$的情况进行比较,本文算法的随机模型更具通用性。

将所有加权残差平方$r_i^2=\dfrac{\left(y_i-\boldsymbol{A}_i\boldsymbol{\xi}\right)^2}{Q_{y_i}+\boldsymbol{\xi}_1^{\mathrm{T}}\boldsymbol{Q}_{A_i^1}\boldsymbol{\xi}_1}$进行升序排列,得到WTLTS估计的准则为:

$$\min:\sum_{i=1}^{h}r_i^2 \tag{7}$$

式中,h为截断参数;$r_1^2<r_2^2<\cdots<r_n^2$。根据粗差比例不能超过50%的假设,需要满足$h>n/2$。式(7)表示算法抛弃残差较大的$n-h$个观测值,在一个不含粗差的子集中进行参数估计,从而消除粗差的影响。与文献[14]比较,WTLTS的协因数矩阵不受到特定结构的限制。

2.2 等变特性

WTLMS估计满足仿射等变性、尺度等变性。令$\boldsymbol{C}$为任意$m_1\times m_1$可逆矩阵,c为任意常数,式(6)的解$\hat{\boldsymbol{\xi}}(\boldsymbol{A},\boldsymbol{y})$满足如下关系:

1)仿射等变性。

$$\hat{\boldsymbol{\xi}}(\boldsymbol{A}^1\boldsymbol{C},\boldsymbol{A}^2,\boldsymbol{y})=\boldsymbol{C}^{-1}\hat{\boldsymbol{\xi}}(\boldsymbol{A}^1,\boldsymbol{A}^2,\boldsymbol{y}) \tag{8}$$

2)尺度等变性。

$$\hat{\boldsymbol{\xi}}(\boldsymbol{A},c\boldsymbol{y})=c\hat{\boldsymbol{\xi}}(\boldsymbol{A},\boldsymbol{y}) \tag{9}$$

证明如下:

$$\begin{aligned}
&\underset{i=1}{\overset{n}{\mathrm{med}}}\frac{\left(y_i-\boldsymbol{A}_i^1\boldsymbol{C}\left(\boldsymbol{C}^{-1}\boldsymbol{\xi}_1\right)-\boldsymbol{A}_i^2\boldsymbol{\xi}_2\right)^2}{Q_{y_i}+\left(\boldsymbol{C}^{-1}\boldsymbol{\xi}_1\right)^{\mathrm{T}}\boldsymbol{Q}_{A_i^1\boldsymbol{C}}\left(\boldsymbol{C}^{-1}\boldsymbol{\xi}_1\right)}=\\
&\underset{i=1}{\overset{n}{\mathrm{med}}}\frac{\left(y_i-\boldsymbol{A}_i\boldsymbol{\xi}\right)^2}{Q_{y_i}+\left(\boldsymbol{C}\left(\boldsymbol{C}^{-1}\boldsymbol{\xi}_1\right)\right)^{\mathrm{T}}\boldsymbol{Q}_{A_i^1}\left(\boldsymbol{C}\left(\boldsymbol{C}^{-1}\boldsymbol{\xi}_1\right)\right)}=\\
&\underset{i=1}{\overset{n}{\mathrm{med}}}\frac{\left(y_i-\boldsymbol{A}_i\boldsymbol{\xi}\right)^2}{Q_{y_i}+\boldsymbol{\xi}_1^{\mathrm{T}}\boldsymbol{Q}_{A_i^1}\boldsymbol{\xi}_1}
\end{aligned} \tag{10}$$

$$\begin{aligned}
&\underset{i=1}{\overset{n}{\mathrm{med}}}\frac{\left(cy_i-\boldsymbol{A}_i(c\boldsymbol{\xi})\right)^2}{Q_{cy_i}+(c\boldsymbol{\xi}_1)^{\mathrm{T}}\boldsymbol{Q}_{A_i^1}(c\boldsymbol{\xi}_1)}=\\
&\underset{i=1}{\overset{n}{\mathrm{med}}}\frac{\left(cy_i-\boldsymbol{A}_i(c\boldsymbol{\xi})\right)^2}{c^2Q_{y_i}+(c\boldsymbol{\xi}_1)^{\mathrm{T}}\boldsymbol{Q}_{A_i^1}(c\boldsymbol{\xi}_1)}=\\
&\underset{i=1}{\overset{n}{\mathrm{med}}}\frac{\left(y_i-\boldsymbol{A}_i\boldsymbol{\xi}\right)^2}{Q_{y_i}+\boldsymbol{\xi}_1^{\mathrm{T}}\boldsymbol{Q}_{A_i^1}\boldsymbol{\xi}_1}
\end{aligned} \tag{11}$$

