

抗差估计理论 及其应用

杨元喜

八一出版社

前 言

抗差估计理论是一门年轻的学科，经过统计学界和应用科学界20多年的耕耘，已在许多领域显示出勃勃生机和强大生命力。她正以多姿的步伐逐渐跨出纯统计学界，并渗透到应用数据处理的各个学科。

本书致力于抗差估计在大地测量中的应用与发展。较系统地阐述了抗差估计的基本理论和方法；讨论了迄今为止测量平差中已归纳起来的常用函数模型和随机模型之抗差估计；详细推导了相应的抗差解法和影响函数；针对大地测量实际，发展了较严格的相关观测抗差估计理论，并设计了应用简便可行的相关IGG III方案；研究了参数具有先验特性的抗差Bayes估计、抗差滤波、抗差拟合推估；提出了M-LS、LS-M和M-M抗差估计法；分析了近几年发展起来的高崩溃污染率抗差估计的优缺点；对有界影响M估计作了较深入的分析，提出了较实用的有界影响抗差估计法；最后对抗差估计在大地测量中的应用潜力作了某些实际的和展望性的研究。

第一章作为引论，主要讨论了抗差估计的概念、任务；抗差估计的背景（模型偏差种种）；异常影响诊断和抗差估计的关系；以及抗差估计的发展史和主要成就。

第二章是后继章节的基础，主要陈述抗差估计的度量、准则，重点是影响函数及其导出量。此外，介绍了几种主要的抗差估计类，以及它们之间的关系；重点是介绍基于广义极大似然原理的M估计类。其中各估计量之间的关系是基于作者对它们形式上的理解而分析的。

第三章不完全属于作者的研究成果，只是作为抗差估计的基础而归纳介绍的。内容包括估计量的效率、渐近效率 (§3.1)、方差变化函数 (§3.2)；基于“有限粗差敏感度”的B抗差估计、基于“有限方差变化敏感度”的V抗差估计、以及基于这两种敏感度的上确界为最小的最大B抗差和最大V抗差估计 (§3.3)；同时也讨论了一定准则下的最优抗差估计 (§3.4) 和最优截尾抗差估计 (§3.5)。这一章的讨论全部基于M估计类。

第四章讨论了等价权原理——经典平差模型的抗差解。内容包括参数平差模型、条件平差模型、混合平差模型、自由网平差模型的抗差解 (§4.1-§4.4)；讨论了 ψ 函数的选择；给出了几种基本的 ρ 函数和 ψ 函数；评述了方差因子的种种抗差解；分析综合了验后协方差矩阵的抗差表式；重点推导了各种平差模型的基本抗差解式及影响函数。所给出的参数解式、影响函数、及验后协方差矩阵，都能在最小二乘的假设前提下复归至经典最小二乘估计。本章的算例只是用来说明抗差估计的基本能力。

第五章是本书的重点之一。主要讨论相关观测的抗差估计。众多的抗差估计和抗差统计学文献几乎都基于观测样本的独立同分布假设，而测量中存在大量的相关观测量。本章首先讨论相关最小二乘估计的误差影响；继而分别讨论由相关转换法、相关 ρ 函数法、相关等价权法定义的相关抗差估计；导出相应的影响函数以及验后精度表式等。所有解式也转换成相关最小二乘表示的形式，从而使相关抗差估计的实际应用成为可能。提出了适应于大地测量计算的IGG III方案；经过多种计算方案比较，IGG III方案均获得满意的抗差结果；§5.8顺便给出了相关条件平差和相关混合平差的抗差解及影响函数。

第六章讨论了抗差贝叶斯估计和抗差滤波。根据不同的误差分布模式，提出了M-LS、LS-M和M-M三种抗差贝叶斯原则；导出了相应的抗差解式、影响函数和验后协方差矩阵；基于第五章发展的相关抗差估计原理，又将抗差贝叶斯模型作了相关性扩

展,使得新的解式更广义化且更具实用性;通过一个实际算例证明了抗差贝叶斯估计在抵制粗差方面的能力;抗差滤波只是作为抗差贝叶斯估计的副产品提出的,附在本章的最后一节。

