Advances in Applied Mathematics 应用数学进展, 2020, 9(9), 1358-1363
Published Online September 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.99161

Hans汉斯

基于信息扩散函数定权的抗差估计及其应用

黄康钰[1]，张　俊[1*]，李屹旭[2]

[1]贵州大学矿业学院，贵州　贵阳
[2]贵州大学农学院，贵州　贵阳
Email: *jzhang512@126.com

收稿日期：2020年8月3日；录用日期：2020年8月25日；发布日期：2020年9月2日

摘　要

利用最小二乘估计的观测向量标准化残差计算信息扩散函数值并构造权阵进行抗差估计。该方法由于事先根据标准化残差估计了残差的实际分布，从而无需任何迭代过程即可获得包含粗差信息的观测值的可靠权阵。试验结果表明：当观测向量中包含粗差时，利用信息扩散函数计算的相应观测量的权值相对较小，从而抑制了粗差的影响。与IGG方案以及拟准检定法相比，新方法仍能取得较好抗差估计结果。

关键词

粗差，标准化残差，信息扩散，抗差估计，拟准检定法

Robust Estimation Based on Weighting of Information Spreading Model and Its Application

Kangyu Huang[1], Jun Zhang[1*], Yixu Li[2]

[1]College of Mining, Guizhou University, Guiyang Guizhou
[2]College of Agriculture, Guizhou University, Guiyang Guizhou
Email: *jzhang512@126.com

Received: Aug. 3rd, 2020; accepted: Aug. 25th, 2020; published: Sep. 2nd, 2020

Abstract

A robust estimation method based on weighting with the information diffusion function value being calculated by the standardized residuals of the observation vectors of least squares estimation is proposed. Compared with the classical robust estimation methods based on mean shift model or va-

文章引用: 黄康钰, 张俊, 李屹旭. 基于信息扩散函数定权的抗差估计及其应用[J]. 应用数学进展, 2020, 9(9): 1358-1363. DOI: 10.12677/aam.2020.99161

riance expansion model, the new method estimates the actual distribution of the residual in advance according to the standardized residual, so that the reliable weight matrix of the observed value containing gross error information can be obtained without any iteration process. The experimental results show that when gross errors are included in the observation vectors, the weights of the corresponding observations calculated by the information diffusion function are relatively smaller than those of others; thus the effect of gross errors is weakened. Compared with IGG schemes and quasi-accurate detection method, the new method can still achieve better robust results.

Keywords

Gross Error, Standardized Residuals, Information Diffusion, Robust Estimation, Quasi-Accurate Detection Method

1. 引言

无论是传统模拟测量还是现代数字化测量技术采集的数据，粗差都在所难免，而当平差系统含有粗差时，经典最小二乘估计由于不具备抵抗粗差的能力而使其应用受到极大限制。针对粗差处理，自上世纪 60 年代以来，提出了很多方法。归纳起来，这些方法大致可分为三类，第一类是均值漂移模型[1] [2]，其思想是将观测向量看作是随机变量，当观测向量中某些观测值包含粗差时，样本均值会发生扭曲，从对结果的影响上来看，可解释为将含粗差的观测值看作是与正常观测值有相同方差和不同期望；第二类是方差膨胀模型[3]-[8]，这类模型是将包含粗差的观测值视为与正常观测值具有相同期望和不同方差来处理，这类方法已经由最初仅能处理独立观测值发展到也能处理相关观测[9]；第三类是我国学者欧吉坤教授提出的拟准检定法，该方法以观测值的真误差为研究对象，通过一套独特的选择拟准观测的实施办法，在附加拟准观测的真误差的范数极小的条件下，求解关于真误差的秩亏方程，根据真误差估值的分布特征来判别粗差，定位粗差，然后改正含粗的观测值[10]，这种方法后来被证明在一定条件下与选权拟合法具有等价性[11]。以上方法尽管从提出至今发展已经非常成熟，但它们都有一个共同特点，那就是要么经过不断重复假设检验过程，逐步定位剔除或修正含粗差观测值，要么经过选权迭代过程削弱粗差影响，而且当样本值较少或粗差占比较大时，其结果的可靠性会有所下降。为此，本文尝试将信息扩散理论引入测量平差，利用信息扩散函数计算最小二乘估计残差的实际分布，将残差向量的概率值近似为相应观测值的权，然后再次进行平差计算，试验表明，这是一种非常高效的抗差估计方法。

