DOI:10.13203/j.whugis1993.01.006

第18卷第1期 武汉测绘科技大学学报 Vol.18 No.1
1993年3月 Journal of Wuhan Technical University of Surveying and Mapping Mar.1993

测量平差抗差化和效率分析

黄 幼 才

摘 要 本文从充分利用可能得到的信息的观点出发,讨论了测量平差的抗差性和优效性及其他们之间的关系。阐述了抗差估计理论和经典估计理论的本质区别。提出了测量平差抗差化的基本原则,并对几种估计的抗差能力和效率作了比较。

关键词 抗差估计;污染分布;有偏;*Fisher* 一致;

1 前 言

广而言之,测量平差是根据从观测值中得到的关于未知参数信息,通过某种统计模型,计算出参数的最优估值。未知参数包括定位参数和尺度参数。在传统的平差理论中,他们分别代表均值和描述数据离散情况的方差。众所周知,最小二乘估计理论是测量平差理论的核心。这种估计基于测量数据来源于正态分布的假设、实践证明,完全符合正态分布的观测数据是不可能得到的。因此,经典的平差方法存在着两个问题:1,不具备抗拒异常值(或粗差)干扰的能力;2,参数估值不具备最优特性(至多是近优的),甚至平差结果是错误的。实际上从观测值中获得的信息不可避免地受到外界不利因素的干扰。笔者认为,按信息所具有的统计性质而言,可将信息分为值的大小和他们排列的位置。按信息的质量来分,可以将信息分为有效信息,可利用信息和有害信息(粗差)。所谓有效信息是指我们能准确知道这部分观测数据分布模式、采用相应的极大似然估计,获得最优估值。可利用信息是指那些服从对称分布的观测数据,虽然这些观测数据的准确分布不知道,但仍然可为提高估值精度作出贡献。抗差估计的基本原则是充分利用有效信息,限制利用可用信息(避免尺度参数估值有偏),排除有害信息。采用了有害信息和不适当的排除了一些有效的或可用信息都会降低估值的效率,特别是前者。困难在于不太可能准确地知道观测值中有害信息和有效信息所占的比例以及他们具体代表的观测值。因此,抗差估计必须是:冒着损失一些效率(指未充分利用有效信息和可用信息)风险,获得可靠的、具有实际意义的最有效估计。由于文章的篇幅有限,本文只讨论测量平差中的直接观测平差,但理论的结论完全可用于间接观测平差。

2 测量平差中两种基本模型及存在的问题

测量平差中广泛采用了两种估计:L_2 估计(最小二乘)和 L_1 估计(一次范数最小,在一维

收稿日期:1992-07-09

的情况下称中位数)。他们分别是正态分布和拉普拉斯分布的极大似然估计。L_1 估计是分位数中的特殊情况(取顺序统计量的中间),它只利用了观测数据中的排序信息,观测数据的大小对这种估计的估值影响不大,因而这种估计天然地具有抗差性。但如果观测数据是有效信息,则这种估计由于没有利用数据的大小信息而效率降低。例如正态数据采用 L_1 估计,其相对效率仅为最小二乘估计的 64%。此外,线性规划法是解算 L_1 估计的最广泛的方法之一。它是从可行集中找出使目标函数为最小的可行解,因而这种方法得出的解是 C_n^m 个解的最优的一个。但因每个可行解没有接纳子样中一切可利用的信息,就整体而言,这种解不是最有效的(虽然如此,线性规划法不失为解算 L_1 的最好方法之一)。对于间接观测平差,存在设计空间和观测空间两方面的抗差问题,L_1 对设计空间不具备抗差能力。

最小二乘估计在假设观测数据来源于正态分布的前提下,集纳了观测值中的全部信息,包括有害信息,一旦观测值出现了粗差或异常值,这种估计的可靠性和效率大大降低,甚至结果完全是错误的。

综上所述,统计学家想到了一种折衷的方法,把上述两种方法"有效"地结合起来,即在数据中部采用 L_2 法,两尾采用 L_1 法,构成了抗差估计最基本的模型——Huber 法。较好地解决了 Fisher 和 Eddington 1914 年关于 L_1 和 L_2 的争论。