式中,$\boldsymbol{Q}_{A_i^1\boldsymbol{C}}$和$Q_{cy_i}$分别为$\boldsymbol{A}_i^1\boldsymbol{C}$的协因数矩阵和$cy_i$的协因数。等变性是参数估计中的重要性质[12],

仿射等变性展示对 $\boldsymbol{A}_1$ 进行线性变换如何影响WTLMS的估值,尺度等变性意味着WTLMS估值独立于 $\boldsymbol{y}$ 的单位。将式(10)~(11)的 $\underset{i=1}{\overset{n}{\text{med}}}(\cdot)$ 算子更换为 $\sum_{i=1}^{h}(\cdot)$ 算子,可证明WTLTS估计也具有上述两个性质。

2.3　崩溃污染率

崩溃污染率表示一种估计方法能承受的最大粗差比例,是衡量算法稳健性的重要指标。在 $\boldsymbol{A}$ 的任意 $m\times m$ 子矩阵可逆的情况下,那么WTLMS估计和WTLTS估计有限样本下的崩溃污染率 γ 为:

$$\gamma=(\lfloor(n-m)/2\rfloor+1)/n \tag{12}$$

式中,符号 $\lfloor\cdot\rfloor$ 表示向下取整。文献[12]给出了LMS估计和LTS估计的崩溃污染率的详细证明,本文两种算法的崩溃污染率与LMS和LTS算法是一致的。通过启发式的方法证明,当仅来自同一观测方程的观测值存在相关性时,无论是在系数矩阵 $\boldsymbol{A}$ 中或者观测向量 $\boldsymbol{y}$ 中的粗差,都是使对应的观测方程偏离实际的模型,故系数矩阵的粗差可以转换为对应观测向量中的粗差,本文算法可以达到式(12)的最大崩溃污染率。

将残差进行升序排列,使得 $r_1^2\leqslant r_2^2\leqslant\cdots\leqslant r_n^2$,令 $h=\lfloor(n-m)/2\rfloor+1$,WTLMS和WTLTS估计分别最小化 r_h^2 和 $\sum_{i=1}^{h}r_i^2$,这是满足等变性的估计方法所能达到的最大崩溃污染率。准确地说,WTLMS估计是一种特殊的分位数估计方法。WTLTS估计的 h 可以根据粗差比例 ω 设置为 $h=\lfloor n\omega\rfloor$,更大的 h 降低了崩溃污染率,但能够增加估计的有效性。一般情况下,当 n 远大于 m 时,根据式(12)可知本文两种估计方法具有50%的渐进崩溃污染率。

2.4　求解步骤

两种估计方法的几何意义十分明确,WTLMS估计表示数据点到超平面的加权平方残差中位数最小,WTLTS估计最小化数据点到超平面的前 h 个较小的加权平方残差和。然而对应的目标函数不可导,基于梯度的优化算法并不适合求解两种算法。

2.4.1　WTLMS的求解步骤

借鉴PROGRESS程序[12]求解LMS估计的流程,将重采样算法结合截距校正算法[17]和空间精化算法[18]应用到EIV模型中,WTLMS估计通过内外两重循环进行求解,具体步骤如下:

1) 计算 $h=\lfloor(n-m)/2\rfloor+1$,令 $i=0$ 和 $j=0$ 记录外循环和内循环迭代次数,内外循环最大迭代次数记为 n_r 和 N_r。

2) 开始外循环:(1) $i=i+1$ 并且 $i<N_r$;(2)从观测方程 $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{\xi}$ 中随机选择 m 个子集计算其精确解 $\hat{\boldsymbol{\xi}}^0$;(3)进行空间精化,令 $\hat{\boldsymbol{\xi}}^1,\hat{\boldsymbol{\xi}}^2,\cdots,\hat{\boldsymbol{\xi}}^{n_r}$ 是端点为 $\hat{\boldsymbol{\xi}}^0$ 和 $\tilde{\boldsymbol{\xi}}$ 的直线上的 n_r 个解,其中 $\tilde{\boldsymbol{\xi}}$ 是当前的最优解。