2. 信息扩散原理及其估计

2.1. 信息扩散原理

设母体 $\boldsymbol{K}$ 的概率密度函数为 $f(x)$，$\boldsymbol{W}$ 为给定的来自母体 $\boldsymbol{K}$ 的样本。当由 $\boldsymbol{W}$ 不能完全精确地认识 $f(x)$ 时，称 $\boldsymbol{W}$ 对 $\boldsymbol{K}$ 是非完备的。从概率统计的角度来看，这种不完备主要是因为样本集 $\boldsymbol{W}$ 中样本点数目太少，不足以反映 $\boldsymbol{K}$ 的全部特征。因此，可以通过增加样本数目来改善不完备程度。不难理解，当通过改善 $\boldsymbol{W}$ 使其趋于或达到可以完备描述 $\boldsymbol{K}$ 时，必定会经历一个由模糊到清晰的过程，即 $\boldsymbol{W}$ 从非完备到完备具有一种过渡趋势。当 $\boldsymbol{W}$ 非完备时，这种趋势表现在 $\boldsymbol{W}$ 的样本点上，就是每一个样本点都

有发展成多个样本点的趋势，使每一个样本点都充当“周围未出现之样本点的代表”[12]。若设 w_i 的观测值为 y_i，则 $\boldsymbol{W}$ 在 y_i 点提供的信息应可以被其周围点分享，而周围点分享的来自 y_i 信息量的多少与其属于 y_i 点周围的程度有关，显然，越靠近 y_i 的点分享的信息就越多。若记 y_i 点自身信息量为 1，则其周围的点从 y_i 获取的信息量介于 0 到 1 之间。这种 y_i 点的信息向其周围扩散的过程称为信息扩散过程，简称信息扩散。

2.2. 信息扩散估计

设 $\boldsymbol{W}=\{w_1,w_2,w_3,\cdots,w_n\}$ 是知识样本，Y 是基础论域。设 $x=\Phi(y-y_i)$，则当 $\boldsymbol{W}$ 非完备时，存在函数 $\mu(x)$，使点 y_i 获得的量值为 1 的信息可按 $\mu(x)$ 的量值扩散到 $y_j\,(j\neq i)$ 上去，且扩散所得到的原始信息分布 $Q(y)=\sum\mu\left[\Phi(y-y_i)\right]$ 能更好地反映 $\boldsymbol{W}$ 所在总体的规律。若设 $\mu(x)$ 为定义在 $(-\infty,+\infty)$ 上的一个波雷尔可测函数，$\Delta_n>0$ 为常数，则称

$$f(y)=\frac{1}{n\Delta_n}\sum_{i=1}^{n}\mu\left[\Phi\left(\frac{y-y_i}{\Delta_n}\right)\right] \tag{1}$$

为母体概率密度函数 $f(y)$ 的一个扩散估计，式中 $\mu(\cdot)$ 称为扩散函数，Δ_n 称为窗宽。

2.3. 信息扩散函数及窗宽确定

由(1)式知，实现母体概率密度函数 $f(y)$ 的扩散估计的关键是 $\mu(\cdot)$ 的具体形式难以确定，当考虑样本分布符合正态分布时，可借用分子扩散理论导出正态扩散函数为[12]：

$$\mu(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{2}$$

由此，将(2)式带入(1)式可得母体概率密度函数的正态扩散估计为

$$f(y)=\frac{1}{nh\sqrt{2\pi}}\sum_{i=1}^{n}\exp\left(-\frac{(y-y_i)^2}{2h^2}\right) \tag{3}$$

