DOI:10.13203/j.whugis2004.10.008
第29卷 第10期 武汉大学学报·信息科学版 Vol.29 No.10
2004年10月 Geomatics and Information Science of Wuhan University Oct., 2004
文章编号:1671-8860(2004)10-0881-04 文献标识码:A

基于相关分析的粗差可区分性

陶本藻[1] 姚宜斌[2] 施 闯[3]

(1 武汉大学地球空间环境与大地测量教育部重点实验室,武汉市珞喻路129号,430079)
(2 武汉大学测绘学院,武汉市珞喻路129号,430079)
(3 武汉大学GPS工程技术研究中心,武汉市珞喻路129号,430079)

摘 要:提出了应用偏相关系数来区分多维粗差和用复相关系数对多维粗差总体显著性进行检验并定位的方法。反复计算这些相关系数并进行检验,可以相对正确地进行多维相关观测的粗差定位。

关键词:相关分析法;偏相关系数;复相关系数;粗差可区分性
中图法分类号:P207

近年来,随着测量数据采集的现代化、自动化和高速化,测量界逐渐认识到粗差也是一种不可避免的观测误差。将粗差纳入函数模型,沿着巴尔达提出的粗差检测理论[1],已取得一大批开创性研究成果。将粗差纳入随机模型,引进统计学中的稳健估计,结合测量平差实际研究,已形成一套稳健最小二乘估计方法[2]。

巴尔达粗差检测理论主要研究了粗差检测统计量的分布、数据探测、粗差的估计和大地网的可靠性等理论;此后,Koch[3]、Pelzer[4]等学者对多维粗差的检测方法作了较多研究;李德仁[5]系统地论述了巴尔达粗差理论的进展,并提出了多维粗差的可区分理论,为粗差理论的研究扩展了思路;於宗俦等[6]直接从最小二乘平差系统中观测误差与其残差的关系式出发,提出了一种多维粗差的同时定位定值方法;欧吉坤[7]从测量误差与残差的关系式出发,给出了检测粗差的拟准检定法;施闯针对相关观测的粗差探测,提出了一种基于相关分析的粗差理论[8~10],为多维相关观测粗差的检测开辟了一条新途径,其特点是定义了测量误差作用于残差的影响向量,利用影响向量间的相关系数定位粗差,给出了同时探测多个粗差和逐个探测多个粗差的相关分析步骤和程序,通过模拟试验并用于国家高精度GPS B级网的平差计算中,取得了满意的结果。文献[10]还提出了粗差的可检测性和可区分性问题。

本文是在文献[8~10]基础上的补充研究,提出了通过分析影响向量的偏相关系数来区分难以识别的粗差,并采用复相关系数分析来判定多维粗差总体的显著性。通过利用相关系数、偏相关系数和复相关系数的相关分析方法,对相关观测多维粗差进行检测、定位以及进行粗差的可检测性和可区分性分析,初步形成了研究粗差的一种基于相关分析的粗差探测理论体系。

1 基于相关分析的粗差检测原理

在最小二乘平差系统中,测量误差 ε 与残差 $\boldsymbol{V}$ 的关系式为[11]:

$$-\underset{n,1}{\boldsymbol{V}}=\boldsymbol{Q}_{VV}\boldsymbol{P}\underset{n,1}{\varepsilon}=\boldsymbol{R}\underset{n,1}{\varepsilon} \tag{1}$$

式中,$\boldsymbol{R}$ 为可靠性矩阵;$\boldsymbol{V}$ 是观测值改正数向量;$\boldsymbol{P}$ 为观测量的权矩阵;$\boldsymbol{Q}_{VV}$ 是观测值改正数的协因数矩阵;ε 是观测值误差。其中,

$$\boldsymbol{R}=\begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & & \vdots \\ r_{n1} & \cdots & r_{mn} \end{bmatrix}=\begin{bmatrix} \boldsymbol{F}_1 & \boldsymbol{F}_2 & \cdots & \boldsymbol{F}_n \end{bmatrix} \tag{2}$$