第七章是相关观测抗差估计和抗差贝叶斯估计的续篇。主要将近代平差理论中的拟合推估模型抗差化。也同样提出了M-LS、LS-M和M-M估计;给出了解式和影响函数;讨论了实际解算和协方差函数的抗差化问题。与第六章相似,抗差拟合推估也基于相关观测抗差估计理论作了扩展。

第八章主要介绍近几年统计学中才提出的高崩溃污染率的抗差估计。理论上作了适于测量平差模型的推演;并作了较详细的分析;通过对统计学中的高崩溃污染率估计法的推演和分析,得出的结论是,它们几乎都不能直接用于大地测量计算。

为了提出适于大地测量实际计算的高抗差性估值,第九章重点结合大地测量观测方案的设计与网的结构等问题,详细地讨论了观测误差和设计空间的影响,J矩阵、保差观测、解差函数等问题;提出了实用的有界影响抗差估计法;并给出了成功的算例。

第十章主要介绍自适应抗差估计理论和方法。首先介绍了最优抗差估计和观测误差分布的关系;进而讨论了自适应抗差M估计、方差因子的自适应估计和自适应抗差滤波。

第十一章是作为应用举例而写的。为应用方便,首先介绍了抗差M估计的计算方法,列出了各种 ρ 函数、 ψ 函数和权函数;基于影响函数讨论了大地网的三类设计新方法,使大地网的三类设计有了可靠的理论基础,即从抗差估计理论出发,要求大地网不仅有较高的内部符合精度,而且具有较好的可靠性,粗差不会被转移。拟稳抗差估计也是作为抗差估计的潜在应用而讨论的,在抗差估计理论的支持下,有可能将拟稳点的选择和粗差抑制同时得到较好的解决。空间网与地面网的联合平差也存在抗粗差和抗形变点问题,本章第五节讨论空间网与地面网联合平差的抗差

联解方案，给出了抗差联解方程和计算程序框图；§11.6 讨论了大地直角坐标系统的转换，提出了一种有效的高崩溃污染率抗差坐标转换法，并利用我国空间36个多普勒点和地面网同名点1954年北京坐标系坐标进行了多种方案的抗差计算分析；§11.7 讨论了垂线偏差的抗差内插法。

本书是在我国著名大地测量学家周江文教授的指导下完成的，写作过程中得到了熊介教授以及军事测绘杂志社、解放军测绘学院教务部的领导和同事们的鼓励、支持与帮助，在此，作者向他们及其他对作者有过帮助的同志们表示衷心的感谢。

由于作者水平有限，错谬之处实属难免，恳请读者批评指正。

作 者

目 录

前言

第一章 引论	(1)
§1.1 抗差估计概念及其任务.....	(1)
§1.2 模型偏差及其分类.....	(5)
§1.3 异常影响诊断与抗差估计.....	(7)
§1.4 抗差估计发展简史及主要成就.....	(13)
第二章 基本抗差估计类及其影响函数	(17)
§2.1 引言.....	(17)
§2.2 抗差性度量.....	(18)
§2.3 极大似然型估计 (M估计)	(23)
§2.4 顺序统计量线性组合估计 (L估计)	(27)
§2.5 秩检验型估计 (R估计)	(28)
§2.6 几种抗差估计之间的关系.....	(31)
§2.7 其它抗差估计类.....	(33)
第三章 最大、最优抗差估计	(36)
§3.1 估计量的效率.....	(36)
§3.2 方差变化函数.....	(40)
§3.3 B抗差性、V抗差性以及最大抗差估计.....	(43)
§3.4 最优抗差估计.....	(45)
§3.5 最优截尾M估计.....	(49)
第四章 等价权原理——经典平差模型的抗差解	(56)
§4.1 多维M估计.....	(56)
§4.2 参数平差、条件平差模型的抗差解.....	(59)