上式中 $h=\sigma\Delta_n$ 为窗宽，一般可根据择近原则导出的以下经验公式确定[12]：

$$h=\frac{\alpha(b-a)}{n-1} \tag{4}$$

其中 $a=\min(y_i)$，$b=\max(y_i)$，$i=1,2,\cdots,n$；a 是 n 的函数，具体随 n 取值变化情况参见文献[12]。当 $n\geq 17$ 时，取 $a=1.420693101$，这是因为测量中总是要求要有多余观测，$n\geq 17$ 一般都可以满足，若 $n<17$ 时，可根据样本数目，按文献[12]给出的相应数值取值。

3. 利用信息扩散函数计算观测值的权

由上文可知，信息扩散估计本身一般只适合于一维参数估计，且要求样本具有相同的数学期望，而测量中的参数估计几乎全是多维参数估计的情况，且常常因为观测量种类不同，数学期望既不相同也难以比较，这样就制约了信息扩散估计在测量数据处理中的应用。文献[13]以水准测量为例，探讨了利用一维信息扩散估计进行水准测量平差的问题，但至今对于更为普遍的多维测量平差问题仍然讨论较少。为解决这一问题，本文提出采用观测值最小二乘估计的标准化残差向量 V 作为扩散变量构造 V 的概率密度函数 $f(v)$ 去代替观测值 L 的概率密度函数 $f(l)$，近似估计母体概率密度函数 $f(l)$，然后根据 $f(l)$ 计算观测值 l 的权 p_l，据此再次进行常规最小二乘估计即可获得参数的抗差最小二乘解。具体可采用以下步骤进行：

1) 针对原始观测值，令观测权阵为单位阵(如有确切先验权阵，可采用先验权阵)建立经典高斯-马尔科夫模型，利用最小二乘估计获得观测值的标准化残差向量；

2) 利用式(2)和式(3)估计标准化残差 V 的概率密度函数 $f(v)$，即

$$f(v)=\frac{1}{nh\sqrt{2\pi}}\sum_{i=1}^{n}\exp\left(-\frac{(v-v_i)^2}{2h^2}\right) \tag{5}$$

3) 用标准化残差的概率密度函数代替观测值概率密度函数计算观测值的权 p_l，即

$$p_{l_i}=\frac{f(l_i)}{\sum_{i=1}^{n}f(l_i)}=\frac{f(v_i)}{\sum_{i=1}^{n}f(v_i)} \tag{6}$$

4) 利用步骤(3)中确定的权值再次进行最小二乘估计获得参数的抗差解 $\hat{X}$，即

$$\hat{X}=\left(B^{\mathrm{T}}PB\right)^{-1}B^{\mathrm{T}}PL \tag{7}$$

式中 B 为高斯－马尔科夫模型的系数矩阵，L 为观测向量，$P=\mathrm{diag}\left(p_{l_1},p_{l_2},\cdots,p_{l_n}\right)$。

4. 算例及分析

4.1. 算例描述及解算方案设计

本文算例来自文献[14]中例 7~9 三角网观测数据，网中共有 4 个已知点和 2 个未知点以及 18 个同精度角度观测值，已知数据及角度观测值见表 1 和表 2，图 1 为测角网示意图。为验证本文方法的合理性及其抗差效果，现分别在第 1、5、8、15 和第 16 等 5 个角度观测值上依次模拟大小为−7.0″、7.0″、−5.6″、5.6″、6.8″和−6.8″的粗差，并采用如下 4 种方案进行平差计算：

方案 1：采用模拟粗差后的 18 个观测值，按最小二乘进行平差处理；

方案 2：采用模拟粗差后的 18 个观测值，按 IGGIII 进行平差处理；

方案 3：采用模拟粗差后的 18 个观测值，按拟准检定法进行平差处理；

方案 4：采用模拟粗差后的 18 个观测值，按本文方法进行平差处理。

以上各方案计算结果见表 3，表 3 中 $\hat{X}=\left[\hat{x}_{\mathrm{P}_1},\hat{y}_{\mathrm{P}_1},\hat{x}_{\mathrm{P}_2},\hat{y}_{\mathrm{P}_2}\right]^{\mathrm{T}}$ 为未知点 P_1 和 P_2 的坐标改正数，$\left\|\Delta\hat{X}\right\|=\left\|\hat{X}-X_{\text{真}}\right\|$ 为各方案参数估值与参数真值(本文参数真值取文献[14]中不含粗差的正常最小二乘估计结果)差值的范数，可作为度衡量各方案参数估计的优劣指标。