3) 开始内循环:(1) $j=0$ 并且 $j<n_r$;(2)计算 $\hat{\boldsymbol{\xi}}^j$ 对应的所有数据点加权残差平方,并校正截距,重新计算残差的加权平方;(3)将加权残差平方按升序排列,设置 $m^j=\left(r_h^j\right)^2$;(4)若 $m^j<\tilde{m}$,$\tilde{\boldsymbol{\xi}}=\hat{\boldsymbol{\xi}}^j$,$\tilde{m}=m^j$,其中 $\tilde{m}$ 是当前最佳目标函数值。

理论上应该遍历所有的 C_n^m 子集,对于大量数据可以固定重采样次数,如 $N_r=1\,000$,但不建议增加观测方程数量,这样会增加计算负担,降低其实用性。

2.4.2　WTLTS的求解步骤

一种有效求解WTLTS估计的方法是穷举法,一共需要进行 C_n^h 次,在观测值数量较小的情况下可以使用该方法,大量观测值将会使算法计算量快速增加,失去实用性。借鉴可行集算法[17],其核心是通过交换子集中的观测值使得目标函数下降,文献[19]分为两步将该算法应用于最小协方差行列式(minimum covariance determinant,MCD)估计,本文将该方法应用于线性回归EIV模型中。

首先需要证明通过交换子集元素能够降低目标函数值。在 n 个观测方程中,首先任意选取样本容量为 h 的子集 M_1,使用mixed LS-TLS方法在子集 M_1 中计算出估计结果 $\hat{\boldsymbol{\xi}}_{M_1}$ 及其加权残差平方和 $f_1=\sum_{i=1}^{h}r_{1i}^2$,其中 $r_{1i}^2=(y_i-\boldsymbol{A}_i\hat{\boldsymbol{\xi}}_{M_1})^2/(\boldsymbol{Q}_{y_i}+\hat{\boldsymbol{\xi}}_{M_1}^{\mathrm{T}}\boldsymbol{Q}_{A_i}\hat{\boldsymbol{\xi}}_{M_1})$。将 $\hat{\boldsymbol{\xi}}_{M_1}$ 的结果代入所有观测方程中,得到 n 个加权残差平方,选择其中 h 个残差平方最小的观测方程组成子集 M_2,在子集 M_2 中使用mixed LS-TLS计算得到 $\hat{\boldsymbol{\xi}}_{M_2}$ 及其加权残差平方和 $f_2=\sum_{i=1}^{h}r_{2i}^2$,则下列不等式成立:

$$f_1\geqslant f_2 \tag{13}$$

因为 M_2 中的元素在 n 个残差中选择最小的 h 个 r_{1i}^2,故 $\sum_{i=1}^{h}r_{1i}^2\big|_{i\in M_1}\geqslant\sum_{i=1}^{h}r_{1i}^2\big|_{i\in M_2}$。集合 M_2 中的 $\sum_{i=1}^{h}r_{2i}^2$

是最小的h个残差平方和,得到:

$$f_1=\sum_{i=1}^{h}r_{1i}^2\Big|_{i\in M_1}\geqslant\sum_{i=1}^{h}r_{1i}^2\Big|_{i\in M_2}\geqslant\sum_{i=1}^{h}r_{2i}^2\Big|_{i\in M_2}=f_2 \tag{14}$$

式(13)证明通过交换子集中元素可以降低目标函数值,以此设计迭代求解方法。但是WTLTS估计的目标函数是非凸的,估计结果依赖于迭代初值。可以使用LTS估计等稳健方法的结果作为迭代初值或者设置多个起算点进行计算,选择目标函数值最小的结果作为输出。WTLTS估计的解算步骤为:

1)通过LTS估计得到一个容量为h子集M_1。

2)在M_1计算$\hat{\boldsymbol{\xi}}_{M_1}$和$f_1=\sum_{i=1}^{h}r_{1i}^2$。选取其中残差较小的$h$个观测值,得到集合$M_2$,在$M_2$中计算$\hat{\boldsymbol{\xi}}_{M_2}$和$f_2=\sum_{i=1}^{h}r_{2i}^2$。

3)若$f_1=f_2$,输出结果$\hat{\boldsymbol{\xi}}=\hat{\boldsymbol{\xi}}_{M_2}$;否则,将子集$M_2$赋给$M_1$,重新进行步骤2)的解算。

2.5 精度评定

对于WTLMS估计而言,根据单位权中误差和残差中位数的关系,得到:

$$\hat{\sigma}_0=1.482\,6\sqrt{\underset{i=1}{\overset{n}{\operatorname{med}}}\,r_i^2} \tag{15}$$