§4.3	混合平差模型的抗差解	(65)
§4.4	自由网伪逆抗差解	(68)
§4.5	关于 φ 函数的选择及几种基本的 ρ 函数和 φ 函数	(72)
§4.6	方差因子的抗差估计及评述	(78)
§4.7	验后协方差矩阵的抗差估计	(84)
§4.8	模拟计算	(90)
第五章	相关观测抗差估计	(92)
§5.1	相关最小二乘估计及误差影响	(93)
§5.2	相关观测转换法	(96)
§5.3	相关 ρ 函数法	(97)
§5.4	直接定义法	(101)
§5.5	相关等价权及相关IGG方案	(104)
§5.6	相关观测抗差估计的验后精度	(112)
§5.7	计算与分析	(117)
§5.8	其它平差模型的相关抗差解	(122)
第六章	抗差贝叶斯估计与抗差滤波	(125)
§6.1	概述	(125)
§6.2	抗差贝叶斯解式	(127)
§6.3	解算问题	(130)
§6.4	验后协方差阵的讨论	(131)
§6.5	算例与分析	(133)
§6.6	抗差贝叶斯估计的进一步扩展	(137)
§6.7	离散系统的抗差滤波	(142)
第七章	抗差拟合推估	(147)
§7.1	概述	(147)
§7.2	拟合推估M-M估计	(148)
§7.3	两个特例	(150)
§7.4	抗差拟合推估的解算及协方差函数	

的抗差问题·····	(151)
§7.5 实际计算与分析·····	(153)
§7.6 进一步扩展·····	(156)
第八章 高崩溃污染率和高效抗差估计理论 ·····	(162)
§8.1 引言·····	(162)
§8.2 二次残差中位数极小型估计的性质及其 算法·····	(165)
§8.3 均方差因子极小型估计·····	(169)
§8.4 高崩溃污染率与高效抗差估计·····	(174)
第九章 有界影响抗差估计 ·····	(182)
§9.1 观测误差对参数LS估计的影响·····	(182)
§9.2 设计空间的强影响与隐差·····	(190)
§9.3 设计空间的不完善对M估计的影响·····	(194)
§9.4 有界影响M估计·····	(197)
§9.5 实用有界影响M估计·····	(202)
§9.6 计算及比较·····	(209)
第十章 自适应抗差估计 ·····	(214)
§10.1 概述·····	(214)
§10.2 最优估计与误差分布·····	(216)
§10.3 未知参数的自适应抗差估计·····	(224)
§10.4 测量平差模型的自适应抗差估计·····	(232)
§10.5 均方差因子的自适应M估计·····	(237)
§10.6 自适应抗差滤波·····	(242)
第十一章 抗差估计的计算及其在大地测量中的 应用举例 ·····	(248)
§11.1 抗差M估计的主要计算方法·····	(248)
§11.2 各种 ρ 函数、 ψ 函数及权函数·····	(252)
§11.3 大地网的三类设计——整体影响函数法·····	(256)
§11.4 拟稳抗差估计·····	(261)

§11.5 空间大地网与地面网抗差联解方案·····	(266)
§11.6 坐标转换模型的抗差解法·····	(272)
§11.7 垂线偏差的抗差内插法·····	(284)
参考文献 ·····	(293)

第一章 引 论

§1.1 抗差估计概念及其任务

一、一般性描述

传统的、广为人们接受的最小二乘 (least squares) 估计理论, 在实践中有着广泛而成功的应用。当观测样本服从正态分布时, 参数的最小二乘估计具有无偏、一致、优效性。在泛函分析理论的支持下, 最小二乘估计又呈现出明晰、直观的几何意义和简捷的计算程式^{[105][10]}。但是, 当测值有悖于正态分布假设、样本遭受异常污染时, 最小二乘又具有明显的负面, 即估值不具有抗干扰性, 单个测值的偏差也可能导致解案面目全非。源于此, 一种迥异于最小二乘估计的新理论——抗差估计 (robust estimation) 应运而生。

抗差估计及其依为基础的抗差统计学, 不象最小二乘法那样过分地追求估值的优效性和无偏性等内部性质, 而是着力于估值的实际抗差性和可靠性。尽管抗差估计的问世才不过20多年, 引起广泛讨论只不过10多年, 但它已在许多领域显示出勃勃生机和强大的生命力。

抗差估计在测量界的引入与发展, 为测绘领域打开了一个广阔而深邃的研究天地。研究成果遍及大地、航测和工程测量。但是具有系统性和普适性的研究成果尚不多见。作为应用学科和基础学科并举的大地测量, 它有着不同于数理统计学的数据背景和模型方程。我们不能也不可能照搬抗差统计学的理论和定则, 而应致力于大地测量实际, 不断灌注抗差估计以丰沛的新内容; 建立