Table 1. Starting data
表 1. 起算数据

点名	坐标(m)		边长	坐标方位角
	X	Y		′ ″
A	9684.28	43,836.82		
B	10,649.55	31,996.50	11,879.60	2,743,938.4
C	19,063.66	37,818.86	10,232.16	344,056.3
D	17,814.63		12,168.60	955,329.1
A			10,156.11	2,164,906.5

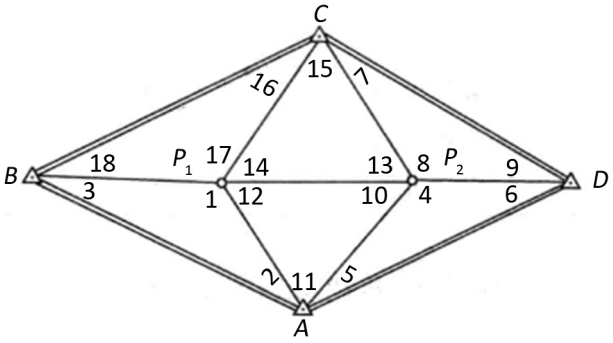

Figure 1. Schematic diagram of angle measuring control network
图 1. 测角控制网示意图

Table 2. Angle observation
表 2. 角度观测值

角度编号	观测值 ° ′ ″	角度编号	观测值 ° ′ ″	角度编号	观测值 ° ′ ″
1	1,261,424.1	7	220,243.0	13	463,856.4
2	233,946.9	8	1,300,314.2	14	663,454.7
3	300,546.7	9	275,359.3	15	664,608.2
4	1,172,246.2	10	655,500.8	16	295,835.5
5	312,650.0	11	670,249.4	17	1,200,831.1
6	311,022.6	12	470,211.4	18	295,255.4

Table 3. Parameter estimation results of each scheme
表 3. 各方案参数估计结果

采用方案	$\hat{X}(\mathrm{dm})$	$\lVert\Delta\hat{X}\rVert(\mathrm{dm})$
$X_{真}$	$[-0.1030, 2.3208, -1.2069, -0.5348]^{T}$	0
方案 1	$[0.1417, 1.6421, -1.4351, -0.8956]^{T}$	0.8384
方案 2	$[-0.3514, 1.7887, -0.9155, -0.5535]^{T}$	0.6558
方案 3	$[-0.1198, 1.9768, -1.1492, -0.6181]^{T}$	0.3590
方案 4	$[-0.0935, 2.3128, -1.0421, -0.5298]^{T}$	0.1654

4.2. 结果分析

1) 结果描述：从表 3 参数估计结果来看，方案 1 参数估计结果与真值相差甚远，其结果是扭曲的，说明最小二乘估计不具备抵抗粗差的能力；方案 2 和方案 3 与最小二乘估计结果相比，都一定程度的削弱了粗差对参数估计结果的影响，方案 4 参数估计结果与真值最接近，抗差效果最好。从方案 2、方案 3 和方案 4 参数估计与参数真值差值的范数分别等于 0.6558、0.3590 和 0.1654 来看，整体上抗差效果由好到差排列顺序依次为方案 4、方案 3 和方案 2；

2) 结果分析：方案 4 采用最小二乘估计残差标准差计算信息扩散函数值并利用式(6)近似计算观测值的权，相当于利用残差的真实分布定权，即便是某些观测值中包含粗差，其偏离正态分布的实际情况也