式中,$\boldsymbol{F}_i=\begin{bmatrix} r_{1i} & r_{2i} & \cdots & r_{ni} \end{bmatrix}^{\mathrm{T}}$。在文献[8]中,$\boldsymbol{F}_i$ 被定义为测量误差作用于残差的影响向量。$\boldsymbol{F}_i$ 由平差系统的结构矩阵 $\boldsymbol{A}$ 和权阵 $\boldsymbol{P}$ 所决定,反映了观测量 l_i 的误差 ε_i 对残差 $\boldsymbol{V}$ 的内在影响关系和作用程度,即 ε_i 通过 $\boldsymbol{F}_i$ 作用于 $\boldsymbol{F}$ 中,而 $\boldsymbol{F}_i$ 起到了对 ε_i 的缩放作用。

收稿日期:2004-06-10。
项目来源:国家自然科学基金资助项目(40004001)。

当 ε 中完全不含粗差时，ε 中各分量对 $\boldsymbol{V}$ 的影响不显著，$\boldsymbol{V}$ 是一组偶然误差的线性组合，呈现偶然性。若残差 $\boldsymbol{V}$ 中有粗差的影响，就要检验粗差来自哪些观测量。如果 $\boldsymbol{V}$ 中含有一个粗差 ε_i，一般通过 $\boldsymbol{F}_i$ 的作用 $\boldsymbol{F}_i\varepsilon_i$ 会显著影响 $\boldsymbol{V}$ 的大小并改变其随机性质。从统计意义而言，此时 $\boldsymbol{F}_i$ 与 $\boldsymbol{V}$ 相对地具有较强的相关性，相关程度高低在一定程度上反映了 $\boldsymbol{F}_i\varepsilon_i$ 对 $\boldsymbol{V}$ 的作用大小，从而为识别粗差提供了一个途径。

$\boldsymbol{F}_i$ 与 $\boldsymbol{V}$ 的相关性可用相关系数度量：

$$\rho_{\boldsymbol{F}_i\boldsymbol{V}}=\frac{\sigma_{\boldsymbol{F}_i\boldsymbol{V}}}{\sigma_{\boldsymbol{F}_i}\sigma_{\boldsymbol{V}}} \tag{3}$$

其估值为：

$$\rho_{\boldsymbol{F}_i\boldsymbol{V}}=\frac{\sum_{j=1}^{n}(r_{ji}-\bar{r}_i)(v_j-\bar{v})}{(\sum_{j=1}^{n}(r_{ji}-\bar{r}_i)^2\sum_{j=1}^{n}(v_j-\bar{v})^2)^{\frac{1}{2}}} \tag{4}$$

式中，$\bar{r}_i=\frac{1}{n}\sum_{j=1}^{n}r_{ji}$；$\bar{v}=\frac{1}{n}\sum_{j=1}^{n}v_j$。

计算 $\rho_{\boldsymbol{F}_i\boldsymbol{V}}$，进行相关系数统计检验。选取显著水平 α，如果 $\rho_{\boldsymbol{F}_i\boldsymbol{V}}>\rho_\alpha$，则相关性显著，可认为 ε_i 为粗差的可能性较大；反之，$\rho_{\boldsymbol{F}_i\boldsymbol{V}}<\rho_\alpha$，则 ε_i 为粗差的可能性较小。

当观测量中含有多个粗差时，改正数向量 $\boldsymbol{V}$ 的变化规律将表现为来自这些粗差对其影响的叠加。各粗差观测量的 $\boldsymbol{F}_i$ 都将显示出与 $\boldsymbol{V}$ 有着相对显著的相关性，相关性的大小取决于粗差个数、分布和粗差值的大小。所以当相关观测量中含有多个粗差时，只要粗差是可测的，通过 $\rho_{F_i\boldsymbol{V}}$ 仍能分析出粗差所在的位置。

基于上述相关分析对粗差探测的思想，文献[8~10]提出了同时探测多个粗差或逐个探测粗差的方法。

2 粗差可区分性的偏相关系数分析法

如果有两个观测量 l_i 和 l_j，其影响向量为 $\boldsymbol{F}_i$ 和 $\boldsymbol{F}_j$，则其相关系数为[8]：

$$\rho_{\boldsymbol{F}_i\boldsymbol{F}_j}=\frac{\sum_{k=1}^{n}(r_{ki}-\bar{r}_i)(r_{kj}-\bar{r}_j)}{(\sum_{k=1}^{n}(r_{ki}-\bar{r}_i)^2\sum_{k=1}^{n}(r_{kj}-\bar{r}_j)^2)^{\frac{1}{2}}} \tag{5}$$