起相对独立且自成体系的理论和方法，演绎出适于应用且计算简便的解式；加强并拓宽这门新学科的外沿，探寻新的广阔的应用前景。

抗差估计与最小二乘估计各有千秋；它们只在特定的前提、特定准则下显示其优越性。不存在后者一致优于前者的结论，反之亦然。在抗差估计内部，各种方法也有优劣，它们各倚为根据的种种假设也不相同。显然，在研究过程中很难给出一个在所有准则下均为最优的抗差估计方法。但我们将本着“两害相较弃其重，两利相较取其大”的比较原则，以决定各种方法的应用范围。

二、抗差估计概念

“抗差估计” (robust estimation) 也称稳健估计，源于统计学中的抗差性 (robustness) 概念。在众多统计学文献中，有关抗差性的陈述并不相同。归纳起来，可作如下一般性描述。

一个估计方法当其依为根据的模型与实际模型有微小差异时，其估计方法的性能只受到微小的影响，即估计方法具有一定的“稳定性”。因若不然，则针对理论模型假设下的估值之优良性不仅没有实际意义，还可能导致使用者误入歧途。这是估计方法抗差性意义的一个方面。

估计方法抗差性的另一个方面的意义是当观测样本中混入少量粗差 (outlier) 时，估计量的数值受其影响不大，即估计方法具有一定的“抗干扰性”。因若不然，则极少数的观测粗差都可能导致估计量面目全非。

这种既能抵制模型偏差又能抗拒异常观测扰动的估值方法称为抗差估计法。

从一定意义上说，上述抗差性的两种含义是相通的。因为当数据受到粗差污染时，该数据可被视为来自另一个总体，其分布与原来模型所规定者不同。反之，模型偏差也可视作污染观测所为。

上述关于抗差性的描述是定性的、相对的。事实上不存在什

么“最抗差”的估计方法。所谓抗差性的优劣,往往是相对于某种范围和准则而言的,如方差 (variance) 准则、偏差 (bias) 准则、标准差 (mean-square error) 准则等,各种准则均有侧重。Tukey曾建议综合各类抗差准则,组成新的目标函数,以达到综合抗差之目的。^[124]

顺便指出,本文采用周江文教授提出的“抗差估计”一词而不采用“稳健”,其基本理由是,在数学学科中,稳健常被看成一组稳定性要求,类似于微分方程中的稳定性^[80]。我们研究“robustness”主要针对估值在模型分布与实际分布有差异时,估值所具有的抵制能力。既然如此,为了与数学上解方程的稳定性 (stability) 相区别,我们认为采用“抗差估计”更确切。

三、抗差估计的任务

在描述抗差估计概念时,曾把估计方法的抗差性理解为其性能对模型的微小变化不敏感,但是一个估值如果只具备这一特性,并不一定有用。例如,不管观测样本如何,总是用一个固定的常数去估计某参数,显然,样本中的异常值不可能影响该参数估值,但这类估值又有何意义!因此,抗差估计所追求的并非只具有抗差性,还应具有其它的良好性能。Huber曾提到抗差估计的三个主要目标:^[89]

(1) 在所假定的模型下,估值具有合理的优效性(最优或接近最优);

(2) 在实际模型与假设的模型有微小差异时,其估值或统计方法(如估值的渐近方差或检验量的检验功效等)所受的影响也较小;

(3) 在实际模型与假定模型有严重偏离时,其估值的性能仍能“过得去”或者说,不致使估值受到破坏性影响。

Hampel等提出了与之相类似的四个抗差估计目标:^[81]

(1) 估值应最优地拟合于观测样本(数据群体);

(2) 估值方法应能识别异常值;

(3) 对于不平衡设计空间, 估值方法应能识别强影响观测,

(4) 估值方法应能处理与假设相关结构有偏离的数据。

在其它文献中, 还有许多抗差估计目标。如高渐近相对效率 (relative efficiency) 或高绝对效率 (absolute efficiency) 等。但这些目标与前面提到的抗差目标相比要次要一些^[88]。

四、进一步说明

现就Hampel等的四个抗差估计目标作如下说明:

Hampel等的第一个抗差目标意味着, 当测值中含有少数异常值时, 其估值的受害应限制在一定范围内。这种少数异常值对估值影响的最大允许比率一般用“崩溃污染率” (breakdown point) 度量。当崩溃污染率达到50%时 (如样本中位数), 即可保证估值只取决于多数优质数据。