在观测值数量较小的情况下,还需要等式右边乘以一个系数进行改正,系数和观测值数量关系参考文献[20]。

对于WTLTS估计,其相当于在子集中进行mixed LS-TLS估计,使用总体最小二乘的精度评定公式,单位权中误差的估计值为:

$$\hat{\sigma}_0=\sqrt{\frac{1}{h-m}\sum_{i=1}^{h}r_i^2} \tag{16}$$

3 模拟实验与真实算例

3.1 模拟实验

设直线方程为$y=-x+5(0<x<5)$,即斜率和截距的真值分别为$\tilde{a}_1=-1$和$\tilde{a}_2=5$。采用均匀分布函数产生20个升序排列的x值,计算对应的y值,得到坐标的真值。按照文献[21]的方法加入期望为0、标准差为0.04的随机误差,同一点x分量和y分量的相关系数ρ_{xy}为0.6,通过$\sigma_{xy}=\rho_{xy}\sigma_x\sigma_y$计算坐标的协方差,得到含随机误差的观测值。在前10个点中随机选择5个,在x分量和y分量上加入绝对值在5~20倍标准差之间的粗差,形成多维粗差,图1为粗差加入前后的误差方案示意图。因为粗差在直线的单侧密集分布,最容易使估计方法崩溃[22],所以本文设置粗差点均在直线上方模拟该最坏情况。

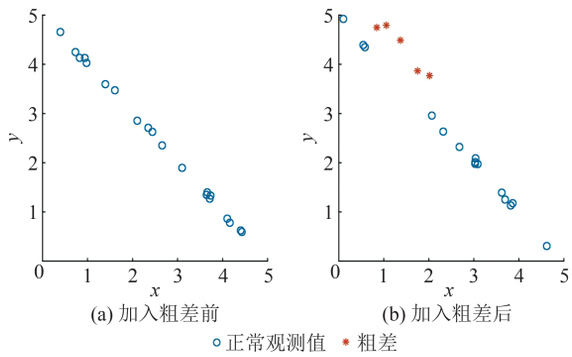

图1 两种误差方案示意图

Fig. 1 Diagram of Two Error Schemes

设置收敛阈值为1×10^{-10},分别在粗差加入前后,采用以下6种估计方法估计直线参数:

方法1:mixed LS-TLS。

方法2:以mixed LS-TLS结果为初值的RWTLS[7]。

方法3:WTLMS,对应§2.4.1。

方法4:WTLTS,对应§2.4.2,其中设置$h=11$以实现最大的崩溃污染率。

方法5:以WTLMS结果为初值的RWTLS。

方法6:以WTLTS结果为初值的RWTLS。

方法5和方法6可以视为EIV模型下的M估计[23],其要求M估计的待估参数和单位权中误差初值具有较高的可靠性。模拟$N=100$组实验,用$\hat{a}_i^j(i=1,2)$表示参数$\hat{a}_i$的第j次估计结果。表1和表2列出不同方案单位权中误差$\hat{\sigma}_0$、斜率和截距的统计指标,包括均值$\bar{a}_i$、标准差σ_{a_i}、均方根误差δ_{a_i},计算公式为:

$$\bar{a}_i=\frac{1}{N}\sum_{j=1}^{N}\hat{a}_i^j \tag{17}$$

$$\sigma_{a_i}=\sqrt{\frac{1}{N-1}\sum_{j=1}^{N}\left(\hat{a}_i^j-\bar{a}_i\right)^2} \tag{18}$$

$$\delta_{a_i}=\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\hat{a}_i^j-\tilde{a}_i\right)^2} \tag{19}$$

在只含正态随机误差的情况下,mixed LS-TLS是目前最优的估计方法。以该方法作为对比,定义算法的数值有效性[24]:

$$\varepsilon_{a_i}=\left(\sigma_{a_i}^0\right)^2\Big/\sigma_{a_i}^2 \tag{20}$$

式中,$\left(\sigma_{a_i}^0\right)^2$和$\sigma_{a_i}^2$分别为mixed LS-TLS估计方法和其他估计方法得到的方差。有效性衡量其他估计方法与最优的估计方法之间的一致性,不同