当 $|\rho_{F_iF_j}|$ 越接近于 1 时，这两个观测量粗差的可区分性越差。$|\rho_{F_iF_j}|=1$ 时，ε_i 与 ε_j 完全不可区分。

当 $|\rho_{\boldsymbol{F}_i\boldsymbol{F}_j}|=1$ 时，说明向量 F_i 与 F_j 完全相关，它们互成线性组合。$\boldsymbol{V}$ 中包含 ε_i 还是 ε_j 的影响确实不可区分。由于残差与观测量 $\boldsymbol{l}$ 间有如下关系[11]：

$$-\boldsymbol{V}=\boldsymbol{R}\boldsymbol{l} \tag{6}$$

如果 $\boldsymbol{F}_i$ 与 $\boldsymbol{F}_j$ 完全相关，说明观测量 l_i 和 l_j 为两个函数相关的观测量，平差中很少出现。如果出现，则需要合并观测量，再进行检验。

当 $|\rho_{\boldsymbol{F}_i\boldsymbol{F}_j}|\neq1$ 且 F_i 与 F_j 强相关时，两个粗差的区分是指判定 ε_i 与 ε_j 中的哪一个是粗差，或是否两个都为粗差。这种区分有一定的难度。研究表明，通过下述的偏相关系数分析可望达到区分粗差的目的。

不失一般性，设经过前述的误差检测的相关分析判定 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_q$ 为粗差，其中包括虽然 $\boldsymbol{F}_i$ 和 $\boldsymbol{V}$ 不相关，但 $\boldsymbol{F}_j$ 与 $\boldsymbol{F}_i$ 强相关，而且 $\boldsymbol{F}$ 和 $\boldsymbol{V}$ 相关的 $\varepsilon_j,\varepsilon_{q+1},\varepsilon_{q+2},\cdots,\varepsilon_n$ 不为粗差，则由式(1)可得：

$$-\boldsymbol{V}=\boldsymbol{F}_1\varepsilon_1+\cdots+\boldsymbol{F}_q\varepsilon_q+\boldsymbol{F}_{q+1}\varepsilon_{q+1}+\cdots+\boldsymbol{F}_n\varepsilon_n \tag{7}$$

此时可令

$$-\delta=\boldsymbol{F}_{q+1}\varepsilon_{q+1}+\cdots+\boldsymbol{F}_n\varepsilon_n$$

δ为偶然误差，并令

$$\boldsymbol{R}_Q=[\boldsymbol{F}_1\quad\boldsymbol{F}_2\quad\cdots\quad\boldsymbol{F}_q]$$

$$\varepsilon_Q=[\varepsilon_1\quad\varepsilon_2\quad\cdots\quad\varepsilon_q]^{\mathrm{T}}$$

则可建立误差方程：

$$\hat{\delta}=\boldsymbol{R}_Q\hat{\varepsilon_Q}+\boldsymbol{V} \tag{8}$$

在 $\hat{\delta}^{\mathrm{T}}\hat{\delta}=\min$ 下，组成法方程：

$$\boldsymbol{R}_Q^{\mathrm{T}}\boldsymbol{R}_Q\hat{\varepsilon_Q}+\boldsymbol{R}_Q^{\mathrm{T}}\boldsymbol{V}=0 \tag{9}$$

算得二次型：

$$\Omega=\hat{\delta}^{\mathrm{T}}\hat{\delta}=(\boldsymbol{R}_Q\hat{\varepsilon_Q}+\boldsymbol{V})^{\mathrm{T}}(\boldsymbol{R}_Q\hat{\varepsilon_Q}+\boldsymbol{V}) \tag{10}$$