3 测量数据分布模式的假设

抗差估计理论认为观测数据可能来自不同分布的母体。观测值中的异常值可能改变数据的分布结构或者看成来源于某种分布的母体。Tukey 于 1960 年提出了污染分布模式。

$$F_\varepsilon = (1-\varepsilon)F + \varepsilon H, 0 \leqslant \varepsilon \leqslant 1 \tag{1}$$

式中 F 是某种标准分布或称基础分布,H 是污染分布,ε 是污染率。基于测量数据大部分来源于正态分布的事实,式(1)可以写为

$$F_\varepsilon = (1-\varepsilon)\Phi + \varepsilon H, 0 \leqslant \varepsilon \leqslant 1 \tag{2}$$

Φ 是正态分布,是数据的主体。式(2)又称污染正态分布。显然,服从 Φ 分布的观测数据是有效信息。H 部分包含了可用信息和有害信息。如果 H 是对称分布,则是可用信息。如果准确地知道 H 的分布模式(不同于 Φ),则所有观测数据为有效信息,可分别采用相应的极大似然估计,综合起来得最优效估计。

4 抗差估计设计原理

估计理论的核心是求最优估值。抗差估计应达到三个条件:1,有效地排除了异常值的干扰;2,充分利用一切可利用的信息;3,估计量的方差最小。抗差估计可分为三种类型:M 估计(广义极大似然估计);L 估计(排序统计量线性组合估计);R 估计(秩估计)。结合测量数据实际,这里只讨论 M 估计。

4.1 定位参数 M 估计

传统的极大似然估计的目标函数可写为

$$\sum_{i=1}^{n}[-\ln f(x_i)] = \min \tag{3}$$

其中 f 是随机独立变量($x_1, x_2, \cdots x_n$)的密度，T_n 是参数的估计量。用函数 ρ 代替上式中的 $\ln f$ 得 M 估计的目标函数

$$\sum_{i=1}^{n}\rho(x_i, T_n) = \min \tag{4}$$

对 ρ 求导，得 M 估计另一种表达形式

$$\sum_{i=1}^{n}\psi(x_i, T_n) = 0 \tag{5}$$

于是，(4)式或(5)式定义了 M 估计。

假设观测数据的分布服从(2)式，且 H 是对称分布，则可得 Huber 估计的 ψ 函数

$$\psi_0(x) = \begin{cases} x, & |x| \leqslant c \\ c\,\mathrm{sign}(x), & |x| > c \end{cases} \tag{6}$$

或

$$\psi_0(x) = [x]_c^c \tag{7}$$

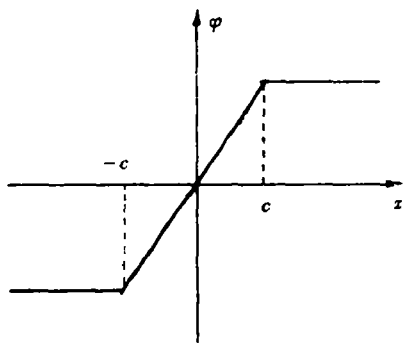

图 1 Huber 估计的 ψ 函数

其中 sign 是符号函数，ψ 函数的下标"0"表示基础分布 F_0，用图表示如图 1。

根据定义，M 估计属于极大似然估计，因此需要知道它的密度函数。由(4)式和(2)式可以写出

$$f_0(x) = (1-\varepsilon)(2\pi)^{-1/2}\exp(-\rho(x)) \tag{8}$$

式中

$$\rho(x) = \int_0^x \psi_0(t)\,dt$$

顾及(6)式，则有

$$f_0(x) = \begin{cases} (1-\varepsilon)\varphi(x), & |x| \leqslant c \\ (1-\varepsilon)(2\pi)^{-1/2}\exp\left\{\dfrac{c^2}{2} - c|x|\right\}, & |x| > c \end{cases} \tag{9}$$

式中 $\Phi(x)=2(\pi)^{-1/2}e^{-\frac{1}{2}x^2}$，其中 c 满足

$$2\Phi(c) - 1 + 2\varphi(c)/c = 1/(1-\varepsilon) \tag{10}$$

(10)式给出了 ε 和 c 之间的关系，即给定了 ε，可以计算出 c 值。因而也就建立了 F_ε 与 ψ 之间的关系。为了清楚理解这种关系，现对(10)式证明如下，如图 2 所示：

$$B = \int_c^\infty (1-\varepsilon)(2\pi)^{-1/2}e^{\left\{\frac{c^2}{2}-c|x|\right\}}dx = (1-\varepsilon)(2\pi)^{-1/2}e^{\frac{c^2}{2}}\int_c^\infty e^{-c|x|}dx$$