表1　加入粗差前参数估计值的统计结果

Table 1　Statistics of Estimated Parameters Without Outliers

方法	$\bar{a}_1$	σ_{a_1}	δ_{a_1}	$\varepsilon_{a_1}/\%$	$\bar{a}_2$	σ_{a_2}	δ_{a_2}	$\varepsilon_{a_2}/\%$	$\hat{\sigma}_0$
1	−0.999 58	0.008 12	0.008 13	100	4.999 53	0.023 28	0.023 28	100	0.040 15
2	−0.999 14	0.008 82	0.008 87	84.0	4.999 09	0.026 68	0.026 69	76.1	0.039 88
3	−1.001 12	0.018 11	0.018 15	20.1	4.997 66	0.050 20	0.050 25	21.5	0.038 97
4	−1.001 38	0.015 54	0.015 58	27.2	5.006 15	0.045 04	0.045 08	26.7	0.039 06
5	−0.999 32	0.009 32	0.009 34	76.1	4.999 45	0.028 36	0.028 37	67.4	0.040 36
6	−0.999 09	0.008 90	0.008 94	82.7	4.999 11	0.026 66	0.026 67	76.3	0.039 01

表2　加入粗差后参数估计值的统计结果

Table 2　Statistics of Estimated Parameters with Outliers

方法	$\bar{a}_1$	σ_{a_1}	δ_{a_1}	$\bar{a}_2$	σ_{a_2}	δ_{a_2}	$\hat{\sigma}_0$
1	−1.108 74	0.035 78	0.114 48	5.442 08	0.125 93	0.459 67	0.190 42
2	−1.042 09	0.061 01	0.074 12	5.164 68	0.237 22	0.288 78	0.056 53
3	−1.000 38	0.020 75	0.020 77	5.001 73	0.057 73	0.057 76	0.041 22
4	−1.001 65	0.017 68	0.017 74	5.006 95	0.048 70	0.048 73	0.040 63
5	−1.001 67	0.013 46	0.013 56	5.006 97	0.046 04	0.046 57	0.039 36
6	−1.002 78	0.014 28	0.014 55	5.010 79	0.050 09	0.051 24	0.041 69

算法的效率如表1所示。

由表1得到如下结论:(1)当观测值只含随机模型时,mixed LS-TLS估计在6种方法中获得最优的估计结果;(2)其他方法的参数均值和单位权中误差与mixed LS-TLS估计差异并不显著,但均方根误差和标准差略有增加。它们的标准差几乎等于均方根误差,表明估计偏差小到可以忽略不计;(3)对比各种方法的有效性,WTLMS和WTLTS估计的数值有效性分别约为20%和25%,这是两种算法在子集中采用排序方法求解参数估计值,没有使用全部的观测信息。方法5和6的有效性均大于65%,所以本文方法更适合提供初值,结合M估计能够显著提高算法有效性。

对表2进行分析可得:(1)当观测值中存在粗差时,mixed LS-TLS得到的参数估值严重失真,表明mixed LS-TLS不具备稳健性;(2)RWTLS通过对可疑粗差进行降权,一定程度上改善了mixed LS-TLS的表现,但可靠性较差,这是因为M估计迭代结果依赖于参数初值,而由于粗差影响,mixed LS-TLS的参数初值与真值偏差较大;(3)WTLMS和WTLTS的估值和真值差别很小,它们的标准差和均方根误差与表1相比变化不大,表明本文算法的崩溃污染率优于RWTLS估计。方法5和方法6的迭代初值更加准确,相较于方法2,估值更加接近于真值;(4)总体而言,受到粗差的影响,各种方法的标准差和均方根误差均有所增加。综上所述,RWTLS估计在25%污染率的数据中,无法得到可靠的估值,本文给出的两种算法能够得到可靠的估值,同时,结合M估计可以进一步提高估计有效性。

为了验证本文算法具有高崩溃污染率,在上述实验的基础上,加入不同数量的粗差,污染率从5%增加到45%。重复实验100次对比稳健方法2~4的表现,得到结果如图2所示。

由图2可知:(1)在均值指标上,当污染率大于25%时,RWTLS估计得到的参数均值偏差十分明显,而WTLMS和WTLTS估计的参数均值没有受到显著影响;(2)RWTLS在污染较小的情况下,参数的标准差比本文方法小,但是当污染率大于20%时,标准差迅速变大,同时出现明显偏差。WTLTS和WLTMS的标准差一直保持稳定,其中后者的标准差更小,但总体而言两者差异并不明显;(3)本文给出的两种算法的均方根误差随着污染率增加出现小幅度上升,并没有显著的偏差。综上所述,RWTLS在高污染率的数据中不能得到可靠的参数估值,本文算法具有接近50%的崩溃污染率。