偏相关系数是指在 $\boldsymbol{R}_Q$ 中扣除了除 $\boldsymbol{F}_i$ 之外的其他 $q-1$ 个向量影响后的 $\boldsymbol{F}_i$ 与 $\boldsymbol{V}$ 的相关程度，其计算式为[11]：

$$\rho_{\boldsymbol{F}_i}=\sqrt{\frac{\Omega_1-\Omega}{\Omega_1}} \tag{11}$$

设在 $\boldsymbol{R}_Q$ 中扣除了 $\boldsymbol{F}_i$ 后的 $q-1$ 个向量为 $\boldsymbol{R}_{Q-1}$，相应的粗差向量为 ε_{Q-1}，则其误差方程为：

$$\hat{\delta_1}=\boldsymbol{R}_{Q-1}\hat{\varepsilon_{Q-1}}+\boldsymbol{V} \tag{12}$$

类似地，可算得 $\hat{\delta_1}$ 的平方和为：

$$\Omega_1=\hat{\delta_1}^{\mathrm{T}}\hat{\delta_1} \tag{13}$$

因此, 扣除与不扣除 $\boldsymbol{F}_i$ 的误差平方和的差值 $\Omega_1-\Omega$ 与 Ω_1 之比的开方即是偏相关系数。偏相关系数越大, 反映了 $\boldsymbol{F}_i$ 的存在对 $\boldsymbol{V}$ 的影响越大, 或者说 ε_i 为粗差的显著性越高。

通过偏相关系数的检验, 可以对 ε_i 和 ε_j 进行区分。偏相关系数的检验通常可采用 F 分布检验法, 其检验统计量为:

$$T_i=\frac{\Omega_1-\Omega}{\Omega/(n-q)} \tag{14}$$

即使影响向量 $\boldsymbol{F}_i$ 与 $\boldsymbol{F}_j$ 之间或 $\boldsymbol{F}_i$、$\boldsymbol{F}_j$ 与 $\boldsymbol{V}$ 之间的相关系数通过了检验($\rho>\rho_\alpha$), 即相关性显著, 只要偏相关系数 ρ(或 ρ_{F_j})不通过检验, 即式(14)中的 $T_i<F_{\alpha(1,\,n-q)}$, ε_i(或 ε_j)就不能判定为粗差; 反之, $T_i>F_{\alpha(1,\,n-q)}$, 就可判定 ε_i(或 ε_j)为粗差。

3　粗差显著性判定的复相关系数法

式(8)中列入的粗差 ε_Q 在总体上是否显著, 可通过函数模型的有效性进行方差分析。

设对 $\boldsymbol{V}$ 的离差平方和进行分解得:

$$\begin{aligned}\Omega_V&=\sum(v_j-\bar{v})^2=\\&\sum[(v_j-\hat{v_j})+(\hat{v_j}-\bar{v})]^2=\\&\sum(v_j-\hat{v_j})^2+\sum(\hat{v_j}-\bar{v})^2=\Omega+\Omega_\varepsilon\end{aligned}\tag{15}$$

式中, $\boldsymbol{V}$ 是给定值; Ω_V 是不变的, $\Omega=\hat{\boldsymbol{\delta}}^{\mathrm{T}}\hat{\boldsymbol{\delta}}$, Ω 大, 则 Ω_ε 小; 反之, Ω_ε 大, 则 Ω 小。因此, 衡量函数模型的有效性可用如下指标:

$$\rho_R=\sqrt{\frac{\Omega_\varepsilon}{\Omega_V}}=\sqrt{1-\frac{\Omega}{\Omega_V}} \tag{16}$$

此即回归分析中定义的复相关系数[11]。给定显著水平, 查复相关系数表得分位值 ρ_{R_α}, 如果 $\rho_R>\rho_{R_\alpha}$, 则认为函数模型有效, 亦即 ε_Q 总体上显著。

Ω_ε 实质上是函数模型中全部自变量(F_Q)的方差贡献, ρ_R 就是这种贡献在总和中所占的比重, 由于 $0\leqslant\rho_R\leqslant1$, ρ_R 越接近于 1, ε_Q 总体上显著性越高。计算在式(8)中引入不同的观测粗差所得的复相关系数, 比较其大小, 可比较正确地定位多维粗差。