因为

$$(d/dx)(e^{-c|x|}) = e^{-c|x|}(-\,c\,\mathrm{sign}(x))$$

对于 $x>c$，有

$$(d/dx)(e^{-cx}) = e^{-cx}(-c)$$

$$\frac{-1}{c}d(e^{-cx}) = e^{-cx}dx$$

$$\int_c^\infty e^{-cx}dx = -1/c\int_c^\infty d(e^{-cx}) = -\frac{1}{c}e^{-cx}\Big|_c^\infty = \frac{1}{c}e^{-c^2}$$

于是

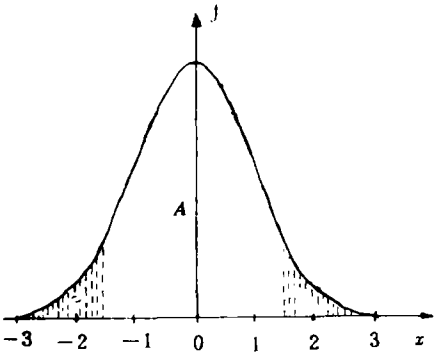

图 2　ε 和 c 关系图

$$B=(1-\varepsilon)(2\pi)^{-1/2}e^{\frac{c^2}{2}}\cdot\frac{1}{c}e^{-c^2}=(1-\varepsilon)\varphi(c)/c$$

$$A=\int_{-\infty}^{c}(1-\varepsilon)\varphi(x)dx-B=(1-\varepsilon)(\Phi(c)-\varphi(c)/c)$$

$$=(1-\varepsilon)(2\Phi(c)-1)$$

两部分和的概率应为 1，故有

$$(1-\varepsilon)(2\Phi(c)-1+2\varphi(c)/c)=1$$

分析 Huber 估计如何有效地利用信息：

1）A 部分只含有正态分布的观测值，保持原观测值不变，采用最小二乘估计。如果 B 部分的数据服从拉普拉斯分布，采用 L_1 估计（取平尾）。两部分均为极大似然估计，估值为最优，这是最佳的情况。

2）根据假设前提，H 是一个未知的对称分布。对 B 部分采用 L_1 估计，利用 H 分布中观测值的排序信息，提高估值的效率。两尾对称取平尾，不会产生有偏问题。

3）不排除 A 中包含有少量服从 H 分布的观测值，B 中含有少量的正态分布的观测值，这对采用 Huber 估计，不会引起定位参数估值偏估，可能影响方差估值的效率，但影响不大。

4）大量实测数据分析证明，带有少量粗差的正态的数据，其分布类似于拉普拉斯分布，所以对尾部采用 L_1 估计一般会获得较好的效果。

很明显，由(9)式定义的分布是分布族 F_ε（H 是对称分布）中最小信息分布，即 $I(F_0)=\inf\{I(F):F\in F_\varepsilon\}$，$I(F_0)$ 是 F_0 的 Fisher 信息。因(9)式中仅仅利用了 H 中的排序信息。另一方面，由 ψ_0 生成的 M 估计是 F_0 的极大似然估计，故有

$$\sup\sigma^2(\psi_0,F)=\inf\sup\sigma^2(\psi,F)\qquad(11)$$

其中 sup 表示上确界，inf 表示下确界。(11)式说明 Huber 估计符合极大极小原则。

现在讨论 H 是非对称分布的情况。对非对称分布数据采用 L_1 估计会引起偏估问题，被视为有害信息，应当完全排除。因此，ψ 函数应具有淘汰这部分信息的功能。Tukey 双权估计，Hampel 三段法估计，Andrews 正弦估计，IGG 估计等都属于这类型的估计。IGG 方案是周江文 1989 年提出的，其 ψ 函数表达式为

$$\psi(x_i,T_*)=\begin{cases}x_i-T_* & |x_i-T_*|<1.5\sigma\\ 1.5\sigma & 1.5\sigma\leqslant|x_i-T_*|<2.5\sigma\\ 0 & |x_i-T_*|\geqslant 2.5\sigma\end{cases}\qquad(12)$$