抗差估计的第二个目标是指抗差估值能自动显示出测值粗差。最小二乘估计常常使粗差转移或被掩盖 (smeared or masked), 而且基于残差算得的均方差也随之增大, 使得粗差探测愈加复杂。尽管在许多理想情况下, 经过仔细分析能够探测出粗差, 但是对于多维、非平衡设计、大样本空间, 最小二乘估计仍无法探测出所有粗差。有些常用的粗差探测准则并不可靠, 甚至使用最普通的学生化残差也只能探测出10%的粗差^[89]。抗差估计则要求能够解决多维、非平衡观测数据的粗差探测。

抗差估计的第三个目标提示人们, 当观测数据平衡, 没有极弱观测结构时, 可以找到非常好的抗差估计方法, 以便在理想模型下获得几乎最优效的估计量。但是当测值中有保差观测时, 仅基于残差的抗差估计法如Huber的M估计^[90]已不能完全达到抗差目的。对于保差观测的处理也应分成两种情况: 如果保差观测是正常测值, 则降低其权的抗差法必然损失效率; 如果保差观测含有粗差, 则不降权处理又必然影响估值的可靠性, 甚至使估值面目全非——崩溃。目前最好的方法只能是分别采取降权法和不降

权法处理保差观测，然后进行比较，如果两种方法的结果相差较大，则表明部分保差观测为异常观测值。显然对于保差观测的处理应十分谨慎。

在大地测量中，保差观测与观测方程的设计矩阵有关，设计矩阵又与观测方案有关，所以在进行粗差探测时，应充分利用设计空间的信息。

经典的粗差探测技术已经顾及到了设计空间和残差向量两个方面的影响，有时考虑两种影响的乘积^[6]，但是这些探测技术通常又是基于最小二乘原理而不是抗差估计原理，故由于隐差(masking)现象的不可避免，使得粗差探测变得相当棘手。

抗差估计的第四个目标告诉人们，长期的观测序列中可能存在相关现象。所有依据独立观测得出的参数估计，置信区间、统计检验可能都不可靠。

应该指出，Hampel等尽管认识到相关观测抗差估计的重要性，然而这方面的研究成果尚不多见。在大地测量、航空摄影测量中，即使是非时序观测也可能存在相关问题。对于这类相关观测的抗差估计问题将是本书要着力讨论的主要问题之一。

§1.2 模型偏差及其分类

在测量中，我们可以将实际模型与假设模型的偏差分为四类：即观测粗差(gross errors或blunders)；系统误差；模型误差；近似独立假设带来的偏差。

一、粗差

粗差主要是由失误引起的，一般以异常值或孤值形式出现。如记录错、计算错、仪器故障。通过认真检查和分析，这类粗差一般能够避免。但异常值有时不是粗差，它可能是由于测值来自不同的子样引起的。如监测网重复观测结果之间往往存在差异，如果将某个历元的部分观测误入其它历元的观测结果中，则会出

现异常值；重尾（long-tailed）分布也是异常值的来源之一，如实践中，人们总是假设观测母体为正态分布，但即使是高质量的观测结果，也不严格服从正态分布。Jeffreys曾统计了九组天文观测数据，每组500个观测值，经分析发现，其中七组为重尾正态分布^[51]。许多学者曾致力于粗差频率的研究，但众说纷纭、结论不一。Tukey认为在工程数据中约有10%的粗差；Cuthbert Daniel认为粗差频率一般在1%至20%之间；对于高质量的观测结果，Freedman等认为粗差约占0.5%至1%；Hampel等人的统计结果则表明粗差约占0.01%，并认为只要认真仔细地进行观测和预后分析，粗差频率可以降到0.01%以下。

测量中，观测方案的设计和实施必须遵循相应的“规范”。实施过程中一般均重复测量，且进行几何检核，故粗差频率一般很低。

二、系统误差

测量数据中的系统误差也是不可避免的。如大地测量中的区域旁折光误差、垂直折光引起的误差、仪器尺度差等。这些误差都使模型分布产生不同程度的偏移，导致参数估值的局部或整体漂移。

三、模型误差

理论上探讨一种估计方法的性质时，总是从一定的模型和假设出发。模型是帮助人们合理进行思考的工具。更具体地讲，所谓模型就是用一种简化的方式表现一个复杂的系统或现象^[52]。显然，模型是通过省略大量信息而简化对复杂问题的处理，故任何假设模型都必然有误差。