3.2　点云数据

上述实验中的观测点具有相同的随机模型,以下通过真实的点云数据进行平面拟合,探究不同精度的观测值下不同算法的表现。选择来自WHU-TLS数据集[25]中校园场景的点云数据,如图3所示。采用RIEGL VZ-400激光扫描仪,其

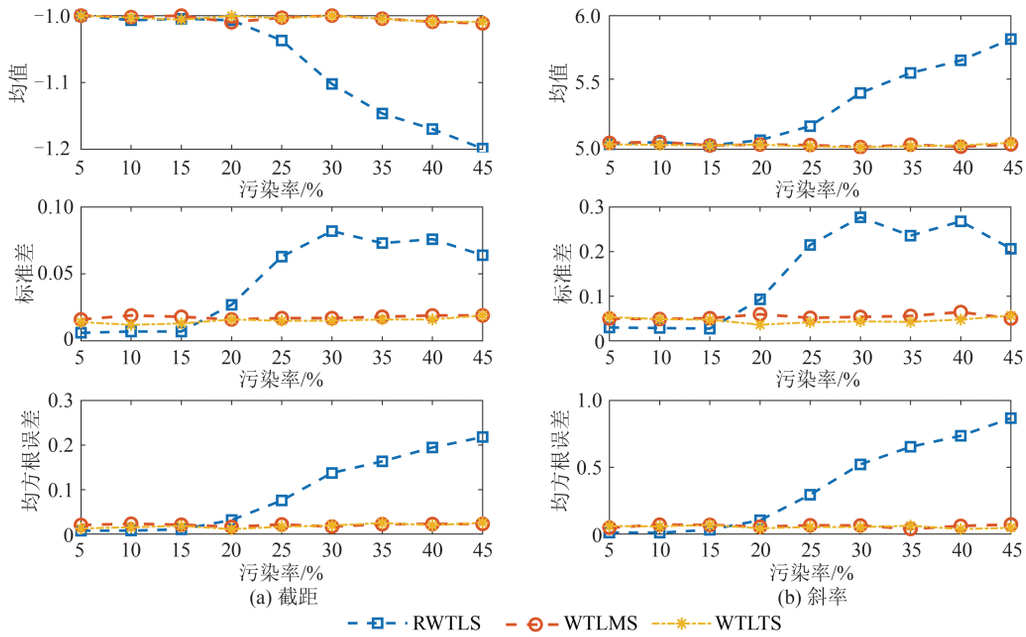

图2 不同污染率下算法的统计指标

Fig. 2 Statistics of Estimated Parameters Under Different Contamination Rates

测程为600 m,扫描范围100°×360°。截选其中一部分数据(图3红框所示,总计3 561个点)进行平面拟合,因为树木遮挡和其他原因,数据中含有大约20%的粗差。对该数据使用方法1~4进行平面拟合,其中观测值采用入射角定权法[26],结果如图4所示。因为点云数据量大,WTLMS估计最大重采样次数为N_r=1 000。