如图 3 所示，大于 2.5σ 的观测值被淘汰。

比较上述两种抗差方案可以看出：对 ψ 函数的假设越多，为了获得 M 估计的渐近正态性，则对 F 的限制越少。图 3 表明，IGG 的 ψ 函数是阶梯函数。对于 Huber，IGG 等这类淘汰型的抗差估计，在转折点处，左右导数不一样，应特别注意。此外，由于 ψ 不是单调递增的，在迭代计算中会产生解收敛于局部峰点，导致了多解和伪解。应采用适当的方法避免这个问题[3]。

4.2　尺度参数（方差）M 估计

定位参数抗差估计的处理方法要比尺度参数抗差估计要容易得多。对于对称分布的数据，就是对数据对称中心进行估计。无论采用 L_1 法对数据两尾取平尾，或是用淘汰法对称地截除

两尾部分都不会引起定位参数的偏估问题。对于尺度参数则不然，它没有像定位参数那样自然的几何对称中心。如果数据完全服从正态，如果我们设计的 ψ 函数对尾数取平尾或截除，则算出来的方差比它应有值要小，这就产生了估计不足的问题，即偏估问题。而方差大小主要决定于数据尾部的分布情况。因此，这里就引出了 M 估计应满足的一个重要条件，即 Fisher 一致条件，用式表达为

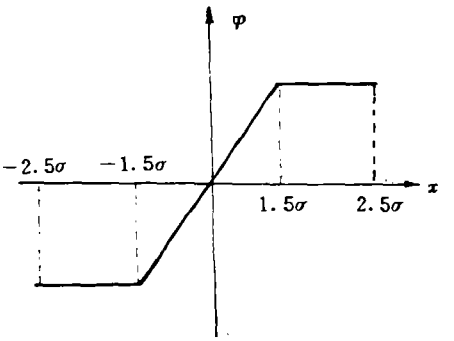

图 3 IGG 法

$$T(F_\theta) = \theta, \text{对于所有在 } \textcircled{H} \text{ 的 } \theta \qquad (13)$$

θ 是待估参数，$\textcircled{H}$ 是 θ 可能取值范围。对于测量数据，$\textcircled{H}=R$。(13)式的含义是：当观测值完全服从正态分布时，用抗差估计的方法所得的参数估值与最小二乘估值一致。以 Huber 估计为例，直接从(6)式引出尺度参数的 ψ 函数为

$$\chi_{(x)} = \begin{cases} x^2, & |x| \leqslant c \\ c^2, & |x| > c \end{cases} \qquad (14)$$

符号 $\chi(x)$ 表示尺度的 ψ 函数以区别定位参数。为了满足 Fisher 一致条件，需要在(14)式中加上改正数，得

$$\chi(x) = \begin{cases} x^2 - \beta, & |x| \leqslant c \\ c^2 - \beta, & |x| > c \end{cases} \qquad (15)$$

式中 $0<\beta<c^2$。$\beta=\beta(c)$ 由下式确定

$$\beta(c) = \int \min(c^2, x^2) d\Phi(x)$$

这样得到的尺度参数在 $F=\Phi$ 处是 Fisher 一致的，因为

$$\int \chi(x) d\Phi(x) = 0$$

这就是 Fisher 一致条件的具体表达式。传统的测量平差方法在 M 估计中称为截尾均值法(见图 4)，其未知参数的 ψ 函数表达式为

$$\psi(x) = \begin{cases} x, & |x| \leqslant c \\ 0, & |x| > c \end{cases} \qquad (16)$$

$$X(x) = \begin{cases} x^2 - \beta, & |x| \leqslant c \\ 0, & |x| > c \end{cases}$$

式中 $\beta=\beta(c)=\int_{-c}^{c} x^2 d\Phi(x)$

4.3 定位参数和尺度参数联合估计

最小二乘平差中均值和方差是分开计算的，这是因为假设观测误差是独立，方差为 σ^2 的正态随机变量，定位参数与尺度参数等完全独立。抗差估计理论假设观测值来源于不同的母体，观测值的方差也不相同。方差不统一对定位参数的估值产生有害影响，因此需要在计算定位参数的同时，也要计算尺度参数。然后利用尺度参数的计算值除观测值使之标准化，统一尺度后再进行平差计算，返复迭代直至收敛为止。