测量平差中的模型分为两类：一类是随机模型，一类是函数模型。因此，模型误差也分为两类：即随机模型误差和函数模型误差。

前面讨论的观测粗差、误差的重尾分布以及误差的近似独立假设和零期望、等方差假设，均属随机模型误差；Andrews等

人曾提到的短尾 (short-tailed) 正态分布也属随机模型误差。但短尾分布不如长尾分布危险, 而且短尾分布也不多见。实际上, 短尾分布一般出现在三种场合: (1) 数据的取值范围本来就十分有限; (2) 数据在较原始状态已进行人为的筛选, 如剔除一切可疑测值等, 因而出现人为的截断分布; (3) 数据与数据之间的相关也可能导致短尾分布。

抗差估计理论也给出了短尾分布的一些推论和定性结论^[1]。

测量平差中函数模型误差与回归分析模型误差不同。这里的函数模型误差一般是由于非线性观测方程的线性化, 以及许多物理因素所致。例如, 我国大地网的布测经历了近40多年的时间, 由于地球重力场的变化、地壳变化等动态地球物理因素的影响, 多期复测结果的静态联合平差模型必然偏离实际的函数模型。即使是顾及速率参数的“动态模型”也难以准确地描述幅员如此辽阔、地壳结构极其复杂的广大疆土上的所有点之观测模型。

四、近似独立假设偏差

在测量平差和其它数据处理中, 常假设观测数据互相独立。因为独立观测数据的处理一般较为容易。目前大量抗差估计文献一般都基于独立同分布假设。但是对于原本是相关的观测序列, 若按近似独立假设进行处理, 则必然会给估计结果带来偏差, 相关观测给抗差估计带来了许多棘手的问题。第五章将重点进行讨论。

§1.3 异常影响诊断与抗差估计

当我们强调抗差估计的重要性的同时, 同样也不应忽视经典统计理论的重要性。对于模型偏差, 异常数据的处理, 经典统计理论中也有较成熟、且使用较广泛的方法——异常影响诊断法 (outlier influence diagnostics) 和误差统计检验法。异常影响诊断与抗差估计具有相同的目的——分析异常数据和模型偏差的影响, 并试图减弱或消除这些影响。

一、异常影响诊断

异常影响诊断指的是异常“观测”数据的影响诊断。作者认为异常影响诊断应比粗差探测含义更广。粗差一般指过失误差，而异常数据的来源除过失误差外，还有模型偏差、物理因素的影响等。异常数据的特点是，单个数据或成组数据远离主体数据群。

异常影响诊断有两个主要目的：一是发现和剔除粗差以保证参数估计和统计分析的“安全”性；二是分析异常值产生的背景，发现函数模型和随机模型的偏差，并加以调整。

第一个目的基于观测数据中含有粗差的情形。粗差显然是异常值，它对最小二乘估计危害极大。毫无疑问，这类异常数据应该从观测样本中予以剔除。一般情况下，粗差测值可视为来自别的分布子样，因此，剔除这类异常值才能保证所使用的统计模型的可靠性。然而，发现并剔除这类异常数据并不容易。粗差测值并非总是显而易见，常遇到剔除不彻底或误剔现象。剔除不彻底，则影响参数估计的安全性；误剔优质数据，又会造成估计效率的损失。

异常数据影响诊断的第二个目的是基于“异常”数据为非粗差测值情形。它较第一个目的更有意义。它不仅要保证所使用的模型和方法的“安全”，而且着重寻找出现“异常”的原因，为改正粗差测值、调整或改进模型提供依据。在许多情况下，异常数据（或异常数据群）可能来自不同的函数模型或随机模型，当所使用的函数模型不十分准确或非处处可靠时，则易导致某些观测呈现孤值或“异常”。例如，三角网中某一方向出现异常，如果观测未发生任何明显失误，则可能是由于该方向具有较强的旁折光影响或其它区域性影响。这时可认为该方向所使用的模型有偏差，即缺少系统改正项。对此，可以通过局部改进函数模型使其顾及局部旁折光影响，即在函数模型中加入表征旁折光影响的未知参数。

又例如，若观测数据中存在着局部相关性，而所使用的随机