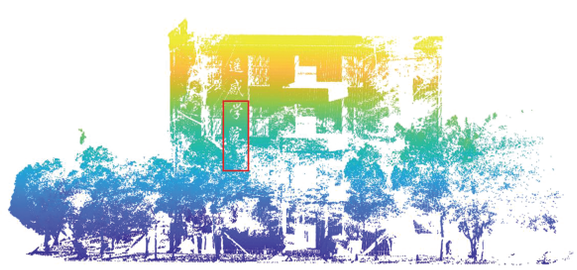

图3 校园场景扫描图

Fig. 3 Scanning Results of the Campus

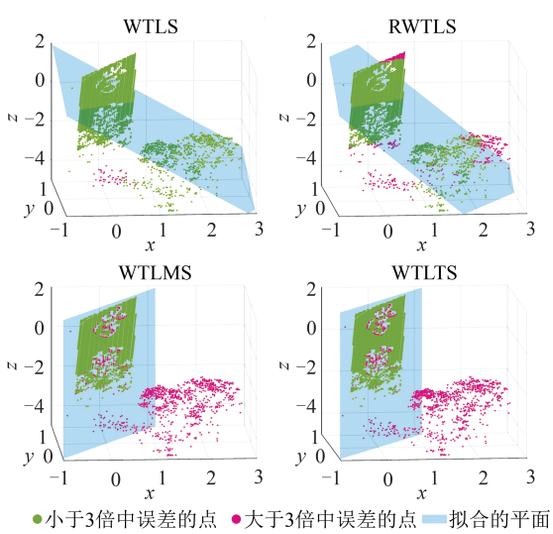

图4 4种方法拟合结果

Fig. 4 Fitted Results by 4 Methods

由图4可知:(1)从平面拟合结果上看,mixed LS-TLS受到粗差的影响,估计结果严重失真,RWTLS虽然具备一定的稳健性,但在实验中也出现明显的偏差,WTLMS和WTLTS得到理想的拟合平面,表明本文算法的崩溃污染率相较于方法1和方法2更高;(2)计算4种方法的单位权中误差,分别为0.789 6 m、0.517 3 m、0.006 1 m和0.005 3 m,可以看出方法1和方法2的拟合精度较差。将加权残差大于3倍中误差的点标记为红色,可以看出方法1和方法2并没有区别出正常观测值和粗差,而本文方法除了成功将所有的树叶数据识别粗差,还识别出墙面上一些粗差,验证了本文算法的可靠性。综上所述,本文算法在不等权的点云数据中也能得到可靠的平面拟合结果。

4 结 语

M估计的优势在于其较高的计算效率和简单的方法原理,然而受多维粗差、杠杆观测权等因素影响,其崩溃污染率并不高,因此本文提出两种适用于线性回归加权EIV模型的具有高崩溃污染率的估计算法,即WTLMS估计和WTLTS估计。本文分析两种算法的等变性和验

证它们具有50%的渐进崩溃污染率。因为本文算法的目标函数不可导,求解复杂程度相对M估计要高。通过实验验证两种算法具有良好的稳健性,可以以此作为初值结合M估计进一步提高算法的有效性。如何提高两种算法的计算效率和有效性成为下一步研究的方向。

参 考 文 献

[1] 刘经南,曾文宪,徐培亮. 整体最小二乘估计的研究进展[J]. 武汉大学学报(信息科学版),2013,38(5):505-512.
LIU Jingnan, ZENG Wenxian, XU Peiliang. Overview of Total Least Squares Methods[J]. *Geomatics and Information Science of Wuhan University*, 2013, 38(5): 505-512.

[2] 鲁铁定,陶本藻,周世健. 基于整体最小二乘法的线性回归建模和解法[J]. 武汉大学学报(信息科学版),2008,33(5):504-507.
LU Tieding, TAO Benzao, ZHOU Shijian. Modeling and Algorithm of Linear Regression Based on Total Least Squares[J]. *Geomatics and Information Science of Wuhan University*, 2008, 33(5): 504-507.

[3] WANG B, YU J, CHEN Y, et al. Efficient and Robust Solution to Universal Symmetric Transformation for 3-D Point Sets[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 1-15.

[4] 官云兰,刘绍堂,周世健,等. 基于整体最小二乘的稳健点云数据平面拟合[J]. 大地测量与地球动力学,2011,31(5):80-83.
GUAN Yunlan, LIU Shaotang, ZHOU Shijian, et al. Robust Plane Fitting of Point Clouds Based on TLS[J]. *Journal of Geodesy and Geodynamics*, 2011, 31(5): 80-83.

[5] AMIRI-SIMKOOEI A R, JAZAERI S. Data-Snooping Procedure Applied to Errors-in-Variables Models[J]. *Studia Geophysica et Geodaetica*, 2013, 57(3): 426-441.

[6] WANG B, LI J C, LIU C. A Robust Weighted Total Least Squares Algorithm and Its Geodetic Applications[J]. *Studia Geophysica et Geodaetica*, 2016, 60(2): 177-194.

[7] 龚循强,李志林. 稳健加权总体最小二乘法[J]. 测绘学报,2014,43(9):888-894.
GONG Xunqiang, LI Zhilin. A Robust Weighted Total Least Squares Method[J]. *Acta Geodaetica et Cartographica Sinica*, 2014, 43(9): 888-894.

[8] 龚循强,李志林. 一种利用IGGII方案的稳健混合总体最小二乘方法[J]. 武汉大学学报(信息科学版),2014,39(4):462-466.
GONG Xunqiang, LI Zhilin. A Robust Mixed LS-TLS Based on IGGII Scheme[J]. *Geomatics and Information Science of Wuhan University*, 2014, 39(4): 462-466.