下面以 Huber 法，IGG 法和截尾均值法为例，分别计算他们在正态分布数据中的效率。取 c

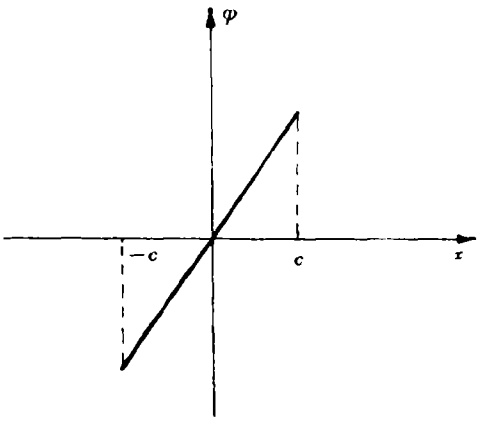

图 4　截尾均值法

$=1.5\sigma$，得 $e=2(1-\Phi(1.5))=0.1336$。

截尾均值的相对效率[2]

$$e=1-\frac{2k\varphi(k)}{M}=1-\frac{2\times 1.5\varphi(1.5)}{1-e}$$

$$=0.5515$$

Huber 估计的相对效率　　$e=0.9642$

从[3]，表 6-1 查得。

IGG 估计的相对效率

$$e=0.9642+(1-0.9977)-\frac{2\times 2.5\times\varphi(2.5)}{1-2(1-\Phi(2.5))}$$

$$=0.9665-\frac{2\times 2.5\times 0.01753}{1-2(1-0.99379)}$$

$$=0.8778$$

在正态的情况下，截尾均值的效率最低，IGG 方案的效率居第二位，Huber 估计的效率最佳。抗差能力的次序则相反。抗差能力和效率是为矛盾的两个方面，抗差估计的设计原则是两者达到最佳平衡。

5　判定观测数据分布的方法

前面提出的测量数据分布的基本模式，除假设数据主体是正态分布符合测量实际外，ε 和 H 都是未知的。平差质量取决于对 ε 和 H 的判定的准确程度。根据笔者的经验，下面简单地谈一下判定数据分布结构的几种方法。

1. 根据测量实际可以判定 ε 是一个很小的数，一般取 $\varepsilon=10\%$，对应的点是 1.65σ。

2. 对整个子样求均值和中位数，如果两者很接近，说明观测数据基本是对称的。否则先少量地对称截去两尾的观测值，再比较均值和中位数，直至两者之差很小为止。此时的截点就是淘汰点。

3. 比较科学的方法是采用[3]中的介绍的数据探测法。用 L_1 法对观测值进行前期平差，得抗差化余差。绘出各种余差图，根据数据探测理论判定余差的结构对称性，离异点等。用几种标准分布曲线与余差分布进行比较，最后确定观测值的分布。数据探测判定观测数据分布的准确性取决于余差是否真实地反映了观测误差。用 L_1 法得到的余差能有效地消除粗差的影响，但不能消除杠杆点对粗差的掩盖，使余差标准化可消除这种影响。

6　结　语

从抗差估计理论到测量平差模型抗差化需要一个理论过渡，内容繁多，本文不可能面面俱到。作者根据自己的研究，从信息观点出发概述了抗差估计基本原理以及实现测量平差抗差化途径。实践证明，不结合测量实际，简单地套用一些抗差估计公式往往效果不佳。由于篇幅有限，这里没有讨论抗差估计的计算和间接观测平差抗差化问题。

参 考 文 献

1 周江文. 经典误差理论与抗差估计. 测绘学报,1989.

2 李庆海,陶本藻. 概率统计原理和在测量中的应用. 北京:测绘出版社,1982,42

3 黄幼才. 数据探测与抗差估计. 北京:测绘出版社,1990.

4 中山大学数学力学系. 概率论及数理统计. 北京:人民教育出版社,1981.

5 Huber P J. Robust Statistice. Wiley,1981.

Analysis of Robustness and Efficiency in Surveying Data Adjustment

Huang Youcai

Abstract This paper deals with the robustness and efficiency of surveying data processing from the point of view of efficiently using information available and relationship between them. The essential difference between the robust estimation and classical estimation has been described. It also presents the principal criterion for robustizing the data adjustment and comparison between ressistance and efficiency with several examples.

Key words robust estimation; contaminated distribution bias; fisher consistency