用下列检验统计量进行 F 分布检验可代替复相关系数检验:

$$T=\frac{(\Omega_V-\Omega)/q}{\Omega/(n-q)} \tag{17}$$

4　结　语

1) 利用相关系数、偏相关系数、复相关系数对多维相关观测进行粗差探测的相关分析法, 为研究粗差理论开辟了一条途径。

2) 通过偏相关系数的计算及检验, 不仅可以进行两个观测量影响向量是否具有强相关的粗差定位, 而且还可以筛选显著性不大的粗差。

3) 反复应用复相关系数检验, 选取复相关系数取最大值时所包含的粗差向量 ε_Q, 可定位出在 $\boldsymbol{V}$ 中最显著的一组粗差。

参　考　文　献

1　Baarda W. A Testing Procedure for Use in Geodesy Networks. Neth. Geod. Comm. New Series, 1968, 2(5)

2　刘经南, 姚宜斌, 施　闯. 基于等价方差-协方差的稳健最小二乘估计理论研究. 测绘科学, 2002(3): 15

3　Koch K R. Parametersch atzung und Hypothesentests in Linearen Modellen. Bonn: DüMMLER, 1980

4　Pelzer H. Some Criteria for the Accuracy and Reliability of Networks. DGK, Reihe B, Nr, 252, 1980, 55~67

5　李德仁. 误差处理和可靠性理论. 北京: 测绘出版社, 1998

6　於宗俦, 李明峰. 多维粗差的同时定位与定值. 武汉测绘科技大学学报, 1996, 21(4): 323~329

7　欧吉坤. 一种统一的可靠性指标研究. 北京: 地震出版社, 1999

8　施　闯, 刘经南. 基于相关分析的粗差理论. 武汉测绘科技大学学报, 1998, 23(1): 5~9

9　施　闯, 刘经南. 国家高精度 GPS 整体平差中的粗差分析. 武汉测绘科技大学学报, 1999, 24(2): 107~111

10　施　闯. 大规模高精度 GPS 网平差与分析理论及其应用. 北京: 测绘出版社, 2002

11　陶本藻. 测量数据统计分析. 北京: 测绘出版社, 1992. 136~144

12　中科院计算中心概率统计组. 概率统计计算. 北京: 科学出版社, 1979. 110~118

第一作者简介: 陶本藻, 教授, 博士生导师, 现从事现代测量数据处理和地壳形变及地球动力学解释的研究。代表成果: 青藏高原地壳运动监测及地球动力学机制研究; 测量平差模型和模型误差的理论; 专著《测量数据统计分析》、《自由网平差与变形分析》等。发表论文百余篇。

E-mail: bztao@sgg.wtusm.edu.cn

Distinguishability of Outlier Based on Correlative Analysis

TAO Benzao[1] *YAO Yibin*[2] *SHI Chuang*[3]

(1 Key Laboratory of Geospace Environment and Geodesy, Ministry of Education, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

(2 School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

(3 Research Center of GPS, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

Abstract: This paper is the sequential research of the "correspondence based outlier analysis". For correlative observations, the partial correlation coefficients are used to distinguish the multidimensional outliers, and the compound correlation coefficients are used to proof-test the holistic prominence of multidimensional outliers. By computing the correlation coefficient time and again, the authors locate the outliers of multidimensional outliers more accurately.

Key words: correlative analysis; partial correlation coefficient; compound correlation coefficient; distinguishability of outlier

---

About the first author: TAO Benzao, professor, Ph. D supervisor. He is engaged in the theoretic research on data processing and geophysics interpretation of crustal deformation. His main achievements indude the project of monitoring the present-day crustal movements and studying its geodynamical mechanism in Qinghai-Tibet Plateau, the project of the theory of adjustment model and its errors, "The Monograph of Statistical Analysis of Surveying Data"; "The Monograph of Deformation Analysis of Free Net Adjustment", etc. His published papers are more than 100.

E-mail: bztao@sgg.wtusm.edu.cn

(责任编辑: 涓涓)

下期主要内容预告

- 论空间信息多级格网及其典型应用
- 高分辨率遥感影像的精纠正
- 一种改进的 ROAM 算法
- 一种用于 GPS 整周模糊度 OTF 求解的整数白化滤波改进方法
- 用遗传算法反演地壳的变密度模型
- 地球重力场与 KBR 系统频谱关系的建立与分析
- 移动 GIS 的原理、方法与实践
- 从平台 GIS 到跨平台互操作 GIS 的发展
- 基于 IKONS 的植被提取中的校正
- 矢量 GIS 平面随机线元误差模型建模机理
- 关于我国采用三维地心坐标系统和潮汐改正的讨论