[9] 刘超,王彬,赵兴旺,等. 三维坐标转换的高斯-赫尔默特模型及其抗差解法[J]. 武汉大学学报(信息科学版),2018,43(9):1320-1327.
LIU Chao, WANG Bin, ZHAO Xingwang, et al. Three-Dimensional Coordinate Transformation Model and Its Robust Estimation Method Under Gauss-Helmert Model[J]. *Geomatics and Information Science of Wuhan University*, 2018, 43(9): 1320-1327.

[10] 李思达,柳林涛,刘志平,等. 多变量稳健总体最小二乘平差方法[J]. 武汉大学学报(信息科学版),2019,44(8):1241-1248.
LI Sida, LIU Lintao, LIU Zhiping, et al. Robust Total Least Squares Method for Multivariable EIV Model[J]. *Geomatics and Information Science of Wuhan University*, 2019, 44(8): 1241-1248.

[11] 陶叶青,高井祥,姚一飞. 基于中位数法的抗差总体最小二乘估计[J]. 测绘学报,2016,45(3):297-301.
TAO Yeqing, GAO Jingxiang, YAO Yifei. Solution for Robust Total Least Squares Estimation Based on Median Method[J]. *Acta Geodaetica et Cartographica Sinica*, 2016, 45(3): 297-301.

[12] ROUSSEEUW P J, LEROY A M. Robust Regression and Outlier Detection[M]. New Work: Wiley, 1987.

[13] FANG X, ZENG W X, ZHOU Y J, et al. On the Total Least Median of Squares Adjustment for the Pattern Recognition in Point Clouds[J]. *Measurement*, 2020, 160: 107794.

[14] LÜ Z P, SUI L F. The BAB Algorithm for Computing the Total Least Trimmed Squares Estimator[J]. *Journal of Geodesy*, 2020, 94(12): 110.

[15] ZHOU Y, FANG X. A Mixed Weighted Least Squares and Weighted Total Least Squares Adjustment Method and Its Geodetic Applications[J]. *Survey Review*, 2016, 48(351): 421-429.

[16] FANG X. Weighted Total Least Squares: Necessary and Sufficient Conditions, Fixed and Random Parameters[J]. *Journal of Geodesy*, 2013, 87: 733-749.

[17] HAWKINS D M. The Feasible Set Algorithm for Least Median of Squares Regression[J]. *Computational Statistics & Data Analysis*, 1993, 16(1): 81-101.

[18] RUPPERT D. Computing S Estimators for Regression and Multivariate Location/Dispersion[J]. *Journal of Computational and Graphical Statistics*, 1992, 1(3): 253.

[19] ROUSSEEUW P J, VAN DRIESSEN K. A Fast Algorithm for the Minimum Covariance Determinant Estimator[J]. *Technometrics*, 1999, 41(3): 212.

[20] YANG L, SHEN Y Z, LI B F. M-Estimation Using Unbiased Median Variance Estimate[J]. *Journal of Geodesy*, 2019, 93(6): 911-925.

[21] 刘春阳, 王坚, 王彬, 等. 基于中位参数法相关观测的抗差加权整体最小二乘算法[J]. 武汉大学学报(信息科学版), 2019, 44(3): 378-384.
LIU Chunyang, WANG Jian, WANG Bin, et al. Robust Weight Total Least Squares Algorithm of Correlated Observation Based on Median Parameter Method[J]. *Geomatics and Information Science of Wuhan University*, 2019, 44(3): 378-384.

[22] XU P L. Sign-Constrained Robust Least Squares, Subjective Breakdown Point and the Effect of Weights of Observations on Robustness[J]. *Journal of Geodesy*, 2005, 79(1): 146-159.

[23] YOHAI V J. High Breakdown-Point and High Efficiency Robust Estimates for Regression[J]. *The Annals of Statistics*, 1987, 15(2): 642-656.

[24] MARONNA R A, MARTIN R D, YOHAI V J. Robust Statistics[M]. New Work: Wiley, 2006.

[25] DONG Z, YANG B S, LIANG F X, et al. Hierarchical Registration of Unordered TLS Point Clouds Based on Binary Shape Context Descriptor[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018, 144: 61-79.

[26] 苍桂华, 李明峰, 岳建平. 以入射角定权的点云数据加权总体最小二乘平面拟合研究[J]. 大地测量与地球动力学, 2014, 34(3): 95-98.
CANG Guihua, LI Mingfeng, YUE Jianping. Study on Point Clouds Plane Fitting with Weighted Total Least Squares Based on Incidence Angle Weighting[J]. *Journal of Geodesy and Geodynamics*, 2014, 34(3): 95-98.