会在信息扩散函数中体现，也就是说通过信息扩散函数定权方法能够较好地顾及粗差影响；IGG 抗差估计核心是通过选权迭代过程实现的，其理想情况是若残差小于某一限值(本文取 1.5 倍验后中误差)则相应观测值的权不改变，若介于某一区间，则给予降权处理，若大于某一限值(本文取 2.5 倍验后中误差)，则对相应观测值给予零权处理(即删除不用)。然而，上述选权迭代过程中，观测值权值会受最小二乘估计“均摊”效应影响而失真，即具有相对较小残差的观测值未必不含粗差，反之大残差也未必真包含粗差。不仅如此，这种“均摊”效应在迭代过程中还可能发生转移，这使得观测值定权具有了不确定性，观测值的最终权值是多次“均摊”和“转移”的结果；拟准检定法结果优于 IGG 方案，主要是因为拟准检定法在迭代过程中采用真误差的分群特性确定拟准观测值，而真误差是没有误差的，相对于 IGG 方案减少了不确定因素，但拟准检定法仍然会受初始“拟准观测值”选择不准确的影响，意即尽管拟准检定法在迭代过程中利用真误差分群特征确定“拟准观测值”，但在实施之初也同样会受最小二乘“均摊”效应影响而无法准确确定初始拟准观测，而这种影响会延续至后续过程。

5. 结语

如何处理受到粗差污染的观测数据一直是测量平差的重要研究内容之一。本文首先利用最小二乘估计的观测值标准化残差计算信息扩散函数，解决了多维不同类观测量难以利用信息扩散估计进行数据处理的难题，然后利用标准化残差的信息扩散估计值构造观测值的权阵再进行最小二乘平差，结果表明这是一种成功的抗差估计方法：与 IGG 方案以及拟准检定粗差处理方法相比，信息扩散估计根据残差实际分布定权，较好地避免了最小二乘估计残差的“均摊”效应影响，不仅可以起到良好的抗差作用，而且还可以一定程度地保留含有粗差的观测值信息，并且不需要任何迭代过程，是一种良好的抗差估计方法。

基金项目

贵州省科学技术基础研究计划项目(黔科[2017]1054)；国家自然科学基金项目(41701464)；贵州大学引进人才科研项目(贵大人基合字(2016)51 号)；贵州大学测绘科学与技术研究生创新实践基地建设项目(贵大研 CXJD[2014]002)。

参考文献

[1] Baarda, W. (1968) A Testing Procedure for Use in Geodetic Networks. *Publications on Geodesy. New Series*, Vol. 2, Netherlands Geodetic Commission, Delft.

[2] 李德仁. 误差处理和可靠性理论[M]. 北京: 测绘出版社, 1988.

[3] Huber, P.J. (1981) Robust Statistics. Wiley, New York. https://doi.org/10.1002/0471725250

[4] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., *et al*. (1986) Roubst Statistics: The Approach Based on Influence Functions. Wiley, New York.

[5] 周江文. 经典误差理论与抗差估计[J]. 测绘学报, 1989(2): 115-120.

[6] 黄维彬. 近代平差理论及其应用[M]. 北京: 解放军出版社, 1992.

[7] 杨元喜. 异常影响诊断与抗差估计[J]. 测绘通报, 1994(5): 34-36.

[8] 杨元喜. 自适应抗差最小二乘估计[J]. 测绘学报, 1996(3): 206-211.

[9] 杨元喜, 宋力杰, 徐天河. 大地测量相关观测抗差估计理论[J]. 测绘学报, 2002, 31(2): 95-99.

[10] 欧吉坤. 一种检测粗差的新方法——拟准检定法[J]. 科学通报, 1999(16): 1777-1781.

[11] 王爱生, 徐春艳. 选权拟合与拟准检定[J]. 测绘科学, 2010, 35(5): 142-143.

[12] 王新洲. 基于信息扩散原理的估计理论、方法及其抗差性[J]. 武汉测绘科技大学学报, 1999, 24(3): 240-244.

[13] 王新洲. 用信息扩散估计进行水准网平差[J]. 武汉测绘科技大学学报, 2000, 25(5): 405-408.

[14] 武汉大学测绘学院测量平差学科组. 误差理论与测量平差基础[M]. 第三版. 武汉: 武汉大学出版社, 2014.